# CAUSES AND CONSEQUENCES OF COPY NUMBER VARIATION OF THE HUMAN GLYCOPHORIN GENE CLUSTER

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Walid Khalid Algady MSc
Department of Genetics and Genome Biology
College of Medicine, Biological Sciences and Psychology
University of Leicester

2019

# Abstract


**Causes and Consequences of Copy Number Variation of the Human Glycophorin Gene Cluster**

**Walid Algady**

Human glycophorin proteins expressed on the surface of erythrocytes, and are receptors for invasion of the Plasmodium *falciparum* parasite, which causes malaria in sub-Saharan Africa. The proteins are encoded by the genes *GYPA* and *GYPB* which, together with *GYPE*, reside on a tandemly-duplicated repeat region on chromosome 4q31.21.

Sequence read depth data of the 1000 Genome Project were used to determine the glycophorin variants. The positive control samples of eight variants were Sanger sequenced and the breakpoints of them were identified and analysed, DEL1, DEL2, DEL6, DEL7, DUP14, DUP29, DUP5 and the gene conversion. In this thesis, paralogue ratio test (PRT) assays were developed to type CNV of the glycophorin gene regions in the Benin malaria cohort (n=563), and an allele-specific PCR assay to genotype alleles of the novel SNP (rs186873296), which is related to resistance to severe malaria in the same malaria cohort. This showed that absent of the *GYPB* is not associated with the Benin malaria cohort phenotypes. In addition, rs186873296 SNP is not in strong linkage disequilibrium (LD) with the *GYPB* deletion in this malaria cohort.

Previous genome-wide analysis has shown that a structural variant within this region (DUP4), is identical to the blood group antigen Dantu NE+, and confers a clinically-important protective effect, is common in East Africans and is strongly protective against severe malaria. DUP4 is a complex structural genomic variant that carries hybrid (*GYPA*/*GYPB*) fusion genes. Using fibre-FISH, we validate the structural arrangement of the glycophorin locus in the DUP4 variant, and provide evidence of somatic variation in the number of *GYPA*/*GYPB* fusion genes. Subsequently, we have developed a paralogue-specific junction fragment PCR to genotype DUP4. We demonstrate association of DUP4 variant with haemoglobin levels - a phenotype related to malaria - in 962 DNA samples from a Tanzanian village holoendemic for malaria using a family-based association test. Using the family-based association approach implemented in (QTDT), we have found a statistically significant association of the DUP4 variant with haemoglobin levels ($p=0.0054$). This thesis confirms the importance of the DUP4 variant in malaria protection, and raises the intriguing possibility of heightened somatic instability and somatic mosaicism at this locus in DUP4 carriers, which might confer added protection against malaria.

# Acknowledgements

I gratefully acknowledge my PhD sponsor in Saudi Arabia, The Presidency of State Security. This PhD has been a great experience for me and it would not have been possible to do without their financial support and the guidance I received from many people.

At the beginning, I would like to express my sincere gratitude to my wonderful supervisor Dr Edward Hollox for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. My thesis would not have been possible without his constant guidance, encouragement, support, patience and enthusiasm. I could not have imagined having a better advisor and mentor. I am in debt forever for him.

I would like to thank all collaborators: Dr Fengtang Yang and Dr Sandra Gomes (Wellcome Sanger Institute) and Dr Danielle Carpenter for her help on the QTDT analysis. Also I would like to thank my APG thesis committee Prof Mark Jobling and Dr Ezio Rosato for going carefully through my first, second and third year reports giving useful suggestions regarding my project.

I would also like to thank all past and present members of Dr Hollox group and Prof Jobling group at University of Leicester for all the help, encouragement and friendship. Special thanks for Dr Razan Abujaber, Dr Ezgi Kucukkilic for their help and encouragement at the beginning of my PhD.

Big thanks for my best friend in the entire world Mr Abdullah Al-Obaid for managing my personal affairs in Saudi Arabia during my PhD period, which makes my life easier.

Finally, from the deepest place of my heart special thanks to my wife, Rasha Alzurayer and my daughters, Leen and Faten for all of their emotional support. Special thanks for my parents, Khalid Algady and Faten Alshamari who made all this possible with their great continuous emotional support and believing in me. I will not forget my brothers Mohammed and Abdullah and my sisters Layla, Jory and Nouf for their emotional support.

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| aCGH | Array Comparative Genomic Hybridization |
| AIDS | Acquired Immune Deficiency Syndrome |
| AMY1 | Salivary Amylase |
| ANR | Agence Nationale de la Recherche |
| ARMS | Amplification Refractory Mutation System |
| BAC | Bacterial Artificial Chromosome |
| BAM | Binary Alignment Map |
| BLAT | Basic Local Alignment Search Tool |
| bp | Single Base Pair |
| C4 | Complement Component 4 |
| CCL3L1 | Chemokine (C-C Motif) Ligand 3-Like 1 |
| CEPH | Centre d'Étude du Polymorphisme Humain |
| CEU | Utah Residents (CEPH) with Northern and Western European Ancestry |
| CGH | Comparative Genomic Hybridization |
| CHB | Han Chinese in Beijing, China |
| CN | Copy Number |
| CNP | Copy Number Polymorphisms |
| CNV | Copy Number Variations |
| CR1 | Complement Receptor 1 |
| Cy3 | Cyanine 3 |
| Cy5 | Cyanine 5 |
| CYP2D6 | Cytochrome P450 2D6 |
| DEFB | Human Beta-Defensin |
| DEL | Deletion |
| DGV | Database of Genomic Variants |
| DMBT1 | Deleted in Malignant Brain Tumours 1 |
| DNA | Deoxyribonucleic acid |
| dNTPs | Deoxy Nucleotide Triphosphate |
| DTR | Dye Terminator Removal |
| DUP | Duplication |
| ECACC | European Collection of Authenticated Cell Cultures |

| | |
|---|---|
| FAM | Fluorescein Amidite |
| FCGR3B | Fc fragment of IgG, low affinity IIIb, receptor (CD16b) |
| Fibre-FISH | Fibre-Fluorescence In Situ Hybridization |
| FISH | Fluorescence *In Situ* Hybridization |
| FoSTeS | Fork Stalling and Template Switching |
| FREM3 | FRAS1 Related Extracellular Matrix 3 |
| Genome STRiP | Genome STRucture in Populations |
| GP | Glycophorin protein |
| GRChr37 | Genome Reference Consortium Human Build 37 |
| GSTM1 | Glutathione S-Transferase Mu 1 |
| GWAS | Genome Wide Association Study |
| GYP | Glycophorin gene |
| HapMap | Haplotype Map |
| HbAE | Haemoglobin Embryonic Subunit Alpha |
| HbE | Haemoglobin E |
| HbS | Haemoglobin S |
| HDFN | Hemolytic Disease of the Fetal and Newborn |
| HEX | Hexachloro-Fluorescein |
| HinDIII | Haemophilus influenzae D III |
| HIV | Human Immunodeficiency Virus |
| HoxD | Homeobox D Cluster |
| HRC | Human Random Control |
| JPT | Japanese in Tokyo, Japan |
| kb | Kilobase |
| KDa | kilodalton |
| L1 (LINE1) | Long Interspersed Element-1 |
| LCR | Low Copy Repeat |
| LD | Linkage Disequilibrium |
| LNA | Locked Nucleic Acid |
| LTR | Long Terminal Repeat |
| MAFFT | Multiple Alignment using Fast Fourier Transform |
| MAPH | Multiplex Amplifiable Probe Hybridization |
| Mb | Megabase |

| | |
|---|---|
| mCNV | Multiallelic Copy Number Variations |
| MCS | Molecular Combing System |
| MHC | Major Histocompatibility Complex |
| Mi | Miltenberger Antigen |
| MLPA | Multiplex Ligation Dependent Probe Amplification |
| MSP1 | Merozoite Surface Protein 1 |
| NAHR | Non-Allelic Homologous Recombination |
| NEB | New England Biolabs |
| NED | (TAMRA) Tetramethylrhodamine |
| ng | Nanogram |
| NGS | Next-Generation Sequencing |
| NHEJ | Non-Homologous End Joining |
| OR | Odds Ratio |
| PAS | Periodic Acid–Schiff |
| PCR | Polymerase Chain Reaction |
| PGM | Personal Genome Machine |
| PRT | Paralogue Ratio Test |
| qPCR | Quantitative Polymerase Chain Reaction |
| QTDT | Quantitative Transmission Disequilibrium Test |
| RBCs | Red Blood Cells |
| RFLP | Restriction Fragment Length Polymorphism |
| RHD | Rh blood group D antigen |
| ROMA | Representational Oligonucleotide Microarray Analysis |
| ROS | Reactive Oxygen Species |
| SAM | Sequence Alignment Map |
| SCD | Sickle Cell Anemia |
| SD | Segmental Duplications |
| SD | Standard Deviation |
| SDS-PAGE | Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis |
| SINE | Short Interspersed Nuclear Element |
| SNP | Single Nucleotide Polymorphism |
| SRD | Sequence Read Depth |
| SV | Structural Variation |

| | |
|---|---|
| TAE | Tris-Acetate-EDTA buffer |
| TBE | Tris-Borate-EDTA buffer |
| TNF-α | Tumor Necrosis Factor-α |
| UCSC | University of California, Santa Cruz |
| µg | Microgram |
| µl | Microliter |
| µM | Micromolar |
| UV | Ultraviolet Light |
| WB | Water Bath |
| WGA2 | Genome Amplification Kit |
| WHO | World Health Organization |
| YRI | Yoruba in Ibadan, Nigeria |

# Chapter 1: Introduction

## 1.1 Copy number variations

Human genetic variation is the genetic diversity or variation in alleles of genes of humans and represents the total amount of genetic diversity within the human genome at both the individual and the population level (Zarrei *et al*., 2015). The human genome shows extensive variation between different individuals in different forms (Girirajan *et al*., 2011). A variant is defined as any form of DNA variation irrespective of frequency and phenotypic effect (Jobling *et al*., 2014). These include single nucleotide changes, deletions, duplications and even changes in copy number of whole chromosomes (Girirajan *et al*., 2011). Polymorphism refers to genetic variation which has a minor allelic frequency of < 1% in a given population. Single nucleotide polymorphism (SNP) is the widely distributed and frequent variant that has a substitution, insertion or deletion of a single base (Frazer *et al*., 2009). According to Jobling *et al*. (2014), despite the definition of polymorphism, many variants described as SNPs can have less than 1% minor allele frequencies. The last update of the human genomes sequencing reveals that each genome contains ~3-4 million SNPs. On average one SNP every 1006 bp for Africans and 1250 bp for Europeans; moreover, according to the 1000 Genomes Project, a human genome contains on average ~10,000 nonsynonymous (missense) and ~30 nonsense SNPs (Abecasis *et al*., 2010).

A human diploid genome consist a copy of a certain gene on both homologous chromosomes. The major proportion of diversity in human polymorphic genetic variation in terms of bases covered until now is accounted for copy number variations (CNVs) (Craddock *et al*., 2010) contributing to the differences between individual humans (Hastings *et al*., 2009). CNVs are widespread in human genomes and a major source of genetic variation in humans (Sudmant *et al*., 2010; Zarrei *et al*., 2015; Zhang *et al*., 2009; Iafrate *et al*., 2004; Sebat *et al*., 2004). CNVs present in the human genome are at different levels; large microscopically visible chromosome anomalies (several megabases (Mb) or more) and submicroscopic copy number variation of DNA segments (tens to thousands of kb pairs) (Feuk *et al*., 2006; Redon *et al*., 2006). Copy number can be diallelic and multiallelic CNVs. Diallelic CNVs have two alleles and could produce three different genotypes in both deletion and duplication events. The diploid copy number of particular gene could be changed by a simple deletion event result in diploid copy number of two, one or zero and by a simple duplication event result in diploid copy number of two, three, or four (Figure 1.1).

Deletions, insertions and duplications of DNA sequences ranging from several kbs to Mbs in size, in comparison with a reference genome are collectively referred to as CNVs (Conrad *et al*., 2010). A CNV can be simple tandem duplication, or may involve complex gains or losses of homologous sequences at multiple sites in the genome (Figure 1.1). Up to 12% of the genome is subject to CNV (Conrad *et al*., 2010; Handsaker *et al*., 2015; Iafrate *et al*., 2004; Kidd *et al*., 2010; Korbel *et al*., 2007; Redon *et al*., 2006). However, rare CNVs are more likely to be detected with more people analysed, such as glycophorin CNVs (Conrad *et al*., 2010; Leffler *et al*., 2017). It has been reported that copy number varies in different organs and across tissues in the same individual and can arise both meiotically and somatically (Piotrowski *et al*., 2008).

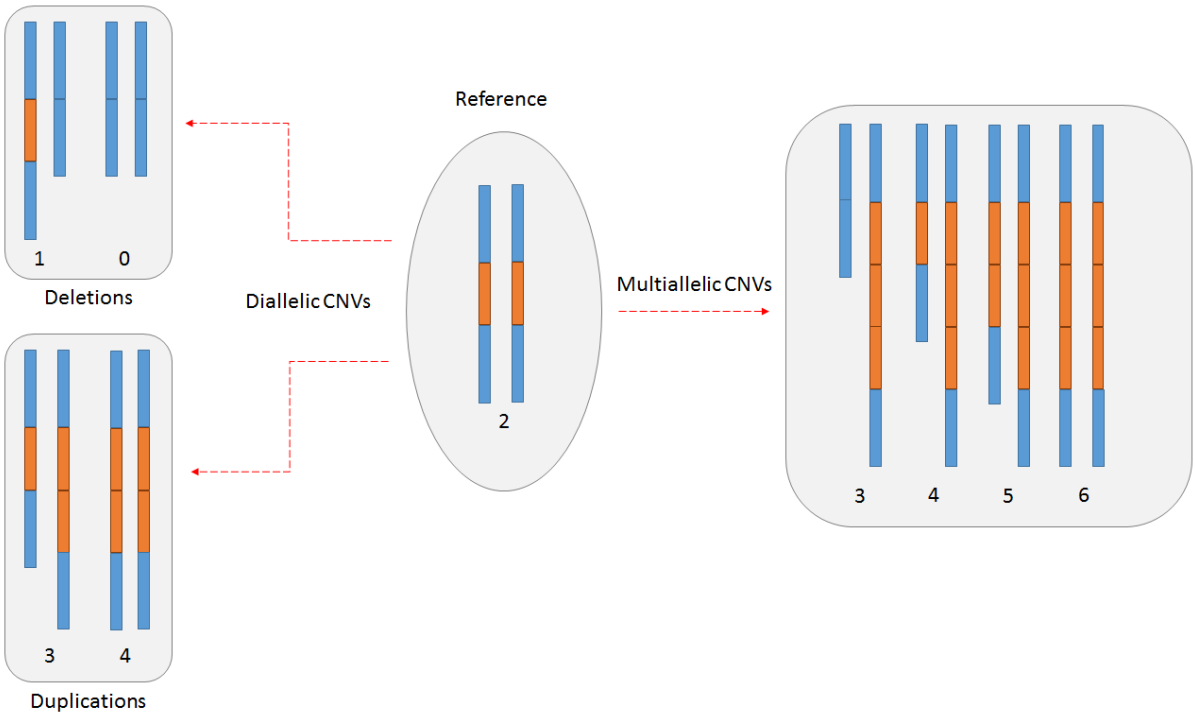



**Figure 1.1: Classification of the CNVs.** The diagram shows that a reference genome copy number could be varied to be deletion or duplication (Biallelic, Multiallelic or Complex). The copy number is shown under each haplotype.

However, the deletion and duplication events in the genome are not always simple (Scherer *et al*. 2007), they could result complex copy number variation, known as multiallelic copy number variants (mCNVs) (Wain *et al*., 2009). Multiallelic CNVs can be produced by successive rounds of duplication of a region of a genome, which results in more than two alleles and produces more than three genotypes (Figure 1.1). For instance, some of these complex CNV may range from 0 to 14 copies and 0 to 11 copies as in chemokine *CCL3L1* gene and *DMBT1* gene, respectively (Aklillu *et al*., 2013, Walker *et al*., 2009, Polley *et al*.,

2015). Multiallelic copy number variations (mCNVs) are considered to be the least characterized amongst other genetic variation. This is due to challenges of detection of this type of variation where loci exist in more states that cannot be explained by segregation of just two structural alleles (Handsaker *et al*., 2015). In general, the deletion and duplication sizes can vary from a few hundred to several million of bases and could cover an entire gene, a part of a gene, a region outside of a gene, or several genes.

## 1.2 Prevalence of CNVs in the human genome

The availability of a complete human genome sequence and significant advances in microarray technologies made it possible to obtain genome-wide maps of approximate locations and frequencies of CNVs. In 2004, two independent research groups investigated the human genome for CNVs. The first group employed Representational Oligonucleotide Microarray Analysis (ROMA) technology in investigating large-scale (>100 kb) CNVs in 20 healthy individuals with 85,000 probes that were 35 kb apart. The result of their studies showed that 221 CNVs were located at 76 CNV loci (Sebat *et al*., 2004). However, the second group used Bacterial Artificial Chromosome (BAC) Comparative Genomic Hybridization (CGH) array with approximately 1 Mb resolution where 55 individuals were investigated (Iafrate *et al*., 2004). 255 clones in the study revealed CNV, 41% of which were present in more than one person, showing that these variations are polymorphic.

Redon *et al*. (2006) studied 270 lymphoblastoid cell lines from the International HapMap project that was established from people of African, European, and Asian ancestry. In their study, they realised 1447 CNV regions occupying a sequence of 360 Mb in size. It was through their finding that CNVs were concluded to occupy 15% of the human genome. It was also discovered that an average of 12 CNVs exist in each person compared to a reference genome (Li and Olivier, 2013). According to Li and Olivier (2013), using a lower limit of 1 kb to define CNVs is discretionary because of the resolution difficulties. Therefore, a change in the threshold setting can radically change the number of CNVs reported.

The controversy of CNV was however summarized by Zhang *et al*. (2009) who claimed that approximately 30% (38,406 genomic variants) of the human genome is covered by CNV. In addition to this, they asserted that CNV is a DNA quantitative variation that exceeds 100bp. However, this value may have been over-exaggerated because of the technology's resolution (array_CGH) limits in screening for CNVs resulting in high rate of false positives in small

sized CNV calls. On the contrary, inexactness in determining CNVs varying between 1-20 kb could have contributed to underestimation of the total number of CNVs.

Conrad *et al*. (2010) designed an experimental strategy to discover CNVs greater than 500 base pairs using a set of 20 NimbleGen arrays (array_CGH), each comprising 2.1 million long oligonucleotide probes covering the assayable portion of the genome across 40 HapMap individuals. They identified 51,997 CNV calls, 11,700 CNV loci and 8,599 validated CNVs, 5,238 loci of which were genotyped allowing to distinguish deletions (0, 1 or 2 diploid copy number), duplications (2, 3 or 4 diploid copy number) and mCNVs (greater than 3 possible diploid copy numbers). This data set has been the core scientific resource on common CNVs for years (Conrad *et al*., 2010).

In 2015, Sudmant and colleagues (2015A) have looked at 1000 Genomes data which was mapped to the reference genome in an attempt to understand the pattern, selection, and diversity of CNVs in the context of the reference human genome. Sudmant *et al*. (2015A) sequenced 236 individual genomes from 125 different human populations and identified 14,467 autosomal CNVs and 545 X-linked CNVs, using a sequence read-depth approach which provided breakpoint resolution to 210 bps. They found that the median size of CNV was 7,396 bp, with 82.2% of events less than 25 kbp. CNVs mapping to SDs were larger on average (median = 14.4 kbp) than CNVs mapping to the unique parts of the genome (median = 6.2 kbp). Around 50% of CNV base pairs mapped within previously annotated SDs. In total, 7 % of the human genome is variable because of CNVs, in contrast to around 1 % resulting from single-nucleotide variations. Duplications were more common (4.4 % of the genome) compared with deletions (2.8% of the genome). When comparing their data set to Conrad *et al*. (2010), 67-73 % of calls were exclusive to their study, whereas they captured 68-77 % of formerly identified CNVs (Sudmant *et al*., 2015A). Another study analysed 1000 Genomes Project phase 3 whole-genome sequencing (WGS) data along with data from orthogonal techniques, including long-read single-molecule sequencing to characterize hitherto unresolved SV classes, the whole genome sequencing-based study, from 1000 genome project suggested that many segmental duplications in human genome carry CNV (Sudmant *et al*., 2015B). In addition, Handsaker *et al*. (2015) have analysed 849 genomes sequenced by the 1000 Genomes Project to identify most large (>5 kb) mCNVs, including 3,878 duplications, of which 1,356 appear to have three or more segregating alleles and they have found that mCNVs gives rise to most human gene-dosage variation and that this variation in gene dosage generates abundant variation in gene expression. Also, they have

described initial strategies for analysing mCNVs via imputation and provide an initial data resource to support such analyses.

## 1.3 Functional consequences of CNVs

The functional importance of many CNVs is relatively clear; reduced copy number of a gene can be correlated with reduced expression level, while duplicated copies of a gene can lead to increase expression level (McCarroll and Altshuler, 2007). 85% - 95% of CNVs in human and mice were reported to be associated with a change in expression of the affected genes (Stranger *et al*., 2007). Handsaker *et al*. (2015) sought to use whole-genome sequence data to understand mCNVs deeply. They analysed 849 genomes sequenced by the 1000 Genomes project to identify most large (>5 kb) mCNVs, including 3878 duplications, of which 1356 appeared to have 3 or more segregating alleles. In addition, Handsaker *et al*. (2015) discovered that mCNVs give rise to most human variation in gene dosage 7 times the combined contribution of deletions and bi-allelic duplications, and this in turn generates abundant variation in gene expression.

There are several ways that CNV can significantly affect phenotype. Firstly, gene dosage effects, where the level of protein is raised due to the alteration of the number of copies of the full gene and the mRNA. Gene dosage effects regulate many of the phenotypes that are related with CNVs, and which affect coding sequence and transcriptional regulation (Jobling *et al*., 2014), for example, (beta-defensin) (Jansen *et al*., 2009). CNV could create a fusion gene as a result of a deletion caused by unequal crossing over between two copies of DNA sequence. However, there would be no effect if both copies were identical, but if they have changed in sequence or regulation, this may lead to novel effects such as the Butyrophilin-like gene (Aigner *et al*., 2013). In addition, CNV can affect the position between a regulatory element (enhancer or repressor) and a gene, which is called position effect, such as *HoxD* in mice (Montavon *et al*., 2012). Furthermore, CNV could alter the number of tandem repeats of exons that code for a protein within a gene, which means more protein coding domains lead to an alteration of their functions. They also will affect the protein final size such as, Complement receptor 1 (*CR1*), the different isoforms vary in size by units of 30kDa. CR1-A (190kDa, also known as CR1-F) and CR1-B (220kDa, also known as CR1-S) are the most frequent alleles CR1-C (160kDa, also known as CR1-F') and CR1-D (250kDa) are rarer (Kucukkilic *et al*., 2018; Wong *et al*., 1989) (Figure 1.2).

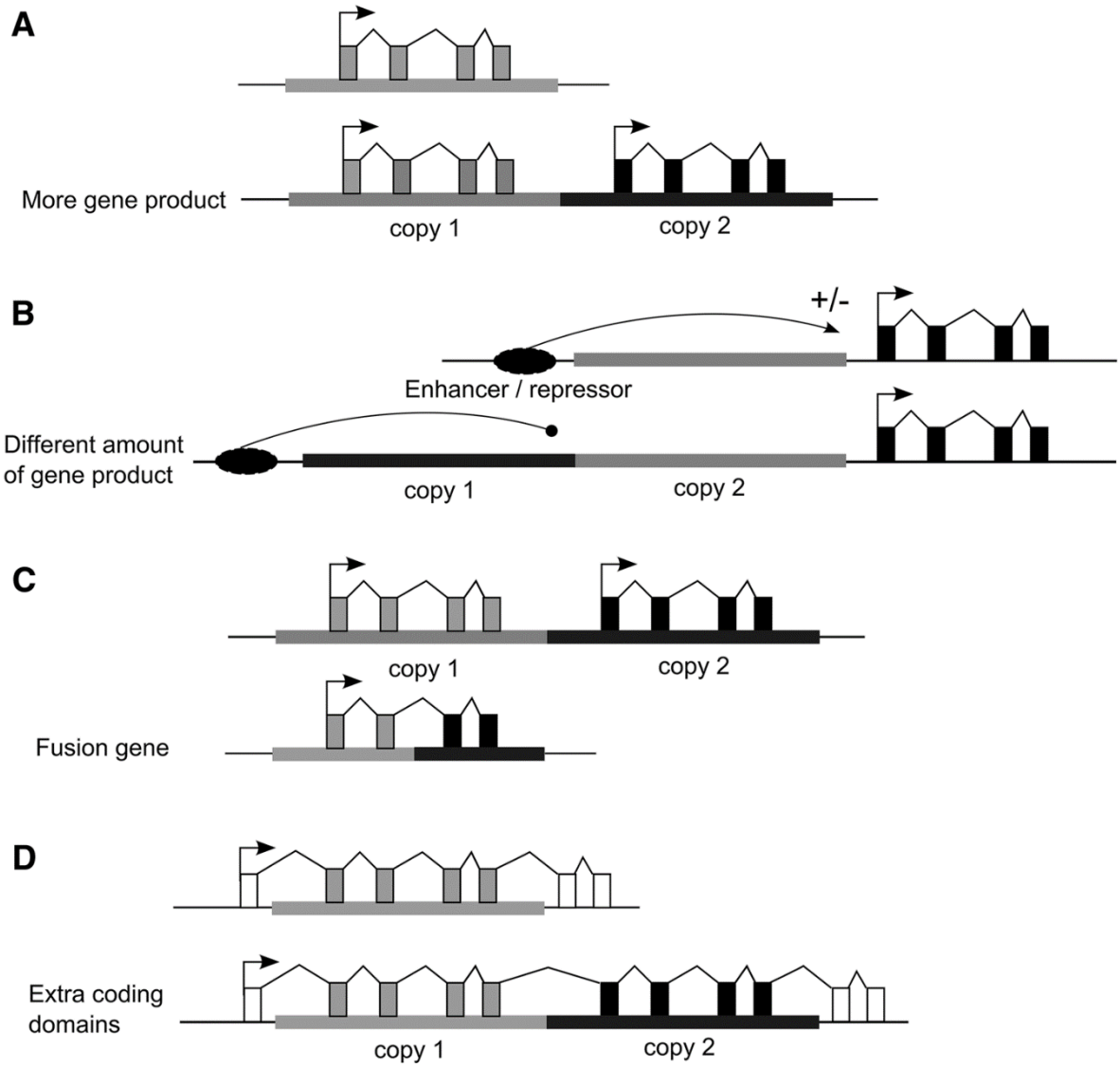



**Figure 1.2: Ways in which the contribution of CNV affects phenotype. (A)** Gene dosage effect, where CNV lead to a change of the total amount of mRNA and protein due to alteration in the number of DNA sequence of the gene. **(B)** Position effect. Copy number alters the distance between a regulatory element and the gene, which could lead to a different amount of gene products. **(C)** Fusion gene. From the figure, a deletion has occurred in two copies of DNA as a result of unequal crossing over between two DNA sequences, and the fusion gene will be created. **(D)** Coding domains effect. Here CNV could alter the number of tandem repeats of exons that code for a protein within a gene, which means more protein coding domains lead to an alteration of protein's function and final size Obtained from (Hollox and Hoh, 2014).

Most CNV is not likely to have phenotypic effect because only 2% of human genome is coding (Zarrei *et al*., 2015); therefore most random CNVs will not contain genes. The majority of the genes with CNVs play a part in the immune system, brain development and brain functioning (Ricklin *et al*., 2010). CNVs were firstly linked to human diseases in the 1980s (Feuk *et al*., 2006; Zhang *et al*., 2009). However; their population incidence was assumed to be not only small but also directly related to certain genomic disorders (Freeman *et al*., 2006; Ghanem *et al*., 1988). For instance, CNV at the α-globin locus was shown to be

the causing agent of α-Thalassaemia (Goossens *et al*., 1980). CNVs may also impair the performance of adjacent regulatory signals that activate or silence genes without directly influencing the copy number of that gene itself (Cahan *et al*., 2009; Henrichsen *et al*., 2009).

CNVs seem to be critical for evolution; some CNV copies sustain their original function whereas paralogues may undergo rapid adaptive evolution to specialize in their functional niche (Inoue and Lupski, 2002). In certain instances, changing in the copy numbers of certain CNV genes offer a selective advantage such as, a reduction in copy number being important has been proposed for the *α-globin* locus. The disorders of *α -globin* gene deletion in homozygotes, for example α thalassaemia, might be stabilized by resistance to malaria for heterozygotes (Higgs *et al*., 1989).

Importantly, there are somatic copy number changes and a distinction should be made are also involved in the formation as well as progression of cancer (Shlien and Malkin, 2009; Volik *et al*., 2006). In support of this, Frank *et al*., (2007) argued that copy number contributes to cancer proneness. Apart from causing cancer, CNVs increase susceptibility to schizophrenia (Ahn *et al*., 2014), epilepsy (Bassuk *et al*., 2013; Mefford *et al*., 2010), autism (Polan *et al*., 2014; Marshall and Scherer, 2012), Psoriasis (Hollox *et al*., 2008), and HIV (Hardwick *et al*., 2012; Larsen *et al*., 2012; Liu *et al*., 2010). Several CNV genes are involved in some known metabolising enzymes, such as *CYP2D6* and *GSTM1*. Others are widely studied such as the beta-defensins at the 8p23.1 genomic location due to their potential clinical relevance for innate immunity, inflammation, and cancer (Hollox *et al*., 2008). While others are potential drug targets such as *CCL3L1*, which may also make significant contributions to pharmacogenomic studies (Ouahchi *et al*., 2006). Recently, James *et al*. (2018) explore of the relationship between human beta-defensin (*DEFB*) copy number, cervicovaginal *HBD2* protein levels and antimicrobial activity in 203 women with risk factors for preterm birth. They have provided evidence that suggests *DEFB* copy number regulates cervical antimicrobial immunity.

According to recent CNV studies, the human genome has CNV distribution hotspots (Sudmant *et al*., 2015B). Structural variant-rich regions are hotspots such as the immunity and cell-cell signalling genes and genes that encode proteins which are involved in interaction with the environment. Another hotspot includes genes that code for retroviruses as well as transposition related proteins (Li and Olivier, 2013).

## 1.4 Imputation of CNVs

Analysis of CNVs, either via direct molecular analysis or statistical imputation (for common CNVs). Imputation from existing SNP data can be used to perform initial genome wide scans to propose specific mCNV loci for more analysis. The relationship between alleles of CNVs and flanking SNPs is not clear, and is likely to differ between different CNVs. For example, SNPs are in linkage disequilibrium with CNV such as the *RHD* (Rh blood group D antigen) deletion (Hardwick *et al*., 2011). On the other hand, some CNVs are only in weak linkage disequilibrium (LD), if at all, with neighbouring SNPs. For example, CNV of beta-defensin (Hardwick *et al*., 2011) and Fc gamma receptors (Hollox *et al*., 2009) are in weak LD with their flanking SNPs. This relation between SNPs and CNVs (LD) can be explained by mutation rate. The rate of CNV mutations when low copy repeat (LCR) involved lie in the range of $10^4$to $10^6$ per generation (1 in 800 bp on average in, 270 individuals, HapMap samples). Therefore, CNV mutation rates are ~3 orders of magnitude higher than those SNPs that this disrupts the LD relationship (Jobling *et al*., 2014).

Indeed, Campbell *et al*. (2011) has shown that a significant proportion of CNV involving segmental duplications (SDs) are not in linkage disequilibrium with nearby single nucleotide polymorphisms. This research has shown that 40% of 192 copy number polymorphisms (CNPs) located in SDs had high correlation to nearby SNPs, in comparison to 70% of 892 CNPs in unique regions of the genome. Therefore, this suggests that, in many cases, flanking SNP genotypes cannot be used as a proxy to type CNVs, but instead CNVs should be typed directly (Campbell *et al*., 2011).

Handsaker *et al*. (2015) have described initial strategies for analysing mCNVs through imputation and provide an initial data resource to support such analyses. They have divided the diploid copy number likelihoods among all potential combinations of copy number alleles and integrated this information with the genotype likelihoods for flanking SNPs in a population framework using the beagle4 imputation software to phase each CNV (Figure 1.3).

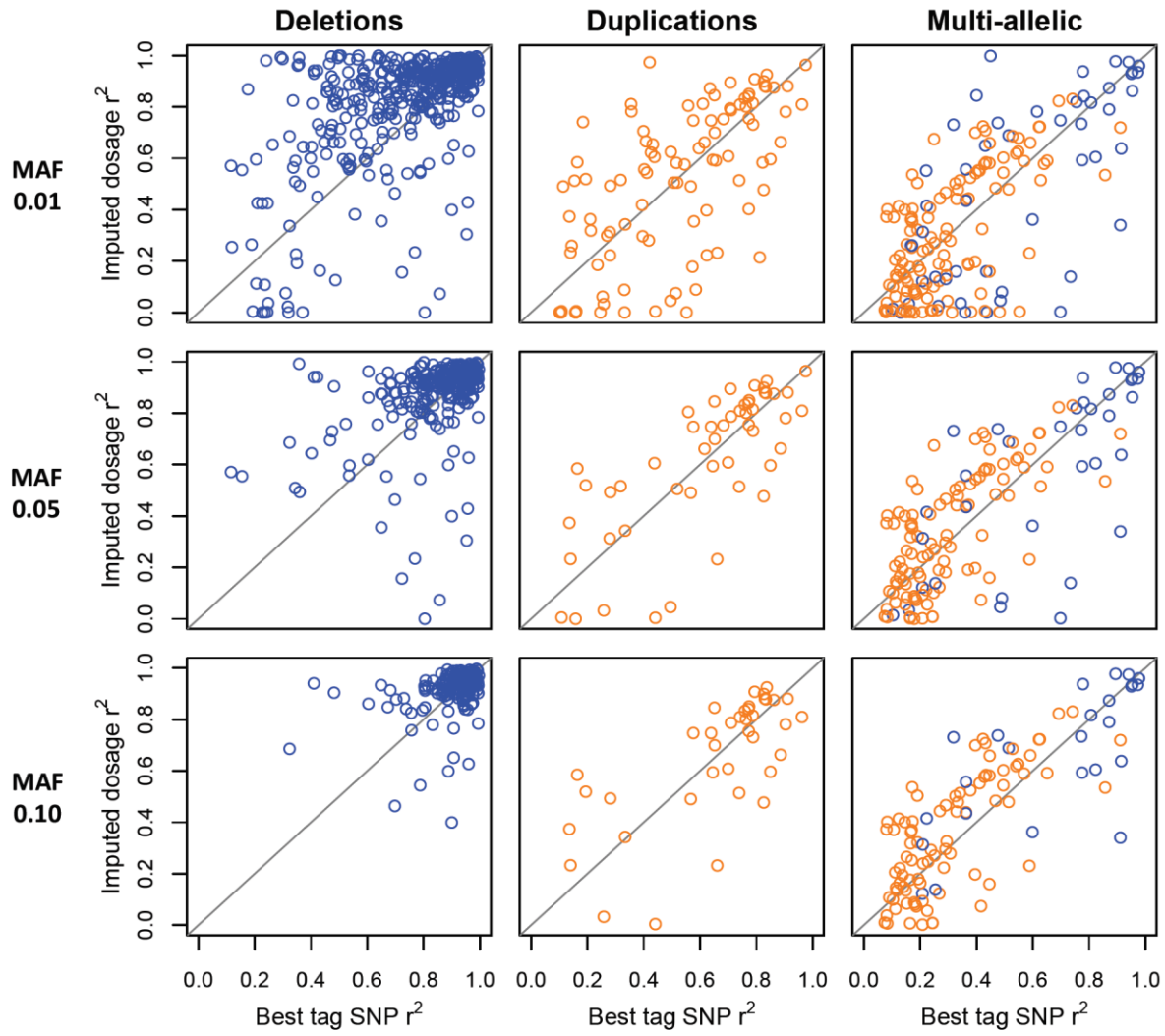



**Figure 1.3: Imputability of deletions, biallelic duplications, and multi-allelic CNVs.** For each CNV with non-modal AF > 1%, a set of leave-out trials was conducted in which 10 samples that were withheld at a time and imputed their allelic copy-number state based on the CNV genotypes from the other samples and flanking SNP genotypes for all samples (Handsaker *et al*., 2015).

At some cases, CNVs can be indirectly typed (imputed) because of LD with flanking SNPs. For example, Schizophrenia is a brain illness with unknown pathogenic mechanisms. Schizophrenia's strongest genetic association at a population level involves variation in the Major Histocompatibility Complex (MHC) locus, but the genes and molecular mechanisms accounting for this have been challenging to recognize. Sekar *et al*. (2016) have shown an association of schizophrenia with the MHC locus arises in substantial part from many structurally diverse alleles of the complement component 4 (C4) genes. They were able to impute C4A and C4B CNV from flanking SNPs and showed that the association with schizophrenia was due to the structural variations such as AL–BS structure. They have indicated that these alleles promoted widely varying levels of C4A and C4B expression and

associated with schizophrenia in proportion to their tendency to promote greater expression of C4A in the brain.

## 1.5 Mechanism of structural variation

There are four known mechanisms for the formation of copy number variation. These are nonallelic homologous recombination (NAHR), nonhomologous end-joining (NHEJ), FoSTeS, and L1 retrotransposition (Figure 1.4).

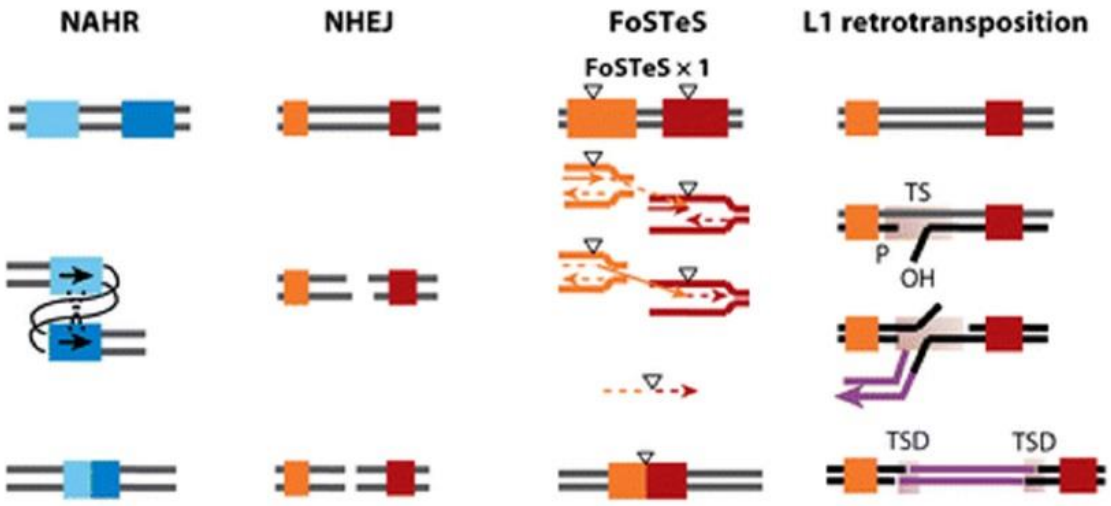



**Figure 1.4: The four major mechanisms for human genomic rearrangements and CNV formation.** Non-Allelic Homologous Recombination NAHR leads to alignment and crossover between two non-allelic DNA sequences, which have a high level of similarity in repeats; Non-Homologous End-Joining (NHEJ) is a mechanism for repairing DNA double strand breaks; Fork Stalling and Template Switching (FoSTeS) is a DNA replication-based model. This method occurs during DNA replication and uses the complementary template microhomology to anneal and prime DNA replication that happened with stalling and switching templates by the active replication fork; and the last one is L1 retrotransposition.

### 1.5.1 Non-Allelic Homologues Recombination (NAHR):

Nonallelic homologous recombination is caused by the alignment and crossover between two nonallelic DNA segments which share highly similar sequences (Zhang *et al*., 2009). The location of sponsoring sequences on the chromosome is important as NAHR between sequences repeated on same chromosome in the same orientations can lead to duplication or deletion, whereas inverted repeats cause inversion of genomic intervals surrounded by repeats. NAHR between sequences repeated on different chromosomes can cause chromosomal translocation (Stankiewicz and Lupski, 2002). Segmental duplications (SD) can be as subunits for NAHR because several reasons, they have extended homology for genetic rearrangements, they are more than 10Kb in length and the similarities range is (>95%)

(Carvalho and Lupski, 2016). However, different genomic rearrangements can be caused depend on the occurrence of the NAHR, meiosis or mitosis. For example, the occurrence of NAHR at meiosis, could leads to inherited genomic disorders (Lupski and Stankiewicz, 2005), whereas if it has occurred at mitosis, it will lead to mosaicism in the somatic cells carrying copy number or structural variant (Flores *et al*., 2007).

### 1.5.2 Nonhomologous End-joining (NHEJ)

NHEJ is a DNA repair system which fixes the DNA double-strand breaks caused by ionizing radiation or reactive oxygen species (ROS) such as hydrogen peroxide ($H_2O_2$). In contrast to NAHR, NHEJ does not need a substrate with extended homology and it can leave an information scar in the form of loss or addition of several nucleotides at the junction point (Carvalho and Lupski, 2016).

### 1.5.3 Fork Stalling and Template Switching (FoSTeS)

FoSTeS is a model for genomic rearrangements. In this model, the replication fork can delay while the lagging strand is disengaging from the original template and switching it to another replication fork. In the new fork it restarts DNA synthesis by priming it by way of the micro-homology between the switched template site and the original fork (Carvalho and Lupski, 2016; Lee *et al*., 2007).

### 1.5.4 L1 retrotransposition

Long interspersed element-1 (LINE-1 or L1) elements which comprise 16.89% of human genome (Zhang *et al*., 2009) are the only independent active nuclear transposons in the human genome. L1 transposition, which occurs via an RNA intermediate (transcribed by RNA polymerase), is followed by reverse transcription and integration. The resultant insertion is bounded by duplicated target sites (Zhang *et al*., 2009).

### 1.5.5 Recurrent and non-recurrent CNVs

CNV is classified based on the mutational origin and the formation molecular mechanism into two classes; frequently termed “recurrent” and “non-recurrent” CNVs. The mutation rates are thought to be different for recurrent and non-recurrent CNVs (Hollox and Hoh, 2014). Recurrent CNVs exist in large segmental duplications (SDs) regions within the genome. In an unrelated individual, the recurrent rearrangement of structural variation has the same size and genomic contents due to the recurrent endpoints which are mostly confined to a few genomic positions (Carvalho and Lupski, 2016, Hastings *et al*., 2009). Recurrent CNVs are fundamentally created by a non-allelic homologous recombination mechanism of CNV

formation. Recurrent CNVs represent 20-40% of normal polymorphic CNVs (Conrad et al. 2010), and can occur anywhere in the genome especially in subtelomeric and pericentromeric regions (Conrad and Hurles, 2007; Redon *et al*., 2006). However, non-recurrent CNVs involve large genomic regions and break-point analysis with minimal or no-homology is required for non-recurrent CNV formation (Conrad *et al*., 2010). Non-recurrent CNVs are mainly created by non-homologous end joining (NHEJ) or fork-stalling and template switching (FoSTeS) mechanisms (Carvalho and Lupski, 2016) and many of them are rare because negative selection acts to rapidly remove the deletion from the population (Hollox and Hoh, 2014). Non-recurrent CNVs are often large, are more likely to affect genes and hence more likely to have an extremely deleterious phenotypic effect (Arlt *et al*., 2012).

## 1.6 CNV detection methods

### 1.6.1 Array Comparative Genomic Hybridization (aCGH)

Comparative genomic hybridisation (CGH) is a technique that allows the detection of changes in chromosomal copy number, balanced and unbalanced structural and numerical chromosomal abnormalities (Baris *et al*., 2007; Jaillard *et al*., 2010) without the need for culturing cells. CGH is a fast screening technique that can detects specific chromosomal regions that might play a role in the pathogenesis or progression of tumours (Weiss *et al*., 1999). Agilent and NimbleGen were the first developers of commercial comparative genomic hybridization arrays (aCGH) platforms for calling copy number. The aCGH works basically with equivalent quantities of both the target and reference DNAs that are labelled with different fluorescent dyes, usually using Cyanine 3 (Cy3) and Cyanine 5 (Cy5) and co-hybridized on a probe array. The created slide is scanned using a microarray scanner and the spot intensities are measured and analysed for copy number analysis (Ahn *et al*., 2013; Jaillard *et al*., 2010; Baris *et al*., 2007; Feuk *et al*., 2006). In cases where the concentrations of the fluorescent dyes correlate with one probe, the region of the patient's genome is claimed to have equal quantity of DNA in the test as well as the reference samples. If the ratio is altered, it indicates relative losses or gains in a target sample (Figure 1.5).

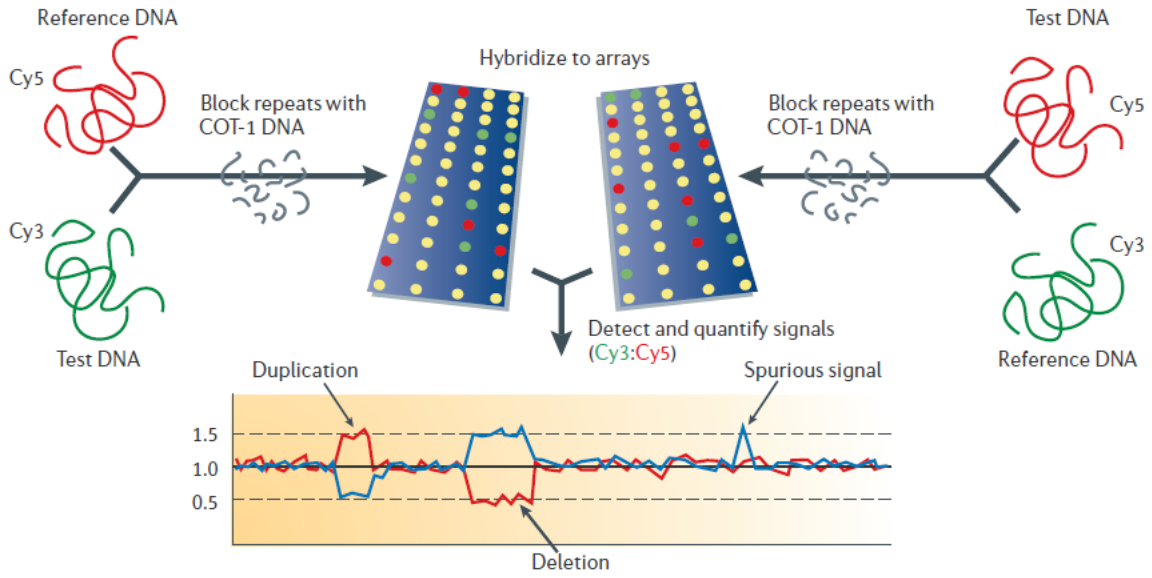



**Figure 1.5: Schematic picture of array-based comparative genome hybridisation (array-CGH).** Typically, in array-CGH, the initial labelling of the reference and test DNA samples reversed for a second hybridisation ('dye-swap') (left and right sides of the panel) to detect spurious signals. The red line represents the original hybridisation and the blue line represents the reciprocal hybridisation. Obtained from (Feuk *et al*., 2006).

### 1.6.2 Fibre-FISH (Fluorescence *in situ* hybridization)

Fibre-FISH is a modified FISH technique developed for high resolution mapping of genes and chromosomal regions on fibres of chromatin or DNA. Fibre-FISH permits physical ordering of DNA probes to a resolution of 1000 bp. The high resolution of fibre-FISH allows assessment of gaps and overlaps in contigs and analysis of segmental duplications and copy number variations. In fibre-FISH, the chromatin/DNA fibres are released from interphase nuclei and are stretched on a glass slide by means of salt or solvent extraction. After stretching, the DNA fibres are fixed on a microscope slide before hybridization. The stretching uniformity and reproducibility of DNA has improved significantly after implementation of the molecular combing protocol (Bensimon *et al*., 1994). In the molecular combing protocol the action of a receding air/water meniscus is used to extend and align DNA molecules at one end to a glass surface. Fibre-FISH allows the determination of copy number per allele which is important for studies of inheritance and diseases. The Salivary amylase gene (*AMY1*) copy number was successfully genotyped using fibre-FISH methods (Perry *et al*. 2007). The fibre-FISH requires a labour intensive workflow, low throughput and a high quality sample requirement and due to overlapping signals, highly variable regions are difficult to interpret (Cantsilieris *et al*., 2012).

### 1.6.3 Paralogue Ratio Test PCR (PRT-PCR)

The Paralogue Ratio Test (PRT) is a comparative PCR method based on amplifying dispersed repeat sequences that was developed by Armour *et al*. (2007) to determinate the copy number of a certain gene. The PRT primer pair co-amplifies a test region, which is copy number variable and a reference region, which is not variable in copy number using PCR (Deutsch *et al*., 2004; Armour *et al*., 2007; Hollox *et al*., 2008). The identical primer pairs improve reproducibility by making the amplification kinetics of test and reference loci very similar. The PCR amplicons can be distinguished easily by capillary electrophoresis based on their small size difference. The strategy of PRT reduces the problems that occur as a result of comparison of dissimilar amplicons with different amplification efficiencies. Experiments can be performed in duplicate by using two different fluorescent dyes to label the same primer, and then run both products on the same capillary (Figure 1.6).

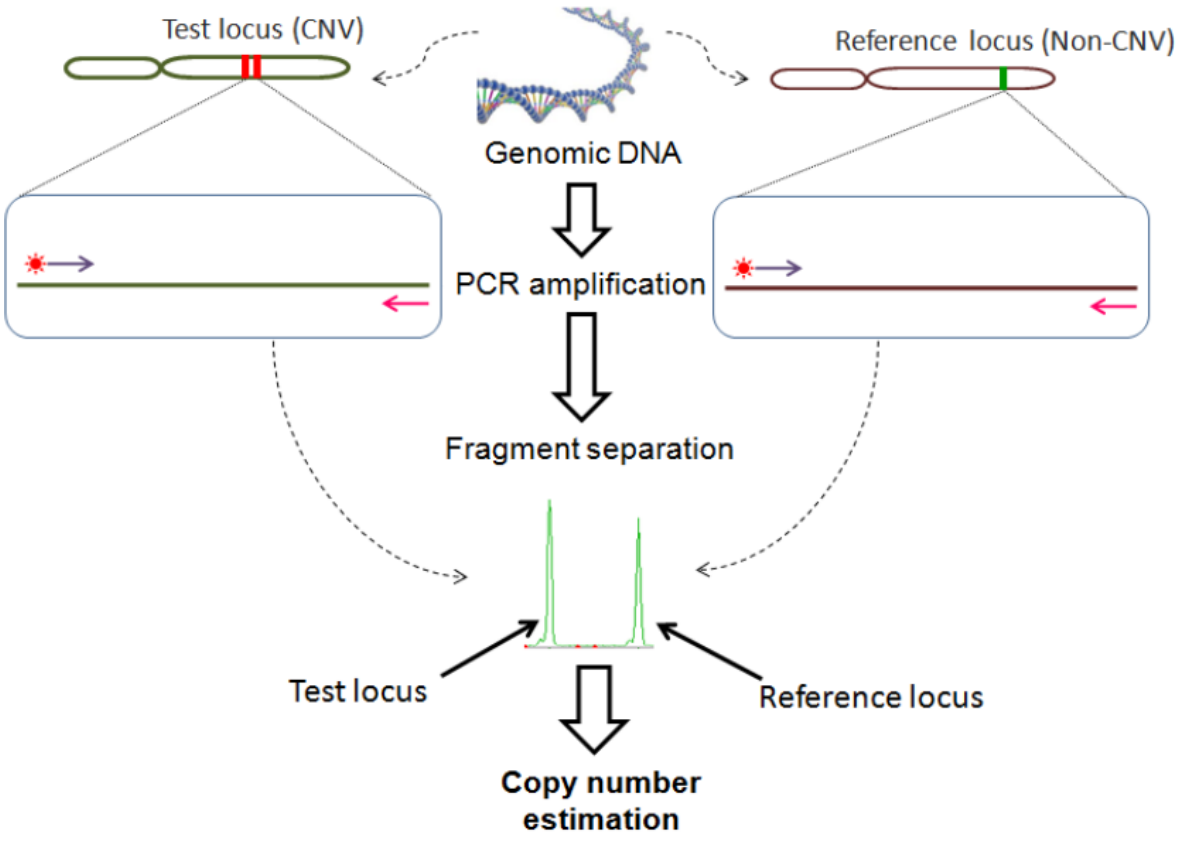



**Figure 1.6: Schematic picture describing different steps of the paralogue ratio test (PRT) assay.** The amount of PCR products are quantified with capillary electrophoresis and copy number is estimated by comparing amount of test products with reference products.

The PRT primer design is the main challenge of this technique. The primer must anneal to only the reference locus and copy number variable locus. This can be achieved with the help of an algorithm, which can quickly design couple of primers, suitable for PRT methods. The

algorithm BLASTs the region of interest with the entire genome sequence, masking repeated regions, in order to find specific and unique paralogous regions and in combination with primer design software, selects the oligos annealing only for those (Veal *et al*., 2013). Raw copy integer number can be determined by calculating the peak area ratio between the test and the reference amplicon (Aldhous *et al*., 2010; Abu Bakar *et al*., 2009; Armour *et al*., 2007).

PRT is robust, accurate, inexpensive, rapid and relatively high-throughput method for identifying gene copy number at a single loci. PRT has the advantage of high-throughput analysis for CNV typing of large cohort effectively using small amount (5-10 ng) of DNA (Armour *et al*., 2007; Hardwick *et al*., 2014; Machado *et al*., 2013; Aklillu *et al*., 2013; Abu Bakar *et al*., 2009; Hollox, 2008; Wain *et al*. 2014; Hardwick *et al*. 2012; Hollox *et al*., 2008). Furthermore, PRT is very powerful for analysing complex regions in the genome with SDs (Aldhous *et al*., 2010; Armour *et al*., 2007). PRT has shown more accuracy in determining integer CN and CN measurement >4 than qPCR (Fode *et al*., 2011). Since the appearance of the PRT technique, it has been widely applied for analysing mCNVs in several genes such as beta-defensin, *CCL3L1*, *FCGR3B* and Complement Component 4 (*C4*) (Hollox *et al*., 2009; Carpenter *et al*., 2011; Fernando *et al*., 2010).

### 1.6.4 Sequencing based methods

High throughput sequencing enables rapid sequencing of the base pairs in DNA samples. Next generation sequencing is a category of sequencing methods developed since 2005. The most commonly used NGS platforms are Illumina (HiSeq and MiSeq), and Ion Torrent (PGM and Proton), which can be classified as second-generation methods and platforms developed by Pacific Biosciences and Oxford Nanopore, which can be classified as third-generation methods.

The second-generation sequencing platforms rely on short fragments (reads) of 150-400 bp, each of which can be considered as an independent experiment. Due to this, every nucleotide is sampled several times by several reads spanning that specific position in the genome (coverage). This strategy raises the amount of data produced at each sequencing run since multiple experiments are running in parallel at the same time. In the other hand, third-generation sequencing platforms have been optimised in order to maximise the read length based on the length of DNA fragments. Pacific Biosciences was reported to produce 20 kb read length data (Korlach *et al*., 2017), while Nanopore promises to generate 882 kb read

length data (Jain *et al*., 2018). With this approach, unamplified DNA samples can be directly sequenced avoiding any PCR step. NGS platforms can use different methods, such as pyrosequencing, which relies on emulsion PCR (Roche 454 and Ion Torrent) and sequencing-by-synthesis, which relies on bridge PCR (Illumina) (Figure 1.7).

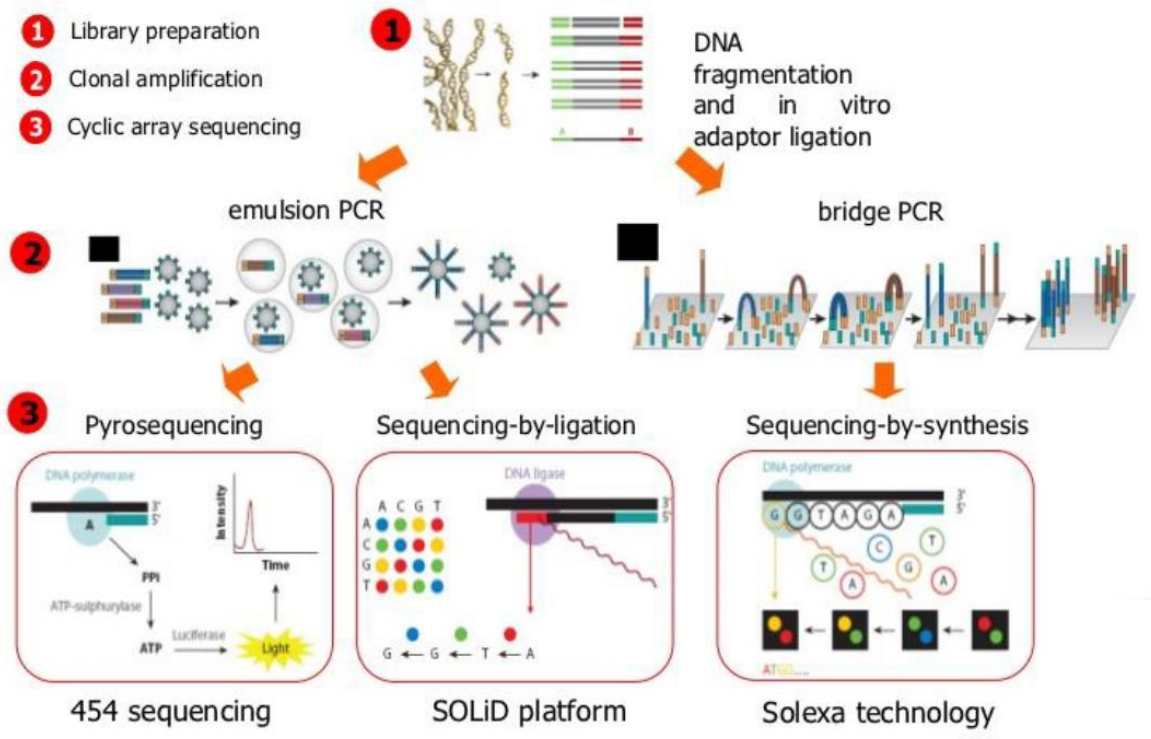



**Figure 1.7: Schematic diagram showing the workflow of NGS technologies.** The first step of this method is to fragment genomic DNA to a uniform size. Sequence enrichment could be performed for targeted or exome sequencing; whole genome sequencing does not require enrichment. To make the library sequencing-ready, adapters are ligated to both ends of the DNA. The different coloured adapters (green and red) reflect a sequencing adapter and a barcoding adapter. The library is then immobilised to an array where bridge amplification occurs to generate clusters (clonal libraries). Sequencing is then done by the various available methods as per the manufacturer's orders (Shendure and Ji, 2008).

### 1.6.4.1 Read-Pair (Mate-pair)

The utility of high-throughput sequencing data for CNV detection was first demonstrated by read pair (RP) methods (Pirooznia *et al*., 2015). There two kinds of reads for paired-end sequencing, short‑insert paired‑end reads (SIPERs) and long-insert paired-end reads (LIPERs), which also called mate pair. These two variants differ mainly between each other by the length of the insert. The SIPERs are 200‑800 bp long, while the (LIPERs) can be longer (2-5 kb) (Zeitouni *et al*., 2010). Mate pair sequencing involves generating long-insert paired-end (LIPERs) DNA libraries useful for a number of sequencing applications, including structural variant detection and identification of complex genomic rearrangements, the

technique creates libraries with much larger inserts than the standard fragment library prep, even 10Kb or more (Gillet-Markowska *et al*., 2014). Combining data generated from mate pair library sequencing with that from short-insert paired-end reads provides a powerful combination of read lengths for maximal sequencing coverage across the genome. Read pair method (mate-pair) compares the average insert size between the actual sequenced read-pairs with the expected size based on a reference genome. In paired-end sequencing, the DNA fragments are expected to have a specific distribution around insert size (Korbel *et al*., 2007). As such, the discordance between mapped paired-reads whose distances are significantly different from the predetermined average insert size is utilized by read pair to identify CNVs. Tools, which use the RP method include PEMer, Hydra, Ulysses, and BreakDancer (Pirooznia *et al*., 2015).

#### 1.6.4.2 Split Read

Split read (SR) is a method that uses reads from the pair end sequencing, where only one read of the pair has a reliable mapping and the other one either completely or partially fails to map to the genome (Zhang *et al*., 2011). The unmapped reads are a possible source of breakpoints at the single base pair level. Mapping of reads that span across a breakpoint of an SV provides the precise start and end positions of the segments that are INDEL events (Pirooznia *et al*., 2015). SR based methods, Pindel, Gustaf, SVseq2, and Prism are able to identify these breakpoints but they have limited ability to identify large-scale SVs. However, according to Jiang *et al*. (2012), Prism seems to substantially overcome this limitation by employing a modified Needleman–Wunsch alignment algorithm.

#### 1.6.4.3 Sequence read depth (SRD)

Read depth (SRD) method relies on the hypothesis that there is a correlation between depth of coverage of a genomic region and the copy number of the region (Teo *et al*., 2012). SRD methods can be in single sample, where the absolute copy number will be called, in presence of controls, where the relative copies compared to controls will be reported and in population based studies, where the overall mean of the SRD will be used to detect CNVs (Zhao *et al*., 2013). Comparing all these approaches, the exact number of CNVs can be detected the SRD as RP and SR can only report the position of the potential CNVs and not the counts. In addition, RD can work better on large size CNVs, which are difficult to detect with RP and SR (Yoon *et al*., 2009). RD method uses the following steps to estimate CNVs. First reads are aligned to a reference genome and RD will be counted using a predefined window. Then

the counts will be normalized in order to remove potential biases, mainly due to GC content and repeat regions (Pirooznia *et al*., 2015), and a segmentation algorithm will be applied to identify a contiguous set of windows having the same number of CNVs. Finally, the statistical significance of the calls will be predicted and filtering will be applied (Zhao *et al*., 2013). This method has been successfully applied to complex genomic regions containing mCNV (Sudmant *et al*., 2010). Typically, smaller CNV events and those that contain high genomic copy number require deeper sequence read depth (SRD) to achieve accurate CNV measurement. However, there is a limitation of using massively parallel short-read sequencing, which is the inability to uniquely map short reads to regions such as SDs (Sudmant *et al*., 2010).

However, in 2014 Layer *et al.* study has introduced LUMPY, which is a general probabilistic SV discovery framework that integrates multiple SV detection signals, including discordant paired-end (Mate-pair) alignments and split-read alignments. The software is depended on a general potential representation of an SV breakpoint, which enables any number of alignment signals to be integrated into a single process. In addition, LUMPY can detect SV from multiple alignment signals in files from one or more samples (Layer *et al*., 2014)

#### 1.6.4.4 Genome Structure in Populations (genome STRiP)

Genome structure in populations (genome STRiP) is a set of bioinformatics tools for discovering and genotyping structural variations using sequencing data. The methods are designed to detect shared variation using sequence data that are distributed across hundreds or thousands of genomes (Handsaker *et al*., 2011). Genome STRiP looks both across and within a set of sequenced genomes to detect variation and in order to run discovery or genotyping on a single sequenced genome or a small set of genomes, running the data against a background population, such as a set of genomes from the 1000 Genomes Project is required. The useful thing is the background population does not need to be matched to the target individuals. This method can be used for the discovery of novel structural variations or to genotype known variants in new samples (Broad Institute, 2015). The analytical framework for analysing Genome Structure in Populations is shown in figure 1.8 below.

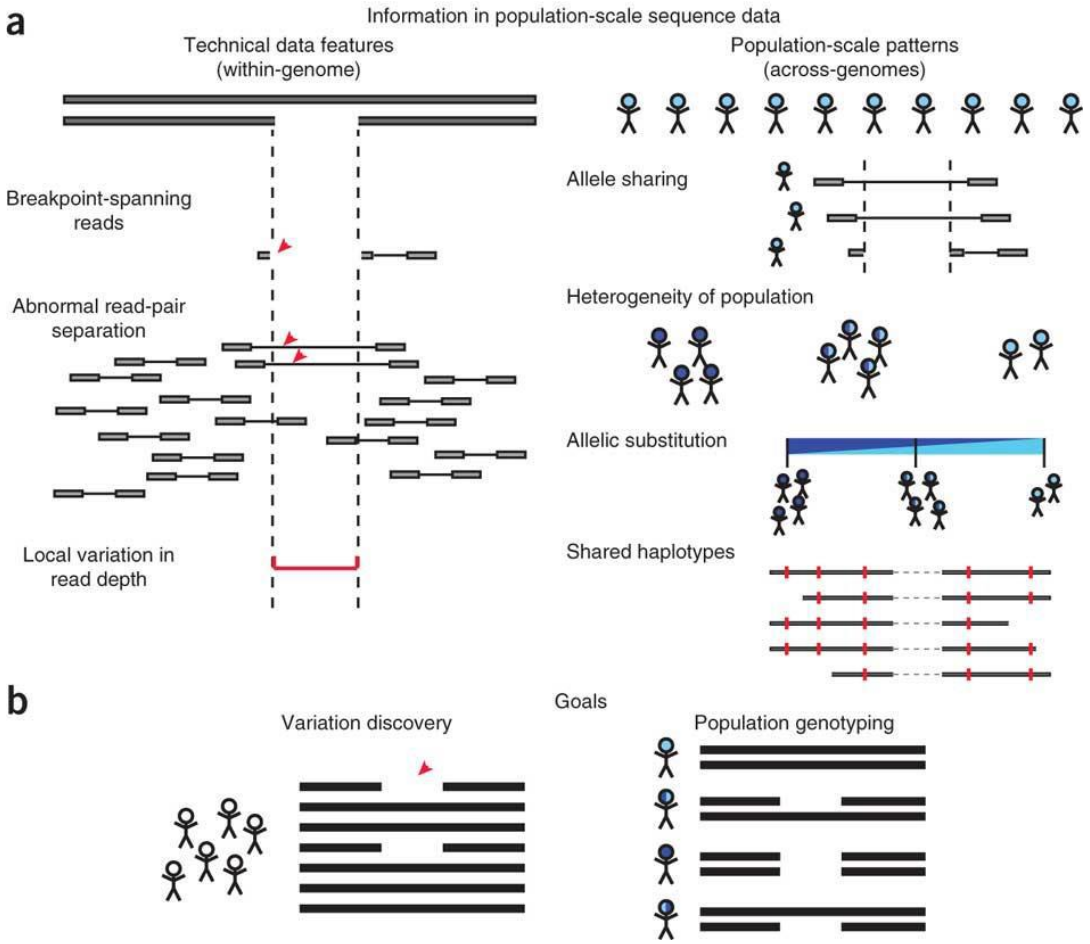



**Figure 1.8: A schematic diagram to explain the analytical framework for analysing Genome Structure in Populations. (a)** Population-scale sequence data contain two classes of information. **(b)** Goals of structural variation (SV) analysis in Genome STRiP (Handsaker *et al*., 2011).

## 1.7 Glycophorins

The glycophorins are a group of transmembrane proteins that contain oligosaccharide chains (glycans, which commonly terminate with sialic acid residues) covalently attached to polypeptide side-chains and the major integral red blood cell membrane proteins that carry a considerable portion of the sialic acid (monosaccharides) present in the extracellular domain (Wassmer and Carlton, 2016). About 10% of total human red cell transmembrane proteins are glycophorins; their sialic acids (carbohydrates) are implicated in the majority of negative charges on the erythrocyte surface. However, erythrocyte carbohydrates on the cell surface might be involved in the stability of the red cell and in the blood group system (Tayyab and Qasim, 1988). Glycophorins are heavily glycosylated N-terminal extracellular domains and C-terminal cytoplasmic domains of different sizes. In the intramembrane domain of the red cell, glycophorins have a single helical segment of a number of hydrophobic amino acid residues (Cartron *et al*., 1993). Glycophorins on the mammalian erythrocyte surface are used by pathogens as major receptors for invasion. In addition, because of the variable density and

diversity of glycophorin structures that are expressed on the red cell surface, their primary function is unclear (Baum *et al*., 2002).

A sialic acid-dependent stain can be applied to human red cell membranes to be separated by SDS-PAGE, such as PAS stain. Four polypeptides will result from the separation to confirm the presence of Glycophorin A (GPA), Glycophorin B (GPB) and Glycophorin C (GPC) and Glycophorin D (GPD). However, GPA and GPB are from homodimeric and heterodimeric complexes that resist separation by SDS. In addition, a further gene (Glycophorin E) *GYPE* has been identified, but the protein product has not been confidently demonstrated in the red cell (Tanner, 1993). Despite glycophorins being similar in their topology, products of glycophorins can be divided into two groups, the products of *GYPA*, *GYPB* and *GYPE* genes family, which are very close together, and the GPC and GPD proteins are products from the same glycophorin gene (*GYPC*), which are unrelated to each other. Other names that could indicate these glycophorins are shown in table 1.1 (Cartron and Rahuel, 1992).

**Table 1.1: A table shows other possible names for some glycophorin proteins.**

| Common name | Other names |
|---|---|
| **GPA** | SGPα and MN sialoglycoprotein |
| **GPB** | SGPδ and Ss sialoglycoprotein |
| **GPC** | SGPβ and sialoglycoprotein D |

The *GYPA*, *GYPB*, and *GYPE* paralogous genes family are located in tandem on chromosome 4q28-31. They are in a tandem fashion in a gene cluster (Figure 1.9) (Rearden *et al*., 1993).

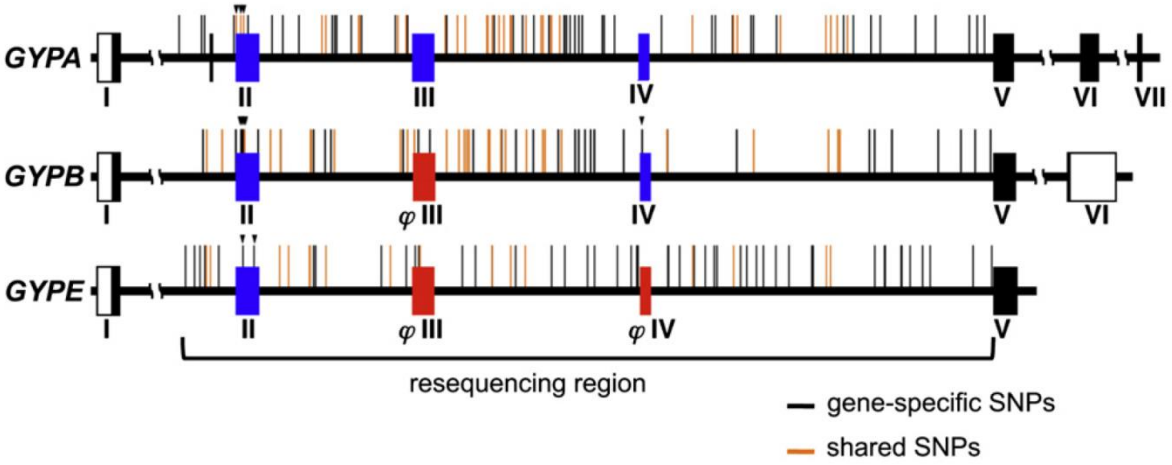



**Figure 1.9: Molecular bases of glycophorin genes.** All exons in the *GYPA* will be translated to mRNA. *GYPA* is about 31kb in length. However, exon III in *GYPB* is a pseudoexon (untranslated to mRNA). *GYPB* consists approximately 23kb in total. The last glycophorin gene is *GYPE*, which contains two pseudoexons (III and IV) and it has 30kb as a total sequence. The *GYPE* gene is more similar to the *GYPB* gene than it is to the *GYPA* gene. The figure highlights the most important exons, where the breakpoints often occur between these three genes. Obtained figure from Ko *et al*. (2011).

### 1.7.1 Glycophorin A (GYPA)

According to the UCSC Genome Browser (GRCh37/hg19) assembly, the *GYPA* gene is located on the long arm of chromosome 4 (145,030,456-145,061,904) (Kent *et al*., 2002). Glycophorin A protein (GPA) is the main and the most abundant sialoglycoprotein of red cells. It has a molecular weight of 36 KDa including its carbohydrate, whereas, according to its molecular sequence, it has 14 KDa without the carbohydrate. The ratio of GPA protein to total membrane protein in the erythrocyte is 1.6% (W/W) (Aoki, 2017). GPA was one of the first transmembrane protein sequences to be available and many studies have determined its amino acid sequence; it contains 131 amino acid residues without disulphide bonds (Denomme, 2004). Northern blotting and immunochemical studies claim that GPA is only expressed on red cells and not expressed in other cells (Tanner, 1993). However, despite that GPA protein is mainly expressed on the erythrocyte surfaces, it also can be found in the bone marrow, spleen and kidney as shown in the human protein atlas (https://www.proteinatlas.org/search/GYPA) and also in the liver (Kim *et al*., 2014).

Seven exons are present in *GYPA* gene (Figure 1.9): exon 1 yields a leader peptide, while exons 2-4 are responsible for encoding extracellular domain. The transmembrane domain is encoded by exon5, whereas exons 6 and 7 encodes the cytosolic domain and the 3'-untranslated region (Denomme, 2004).

The primary structure of GPA can be divided into three different segments: extracellular domain, cytoplasmic domain, and intermembrane domain (Figure 1.10). Among different mammals, the amino acid sequence of GPA shows significant homology in intermediate and C-terminal domains, whereas there is no homology in the N-terminal amino acids sequence (Denomme, 2004). This is possible because the hydrophobic domain of GPA is more conserved than the N-terminal segment, which acts as a receptor for many ligands. The hydrophobic segment (intermediate segment) exists in α-helical as a secondary structure for 27% of the studies. However, 10% indicate that N-terminal has a β structure, while 63% of unordered structures have referred predominantly to N-terminal and C-terminal domains (Aoki, 2017).

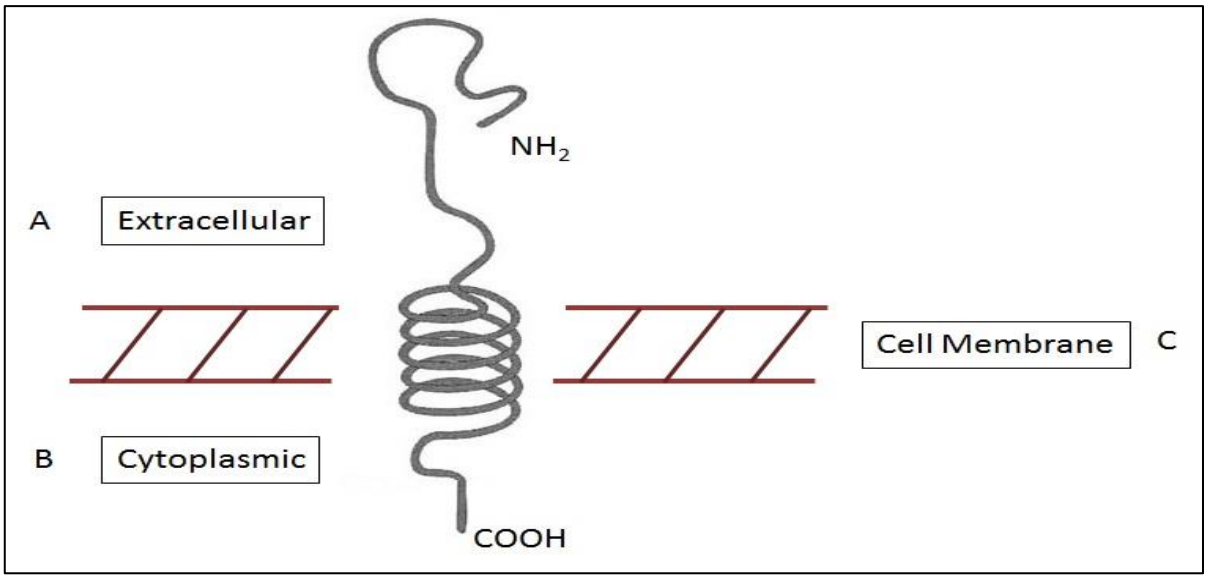



**Figure 1.10: Primary structure of GPA protein. A:** extracellular domain (N-terminal segment), which is on the cell surface, contains all carbohydrates that are usually found in different glycophorins. It contains a significant amount of serine and threonine that link with N-terminal's carbohydrates by O-glycosidic bonds. However, a large portion of extracellular domain is absent in GPB. **B:** cytoplasmic domain (C- terminal segment), which is in the cytoplasm of the red cell. It contains a high level of acidic amino acids, mostly proline. Nevertheless, GPB lacks any substance for C-terminal domain. **C:** intermembrane domain (intermediate segment) connects N-terminal with C-terminal and it passes the lipid bilayer. Hydrophobic amino acids are the content of the intermediate segment.

The exact function of GPA protein is still unclear. However, the intensive negative charge of the red cell surface that is provided from the sugar groups might involve preventing adhesion between red cells and vascular surfaces. Also GPA is implicated in affecting blood viscosity. It is expected that sialic acids organise interaction among red cells or between red cells and endothelial cells. Therefore, if sialic acids are removed from red blood cells, they will be clumped. Nevertheless, some have claimed that GPA might be involved in cell ageing (Blumenfeld and Huang, 1997). In addition, GPA's carbohydrates have been implicated in MN-antigenic activity of the GPA protein (Blumenfeld and Huang, 1995).

A human carcinoma cell marker ($T_n$ antigen) has been exposed when GPA has missed some of the carbohydrates as a result of changes in the *GYPA* gene (deletions in the CN). It passes some essential metals ions to the red blood cells membrane such as $Mg^{++}$ and $Ca^{++}$ (Tayyab and Qasim, 1988). Much evidence has shown that there is an association between GPA and Band-3, which is also a transmembrane protein. It is predicted that $Wr^b$ antigen that is located on GPA is formed by the association between these two proteins. A study has been conducted using co-injection of both *GYPA* and Band-3 mRNAs into oocytes of a *Xenopus*, and the result has shown that the movement of newly synthesised Band-3 has been enhanced by GPA

(Young and Tanner, 2003). However, GPA is not essential for this translocation since Band-3 has moved to the red cell surface in the time of absence of GPA, but slower. Red cells with deficient GPA [En($a^-$)] type exhibit an increase in glycosylation of Band-3, which may be due to the addition of excessive sialic acid, which is usually presented on the GPA molecule (Sim *et al*., 1994). GPA acts as a receptor for several viruses such as encephalomyocarditis and influenza viruses (Denomme, 2004), which could lead to many red cell diseases such as leukopenia and thrombopenia (Aoki, 2017). It is also a receptor for erythrocyte invasion by a malaria parasite (*P.falciparum*) (Baum *et al*., 2002), it will be explained later (section 1.11).

### 1.7.2 Glycophorin B (*GYPB*)

Glycophorin B protein (GPB) is a sialoglycoprotein and similar to GPA except for its exoplasmic domains and cytoplasmic tails (Tanner, 1993). It is a type1 transmembrane protein. The *GYPB* gene yields a protein with 72 amino acids in length, determined by protein and cDNA sequence methods. The *GYPB* gene is located on the long arm of chromosome 4 (144,917,257-144,940,496) according the UCSC Genome Browser (GRCh37/hg19) assembly (Kent *et al*., 2002). The molecular weight for GPB protein has been calculated twice, with its carbohydrates and without the carbohydrates, and this is 20 KDa Daltons and 8 KDa respectively (Aoki, 2017).

Although the *GYPB* gene consists of 6 exons, one of them, exon 3, is a pseudoexon. Exon 1 encodes a leader peptide, while exons 2 and 4 encode the extracellular domain (Figure 1.9). In addition, exon 5 encodes the transmembrane and short cytosolic segment, whereas the 3'-untranslated region will be generated by the last exon. A large portion of the extracellular domain is absent in GPB compared with GPA's extracellular domain. Furthermore, GPB lacks any substance for the C-terminal domain (Figure 1.10). While GPA has an N-glycan chain at position 26 (Asparagine), it is absent in GPB due to exon 3 splicing out. However, GPB has 11 serine-threonine linked oligosaccharide chains. In addition, the *GYPB* gene differs on the 3' side of *Alu* repetitive sequence, which is present in both genes (Figure 1.11) (Tanner, 1993). GPB protein also acts as a *P.falciparum* parasite receptor (Mayer *et al*., 2009), it will be explained later (section 1.11). GPB protein is expressed on the erythrocyte surfaces, GPB RNA can be found in the bone marrow, spleen and kidney as shown in the human protein atlas (https://www.proteinatlas.org/search/GYPB).

### 1.7.3 Glycophorin E (*GYPE*)

Glycophorin E (*GYPE*) is the third gene of the glycophorin family and is located on the same chromosome, (chr4:144,792,019-144,826,716) according the UCSC Genome Browser (GRCh37/hg19) assembly (Kent *et al*., 2002). GPE protein has not been detected on RBCs till now, however, at transcript level, the human protein atlas shows that GPE RNA found in the bone marrow and spleen tissues (https://www.proteinatlas.org/search/GYPE). The expectation of the GPE protein was to be similar to GPA and GPB in that it contains cleavable leader peptides. It is predicted that the *GYPE* gene comprises 6 exons and two of them are pseudoexons (3 and 4) (Figure 1.9). Like *GYPA* and *GYPB*, exon 1 encodes GPE's expected leader peptide and exon 2 encodes extracellular domain. However, exon 5 encodes the predicted transmembrane segment and exon 6 produces the 3'-untranslated region (Figure 1.11) (Denomme, 2004).

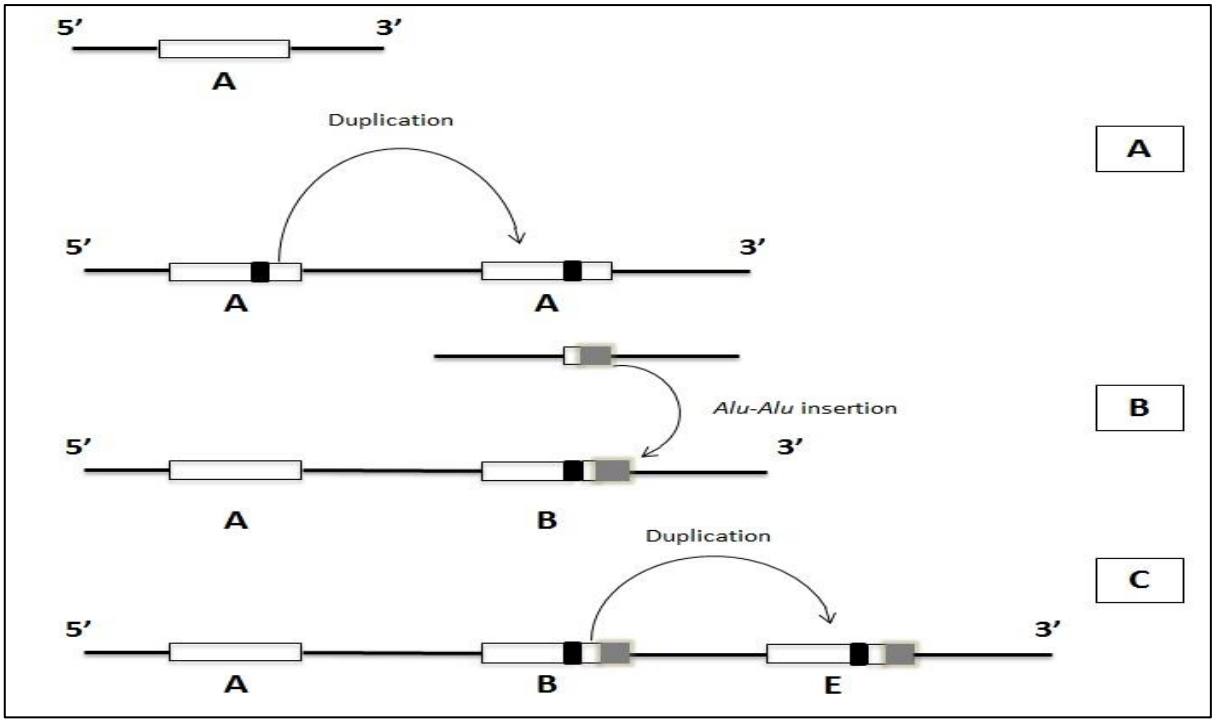



**Figure 1.11: Evolution of Glycophorin A, B and E genes.** (**A):** shows the *GYPA* gene and how it duplicates to produce *GYPB*. (**B):** an insertion of *Alu* repeats element to the duplicated *GYPA* with some deletions. (**C):** *GYPB* has duplicated with some deletions and insertion to generate *GYPE*. Each gene is over 30 kb and they are all very similar to each other in DNA sequence. The *GYPB* gene is almost identical to *GYPA* gene by the first five exons, but differs in the intronic *Alu* repetitive sequence, which is common to both genes, thus GPB lacks cytoplasmic domain. On the other hand, *GYPE* is more similar to the *GYPB* than it is to the *GYPA*. *GYPE* lacks a DNA sequence that codes the amino acid residues 27-39 of GPB, so technically it is about 40 bases shorter. These amino acids are responsible for Ss antigens in GPB. In addition, *GYPE* contains a DNA sequence insertion at exon 5 compared with *GYPA*, which leads to extra amino acids that are not present in GPA and GPB (Original diagram using information from Tanner, 1993).

## 1.8 Variations of glycophorins and antigens (MNS Variants)

MNS, the second blood group system discovered, is probably second only to Rh in its complexity. Many of the serological intricacies of MNS are now understood at the molecular level. MN is polymorphic in all populations tested: the frequencies of the common phenotypes in white people are M+N– 28%, M+N+ 50%, and M–N+ 22%. Phenotype frequencies in white people are as follows: S+s– 11%, S+s+ 44%, and S–s+ 45% (Daniels, 2013). M and N determinants are carried on glycophorin A (GPA), the major red cell sialic acid-rich glycoprotein (sialoglycoprotein, SGP). M differs from N in the amino acid composition of the extracellular tip of GPA: M has serine at position 1 and glycine at position 5; N has leucine at position 1 and glutamic acid at position 5 (Table 1.2). Carbohydrate, especially sialic acid, also plays a part in the expression of M and N antigens. S and s are carried on another red cell SGP called glycophorin B (GPB) (Avent and Reid, 2000). The S/s distinction arises from a single amino acid substitution at position 29 of GPB: S has methionine and s has threonine at position 29 (Table 1.2). The first 26 amino acid residues from the extracellular terminus of GPB are identical to those of N-active GPA ($GPA^N$). Consequently, GPB also demonstrates N activity (often referred to as 'N'), which is detected on the red cells of homozygous *M/M* individuals by some anti-N (Daniels, 2013; Reid, 2009). (GPA on the intact RBCs is trypsin sensitive and α-chymotrypsin resistant, but GPB is resistant to trypsin cleavage and sensitive to α-chymotrypsin).

**Table 1.2: Amino acids associated with M, N, S, and s antigens of GPA and GPB.**

| Antigen | Gene | Glycophorin | Amino acid polymorphisms |
|---|---|---|---|
| M | *GYPA* | GPA | 1 Ser, 5 Gly |
| N | *GYPA* | GPA | 1 Leu, 5 Glu |
| S | *GYPB* | GPB | 29 Met |
| s | *GYPB* | GPB | 29 Thr |

However, nowadays, there are 41 antigens related to the MNS blood group system that have been discovered and the terminology system has been changed to GP (for glycophorin) plus the abbreviated name of the person in whom the phenotype was first described. These low prevalence antigens are generated from a single amino acid change or as a result of hybrid GPA/GPB protein (Reid, 2009). A number of these antigens in the MNS system were grouped together as a subsystem called Miltenberger (Mi) (Dahr, 1992). After the Miltenberger subsystem reached the $11^{th}$ class or antigen, the MNS system became more complex, and because of this the suggestion was made to use GP for glycophorin plus an abbreviation of the person for whom the antigen was first described (Lomas-Francis, 2011).

For example, GP.Bun phenotype was called Mi.VI with the obsolete Miltenberger nomenclature (Reid, 2009). These variant phenotypes could be revealed by altered M, N, S, and s antigens. The incidence of MNS system variations varies dependent on the ethnic or geographic origin (Blumenfeld and Huang, 1997). For example, the variants have low incidence among western Europeans but high incidence among African or Asian populations (Table 1.3) (Blumenfeld and Huang, 1995).

**Table 1.3: MNS blood group system phenotype prevalence. A table shows the spread of variations of MNS blood group system according to ethnic or geographic regions.**

| Phenotype | prevalence |
|---|---|
| GP.Hil (Mi.V) | High in European. 1 in 2,000 in Swiss blood donors. One Taiwanese and a homozygote Spanish-American woman. |
| GP.JL(Mi.XI) | Europeans, people from southern China and in Hispanics. |
| GP.Vw (Mi.I) | South-eastern Switzerland — found a prevalence of 1.43%. |
| GP.Nob (Mi.VII) | Bristol, England (white donors), — found a prevalence of 0.06%. |
| GP.Dane (Mi.IX) | 0.43% in Danes [24]. |
| GP.Mur (Mi.III) | 1- Rare in Caucasians.<br>2- Common in southeast Asia:<br>• 9.6% in Thais, 5% Chinese populations.<br>• 7.3% in the general Taiwanese population.<br>Up to 88% in some indigenous Taiwanese tribes. |
| GP.Hop (Mi.IV) | Two probands have been described with GP.Hop, both Caucasians. |
| GP.Bun (Mi.VI) | Rare in Caucasians. |

The high level of homology and organization of glycophorin genes along the chromosome is implicated in the occurrence of some rearrangements such as unequal crossover and gene conversion among glycophorin genes. This leads to raise the antigenic variants of MNS system. Therefore, according to the rearrangement, the blood group phenotype will be expressed (Table 1.4) (Reid, 2009).

The NAHR between homologous genes (*GYPA* and *GYPB*) can occur once the chromosome misaligns during meiosis. This will produce two types; the first type is called Lepore type, which is the chromosome that carries a hybrid *GYPA-B* gene upstream of *GYPE*, with neither *GYPA* nor *GYPB* genes. However, the Second type is called anti-Lepore type, which contains the reciprocal chromosome that carries a hybrid *GYPB-A* gene and normal *GYPA* and *GYPB* genes (Figure 1.12). Occasionally, double crossover could occur in these homologous genes and result with a chromosome carrying hybrid *GYPA-B-A* gene and normal *GYPB* gene, whereas, the reciprocal chromosome will carry hybrid *GYP(B-A-B)* gene and normal *GYPA* gene (Table 1.4) (Lomas-Francis, 2011).

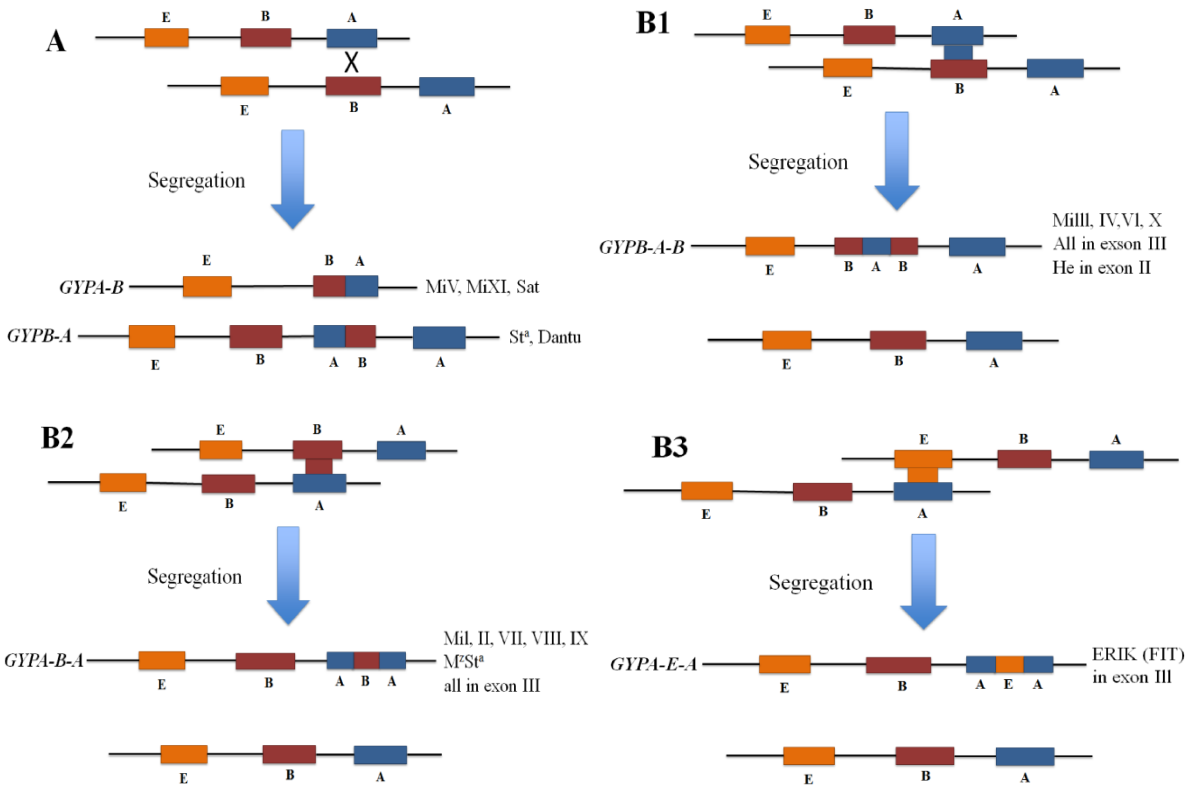



**Figure 1.12: Schematic representation of possible gene rearrangements at meiosis, resulting in glycophorin hybrid alleles. (A)** Unequal crossover; **(B1-3)** Gene conversion possibilities. The three genes within the glycophorin family cluster on the long arm of chromosome 4 are shown: A = *GYPA*; B = *GYPB*; E = *GYPE*. Blue, red and orange boxes indicate to *GYPA*, *GYPB* and *GYPE* respectively.

Gene conversion usually occurs during the DNA repair between homologous genes. This mechanism does not produce a reciprocal chromosome but it reflects a directional transfer of nucleotides from one duplex to another. That will be during meiosis if the paralogues homologous genes have misaligned. Therefore, nucleotides from a gene can be inserted into one strand of DNA of another gene and a hybrid gene will be generated (Reid, 2009) (Figure 1.11). Once this mechanism occurs to glycophorin genes, nucleotides from the *GYPA* gene will be inserted into the *GYPB* gene to generate a hybrid *GYPB-A-B* gene with normal *GYPA* gene. In contrast, if the *GYPB* gene is the donor and *GYPA* gene is the recipient, the product will be a hybrid *GYPA-B-A* gene with normal *GYPB* gene (in Cis) (Table 1.5). Due to gene conversion pseudoexon 3 in *GYPB* could be expressed in some hybrid genes and a novel GP protein sequence will be generated with low incidence antigens of the MNS system (Palacajornsuk, 2006). In a few incidences *GYPE* could be involved in recombination events and result in different hybrid genes and phenotypes such as GP.Mur (Table 1.6).

**Table 1.4: The classification of hybrid glycophorins of the obsolete miltenberger subsystem phenotypes, associated low-prevalence antigens and glycophorin hybrid alleles.**

| **Phenotype** | **GP.Hil (Mi.V)** | **GP.JL (Mi.XI)** | **GP.Vw (Mi.I)** | **GP.Hut (Mi.II)** | **GP.Nob (Mi.VII)** | **GP.Joh (Mi.VIII)** | **GP.Dane (Mi.IX)** | **GP.Mur (Mi.III)** | **GP.Hop (Mi.IV)** | **GP.Bun (Mi.VI)** | **GP.HF (Mi.X)** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Associated Antigens** | Hill and MINY | TSEN and MINY | Mia + Vw | Mia + Hut + MUT | Nob | Hop + Nob | DANE + Mur + a trypsin-resistance M antigen. | Mia + Mur + Hil + MINY + MUT | Mia + Mur + TSEN + MINY + MUT + Hop | Mia + Mur + Hil + MINY + MUT + Hop | Mia + Hil + MINY + MUT |
| **Blood group expressed** | Elevated (s) Weak (M or N) | Altered (S) Weak (M) | Mostly (Ns), Occasionally (NS) and Found once with (MS) | Travels equally with MS or Ns | Associated with Ms and MS | Travelled with Ns within two known families | Often travelled with MS. Rarely found with Ms and in trans with Mk | Always with s. In European ancestry ether Ms or Ns. However, in Thais and Chinese it travel with Ms | Always with S generally travel with M | Always with s and generally travel with M | Presence of M and elevated N. Unusually strong s antigen |
| ***GYP* allele** | *GYP(A-B)* hybrid | *GYP(A-B)* hybrid | *GYP(A-ψB-A)* | *GYP(A-ψB-A)* | *GYP(A-ψB-A)* | *GYP(A-ψB-A)* | *GYP(A-ψB-A)* | *GYP(B-A-B)* | *GYP(B-A-B)* | *GYP(B-A-B)* | *GYP(B-A-B)* |
| **GP protein encoded** | GP(A-B) | GP(A-B) | GP(A-B-A) | GP(A-B-A) | GP(A-B-A) | GP(A-B-A) | GP(A-B-A) | GP(B-A-B) | GP(B-A-B) | GP(B-A-B) | GP(B-A-B) |
| **Mechanism** | Unequal Crossover | Unequal Crossover | Gene Conversion | Gene Conversion | Gene Conversion | Gene Conversion | Gene Conversion | Gene Conversion | Gene Conversion | Gene Conversion | Gene Conversion |
| **Breakpoint** | 5’ end of intron3 of *GYPA* | 3’ end of intron3 of *GYPA* + 7 nucleotides of exon4 of *GYPB* | p.Thr47Met | p.Thr47Lys | In the DNA sequence 68 the Arg changed to Thr and in 71 the Tyr changed to Ser | In the DNA sequence 68 the Arg changed to Thr | In the DNA sequence 65 Ile changed to Asn in position 64 due to an Adelyn nucleotide change | The protein sequence position 57 is Arg. 55 bp insert from exon 3 of *GYPA* | 131 bp insert from exon 3 of *GYPA* | The protein sequence position 57 is Thr. 131 bp insert from exon 3 of *GYPA* | 98 bp insert from exon 3 of *GYPA* |

In GP.Hil and GP.JL, the chromosome that carries a hybrid *GYPA-B* gene downstream of *GYPE* has neither entire *GYPA* nor *GYPB* genes (Table 1.4). However, the reciprocal chromosome of the GP.Hil and GP.JL, carries a hybrid *GYPB-A* gene with normal *GYPA* and *GYPB* genes, which leads for example to GP.Dantu and Sta (not illustrated in table 1.4). In GP.Vw, GP.Hut, GP.Nob, GP.Joh and GP.Dane, different small portions (from 1 to 16 bp) substituted from the exon 3 of *GYPA* by the same number of nucleotides of pseudoexon 3 of *GYPB*; therefore, a part of the pseudoexon is translated. For instance, in GP.Dane a short amino acid sequence encoded by pseudoexon of *GYPB* (54Pro-Ala-His-Thr-Ala-Asn59) (Table 1.4). Both GP.Mur and GP.Bun are encoded by a *GYPB* allele that carries s antigen but differs in the length of the translated *GYPB*. This segment contains a portion of pseudoexon 3 and a portion of intron 3 of *GYPB* gene (Table 1.4). However, GP.Hop is identical to GP.Bun except that GP.Hop is encoded with S rather than s in GP.Bun (Table 1.4). GP.HF's hybrid gene encodes an amino acids sequence differing from GP.Mur and GP.Bun (Table 1.4) by five amino acid residues and six amino acid residues respectively (Blumenfeld and Huang, 1997; Lomas-Francis, 2011; Reid, 2009; Huang and Blumenfeld, 1988 and 1991; Storry *et al*., 2000 and Velliquette *et al*., 2008).

**Table 1.5: Mechanisms giving rise to variant glycophorin genes and their in cis partner genes.**

| Mechanism | Glycophorin variant | Gene(s) in cis |
|---|---|---|
| Single crossover (Lepore) | *GYP(A-B)* | No *GYPA* or *GYPB* |
| Single crossover (anti-Lepore) | *GYP(B-A)* | *GYPA* and *GYPB* |
| Gene conversion | *GYP(A-B-A)* and *GYP(A-ψB-A)* | *GYPB* |
| Gene conversion | *GYP(B-A-B)* | *GYPA* |

Depending on the breakpoint of a hybrid glycophorin gene the antigen/s will be expressed and thus the phenotype will be recognised. For example, although both Gp.Hil and GP.JL are encoded by a hybrid *GYPA-B* gene, the breakpoint of GP.Hil's hybrid gene locates at the 5' end of intron 3 of *GYPA*, while the junction in GP.JL's hybrid gene combines the 3' end and seven bases of exon 4 of *GYPB* gene and each phenotype has different antigens (Table 1.4). There are more glycophorin phenotypes that have been recorded after the Miltenberger Subsystem (Table 1.6).

**Table 1.6: MNS system hybrid alleles, encoded glycophorins (GPs), phenotypes, and associated low-prevalence antigens after the obsolete Miltenberger subsystem.**

| Phenotype | Associated antigens | Glycophorin Allele | Glycophorin protein encoded | Mechanism |
|---|---|---|---|---|
| GP.TK | SAT | *GYP(A-B)* | GP(A–B) | Unequal Crossover |
| GP.KI | Hil | *GYP(A-B-A)* | GP(A–B–A) | Gene Conversion |
| GP.SAT | SAT | *GYP(A-B-A)* | GP(A–B–A) | Gene Conversion |
| GP.M$^{g}$ | M$^{g}$ | *GYP(A-B-A)* | GP(A–B–A) | Gene Conversion |
| GP.Zan (M$^{z}$) | St$^{a}$ | *GYP(A-ψB-A)* | GP(A–A) | Gene Conversion |
| GP.Mar | St$^{a}$ | *GYP(A-ψE-A)* | GP(A–A) | Gene Conversion |
| GP.He | He | *GYP(B-A-B)* | GP(A–B) | Gene Conversion |
| GP.Dantu | Dantu | *GYP(B-A)* | GP(B–A) | Unequal Crossover |

En(a-), S-s-U- and M$^{k}$ phenotypes are caused by deletion variants of glycophorin, which lead to the absences of GPA, GPB, or both of them from the RBCs (Table 1.7). The En(a-) indicates the absence of GPA, whereas, RBCs without GPB will express (S-s- U-). The absence of both GPA and GPB will result in RBCs with M$^{k}$/M$^{k}$ phenotype, which leads to a resistant to malaria infection (Blumenfeld and Huang, 1997).

**Table 1.7: Molecular basis of null phenotype RBCs (table obtained from Reid, 2009).**

| Null phenotype RBCs | Molecular basis |
|---|---|
| En(a-) Fin | Deletion of *GYPA* (exon 2 to 7) and *GYPB* (exon 1) |
| S–s–U– (deletion type) | Deletion of *GYPB* (exon 2 to 6) and *GYPE* (exon 1) |
| M$^{k}$/M$^{k}$ | Deletion of *GYPA* (exon 2 to 7), *GYPB* (exon 1 to 6) and *GYPE* (exon 1) |

RBCs with variable glycophorins can be detected in many ways. The M, N, S, and s antigens are expressed stronger or weaker than controls, and they could be sensitive or resistant to the treatment by different enzymes. For example, the M antigen on GP.Dane is considered as a trypsin resistant. In addition, Mi$^{a}$ has different sensitivities to some enzymes when it is expressed in GP.Vw and GP.Hut than when it is expressed in GP.Mur and GP.Hop (Lomas-Francis, 2011). The mechanisms of the allelic diversity that occur on the glycophorin genes family are very different from other human genes and multigenic families such as the haemoglobin gene cluster that varies due to point mutations rather than recombination events (Blumenfeld and Huang, 1997). However, the Rh blood group system has 50 antigens that depend on the variations in the *RHD* and *RHCE* genes, which have explicit examples in gene recombination such as crossing over and gene conversion, similar to the glycophorin family (Rouillac *et al*., 1995).

The function of glycophorin/MMiltenberger appeared to be a complementary function. However, there is a significant consequence of the presence of GP.Mur, which could increase RBCs resistance to *P. falciparum* (Lomas-Francis, 2011). In addition, it can make the RBCs more resistant to osmotic stress by upregulating the Band 3 proteins amount on the RBC surface (Hsu *et al*., 2009).

The glycophorin polymorphisms (MNS variants) have rarely been implicated in transfusion reactions and haemolytic disease of the fetal and newborn (HDFN), which have been reported for anti S-, s- and U–. Allo-anti-En(-α) has been reported for haemolytic transfusion reactions (Postoway *et al*., 1988), whereas, auto-anti-En(-α) has been implicated in severe and fatal autoimmune haemolytic anaemia (Daniels, 2013; Pavone *et al*., 1981). Furthermore, there is a clinical relevance between antibodies and many low-prevalence MNS antigens such as, anti-$Mi^a$, -Vw, -Hil, -Hut and –Mur. They have been reported with HDFN and sometimes could be more severe. For illustration, because of incompatibility, Miltenberger blood variant may cause a severe haemolytic disease in transfusion. In Taiwan, the blood bank has been obligated to screen 5-8 Miltenberger phenotypes before any transfusion (Hsu *et al*., 2009). Therefore, the CNVs of glycophorin genes either deletion or duplication are presenting different blood groups, which may affect the susceptibility to some diseases.

Two different previous studies have shown the adaptive evolution of these glycophorin genes through primates and balancing selection in *GYPA* (exon 2) in a population from West Africa. Baum *et al*. (2002) have suggested that GPA works as a decoy for pathogens, while Wang *et al*. (2003) have reported that the GPA variations were implicated in evading pathogen invasion. An additional study has been applied to individuals from 15 different African populations with a variety of malaria endemicity levels by sequencing 3.7Kb of *GYP* genes. The study indicates that there are huge genetic variations in glycophorin genes among these populations, 1 tri-allelic in *GYPA*, 2 tri-allelic in *GYPB* and 1 tri-allelic in *GYPE*; and 83 SNPs for *GYPA*, 80 for *GYPB* and 72 for *GYPE*. Moreover, three mutations have been identified in *GYPB* (exon 2) within five high malaria exposure populations (Ko *et al*., 2011).

## 1.9 CNVs within glycophorin genes region

Glycophorin genes reside within a high polymorphic region due to segmental duplications. Therefore, NAHR and gene conversion events are very likely to be occurred in this area, which results in CNVs such as deletion or duplications (Ndila *et al*., 2018). It has been investigated that 121 polymorphisms in 70 candidate severe malaria associated genes. In addition, significant associations between risk for severe malaria and polymorphisms in 15 genes or locations, of which most were related to red blood cells (Ndila *et al*., 2018).

Study conducted by Leffler *et al*. (2017), identified 16 different CNVs within the glycophorins region. These variants were duplications, deletions of entire or part of three main glycophorin genes, *GYPA*, *GYPB*, *GYPE* as well as complex CNVs (Figure 1.13). However, *GYPB* was found to be the most affected gene by the CNV, resulting in five different forms of deletion (Figure 1.13). The study was conducted on 28 different human populations, using 3269 DNA samples. Within 17 African populations, 13 out of total 18 CNVs were found (Figure 1.14). The most common CNVs found within tested African populations was a DEL1, DEL2 resulting in complete deletion of *GYPB* gene and DUP1 with partial duplication of *GYPE* unit. The study discloses that DUP1 (partial duplication of *GYPE* unit) is the most widespread duplication variant within tested African populations. However, two most common CNVs found within tested African populations were DEL1 and DEL2, which both resulting in a deletion of the entire *GYPB* gene (Figure 1.15).

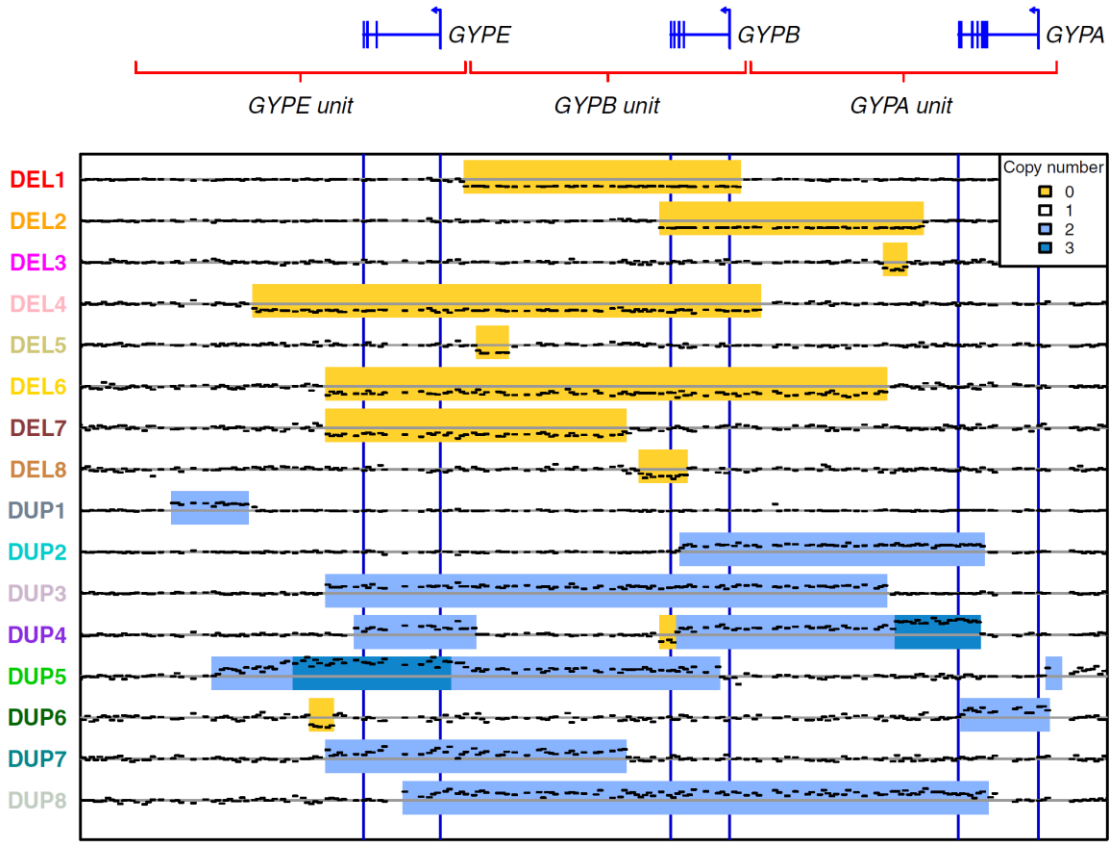
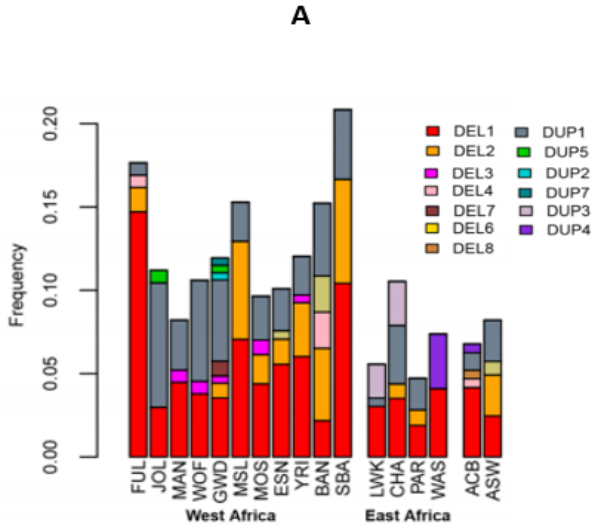



**Figure 1.13: Sixteen CNVs identified in ≥2 unrelated individuals.** A 1600 bp windows of sequence coverage presenting copy number for CNVs present in at least 2 unrelated individuals. Mean sequence coverage occurring in heterozygotes represented by black dashes. Yellow, blue and dark blue means deletion, duplication and triplication respectively. Obtained figure from Leffler *et al*. (2017).




B

| Abbreviation | Population | Country |
|---|---|---|
| FUL | Fula | The Gambia |
| JOL | Jola | The Gambia |
| MAN | Mandinka | The Gambia |
| WOF | Wollof | The Gambia |
| GWD | Gambian in Western Division | The Gambia |
| MSL | Mende | Sierra Leone |
| MOS | Mossi | Burkina Faso |
| ESN | Esan | Nigeria |
| YRI | Yoruba in Ibadan | Nigeria |
| BAN | Bantu | Cameroon |
| SBA | Semi Bantu | Cameroon |
| LWK | Luhya in Webuye | Kenya |
| CHA | Chagga | Tanzania |
| PAR | Pare | Tanzania |
| WAS | Wasambaa | Tanzania |
| ACB | African Caribbean | Barbados |
| ASW | African-American | Southwest US |

**Figure 1.14: CNVs present among 17 African populations.** The figure (A) illustrates the frequency of present CNVs among 17 African populations. The table (B) shows the abbreviations of the populations. Obtained figure from Leffler *et al*. (2017).

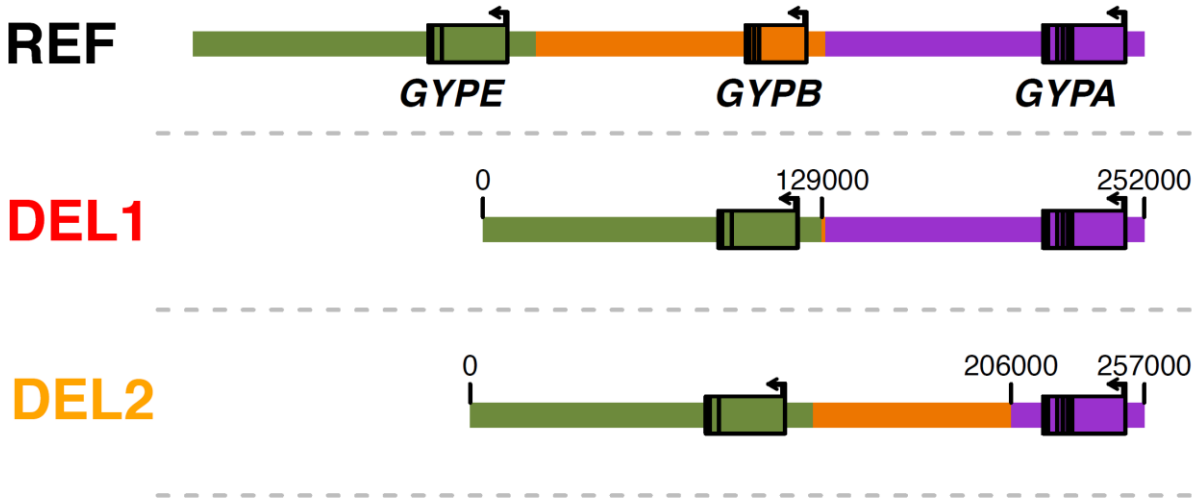



**Figure 1.15: Structural representation of DEL1 and DEL2 in comparison to the reference genome.** Modified figure from Leffler *et al.* (2017).

## 1.10 Malaria

Malaria is a very serious infectious disease that threatens human life. It continues to be a serious public health problem. It causes about one million deaths per year and a high percentage (90%) of malaria deaths occur in Africa (WHO, 2016A). A study conducted by Murray *et al.* (2012), designed to calculate the global malaria mortality rate between 1980 and 2010, took into the account each of the factors that may have an effect, such as age, sex, country and year. The study found that malaria deaths increased from 995,000 in 1980 to 1,817,000 in 2004 and then decreased to 1,238,000 in 2010. In African countries, malaria deaths increased from 493,000 in 1980 to 1,613,000 in 2004 and then decreased by around 30% to 1,133,000 in 2010. The significant increase throughout the 1980s and 1990s, rising to a peak in 2004 was attributed to increasing chloroquine resistance and first-line antimalarial drug resistance.

Since 2004, the healthcare infrastructure in sub-Saharan Africa has improved. This explains the significant decrease in malaria deaths from their global peak in 2004. Investments made by establishments such as the Global Fund to fight AIDS, Tuberculosis and Malaria in African countries where malaria is endemic have led to improvements such as insecticide-treated bed nets, artemisinin-combination treatment and vector control strategies. Insecticide-treated bed nets have been shown to be effective in reducing adult mortality, while artemisinin-combination treatment and vector control strategies are critical for addressing growing antimalarial drug resistance (Murray *et al.*, 2012).

Malaria infection during pregnancy is implicated in between 75,000 and 200,000 infant deaths each year around the world. The susceptibility to malaria and the frequency and the severity of the disease is higher in pregnant women in endemic areas compared to in other adults or non-pregnant women. This is linked to the high frequency and density of *Plasmodium falciparum*, which causes the most lethal and severe malaria. It leads to severe malaria anaemia and cerebral malaria, which are implicated in a huge number of child deaths (around 5 years old) in sub-Saharan Africa, although there are other parasites that can cause malaria infection (Ko *et al*., 2011). Generally, infection is characterized by a significant increase of cytokines, thus, tumor necrosis factor-α (TNF-α) will be increased. In *P.falciparum*, TNF-α is associated with the severity level of several symptoms in diseases such as, fever, leucocytosis, hypoglycaemia, acidosis, and cerebral malaria. However, the rosetting event is dependent on the severity of *P.falciparum* malaria infection (Uneke, 2007).

Anopheles mosquitoes are responsible for the transmission of *P.falciparum* to the erythrocytes. This parasite starts its life cycle in the red blood cells and infects them with malaria. The female mosquito will be infected during its blood meal by ingesting affected male and female gametocytes. After that, sexual reproduction will take place in the mosquito's digestive harct and the sporozoites accumulate in the mosquito's salivary gland. The next life cycle of the parasite will start once the mosquito injects the next host (human) with sporozoites. Thus, the merozoite of *P.falciparum*, which is responsible for the initial step of the RBC invasion, will be generated during the parasite life cycle by multiple forms of a sporozoite (schizogony) inside the body of the host. Consequently, merozoites will be released from schizont-infected erythrocytes and invade normal erythrocytes during the asexual erythrocytic phase of *P.falciparum* life cycle.

The prevalence of malaria in pregnant women is affected by many factors, such as maternal age, gravidity, use of prophylaxis, nutrition, host genetics, level of anti-parasite immunity, parasite genetics and transmission rates. Infants born to mothers with placental malaria are at an increased risk of anaemia, malaria infection and mortality during their first year of life (Tako *et al*., 2005). The host's response to infection may be affected by many factors, such as the intensity and seasonality of malaria transmission, the virulence of the parasite (which may depend on its genetic polymorphism), and host characteristics such as age and genetic make-up (Milet *et al*., 2010). In addition, a group of children who were exposed but non-sensitized to malaria (born to mothers with an infected placenta and a negative cord T-cell response) had a 61% greater risk of infection compared with a group of non-exposed children (negative

placenta and negative cord T-cell response), and a 41% greater risk compared with a sensitized group of children (positive placenta and cord T-cell response). Therefore, it has been shown that children exposed to malaria in the uterus gained a phenotype tolerant to blood-stage antigens which could persist during childhood (Le Port *et al*., 2012).

The World Health Organization (WHO) notes that children under five years of age (WHO, 2016A) are one of the groups most vulnerable to malaria (Figure 1.16). In 2015, there were an estimated 438,000 malaria deaths around the world in 2015, of which 69% were in children under five years of age. In addition, the World Health Organization reported that newborns and infants less than 12 months of age (WHO, 2016B) are some of the most vulnerable groups to be affected by malaria and are at risk of rapid disease progression, severe malaria and death. At the age of three months, infants become vulnerable to *P.falciparum* malaria as their acquired immunity from their mothers begins to decrease.

Malaria imposes its greatest burden on infants and young children in highly endemic areas, whereas in areas of lower transmission many cases occur in older children and adults, which means that there is a shift in cases to older ages with declining transmission (Griffin *et al*., 2014). For example, between 1996 and 2010 in southwestern Senegal and between 2003 and 2007 in western Gambia, a significant decrease in the incidence of malaria was observed, accompanied by an increase in the mean age distribution of clinical cases. Malaria cases in children under five decreased from 34% to 5% between 1996 and 2010 in Senegal, while the mean age of paediatric malaria admissions in Gambia increased from 3.9 years to 5.6 years (Griffin *et al*., 2014).

In high transmission areas, lower transmission areas and low transmission areas of malaria, the parasite prevalence in children aged two to 10 is measured at around 60%, 20% and 5% respectively. In high transmission areas of malaria, 57% of cases are predicted to occur in children less than five years of age. In the lower transmission areas of malaria, 21% of the cases are predicted to occur in children less than five years of age, with another 22% in children aged five to 10 years. In the low transmission areas of malaria, 61% of the cases are predicted to occur in children aged over 15, with another 10% in children aged less than five years (Griffin *et al*., 2014).

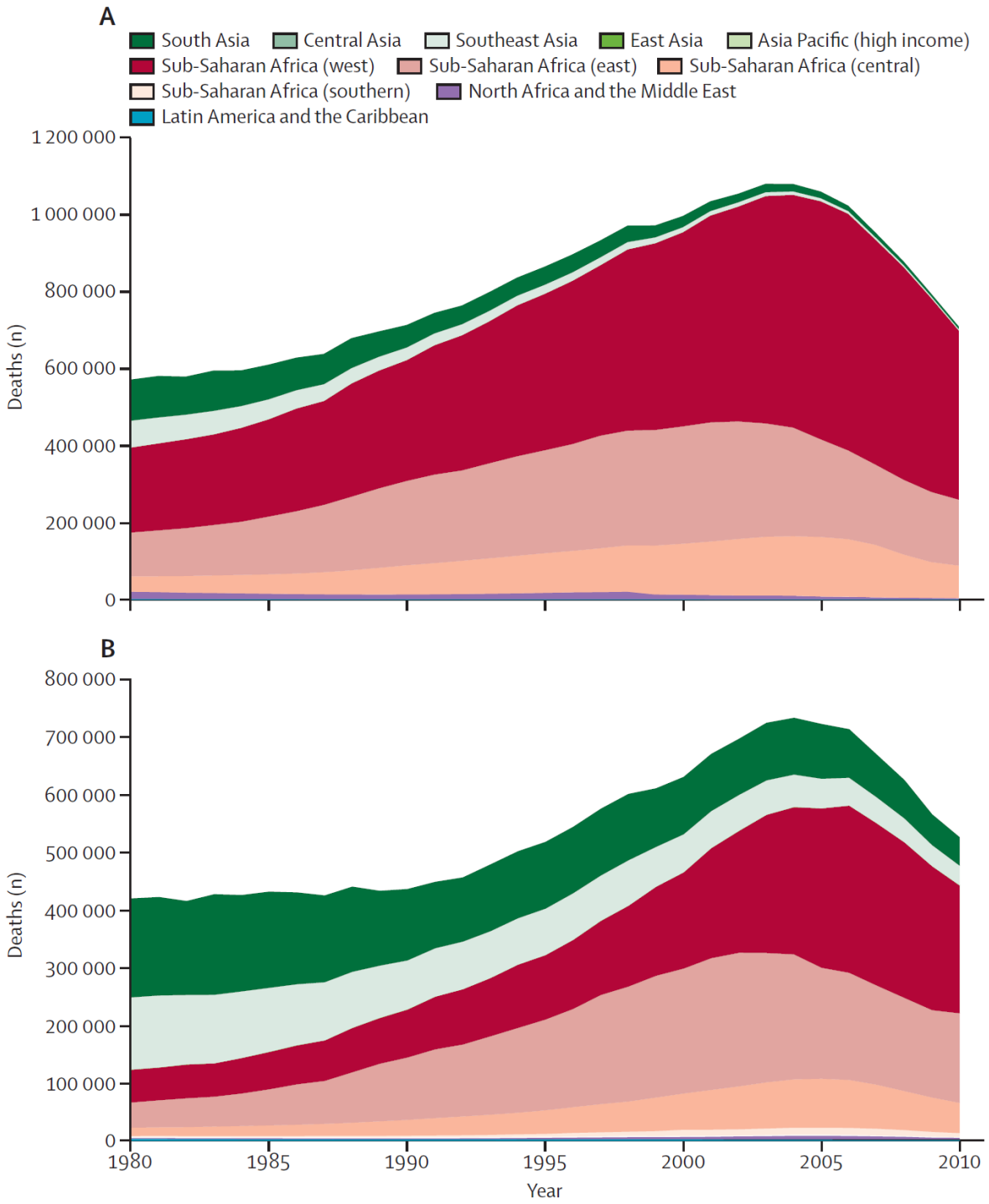



**Figure 1.16: (A) Malaria deaths by global burden of disease study region for children younger than 5 years and (B) individuals aged 5 years or older, 1980 to 2010.** The under-fives group is much smaller than the over-fives group, but has more deaths. Therefore, the chance of death in children under 5 years old is higher (infants <5 years old) but decreases as age increases (>5 years old). Obtained from Murray *et al*. (2012).

## 1.11 The role of glycophorins in malaria

The invasion of erythrocytes needs parasite ligands and erythrocyte receptors. Glycophorin proteins (GPA and GPB) are receptors for the *P.falciparum* parasite that expresses several antigens and binding ligands, which are essential for RBCs invasion. The most common of

these antigens and binding ligands are erythrocyte-binding antigen 175 (EBA-175), erythrocyte-binding antigen 140 (EBA-140), erythrocyte-binding antigen 181 (EBA-181), and erythrocyte-binding ligand 1 (EBL-1) (Wassmer and Carlton, 2016). Sim *et al*. (1994) have discovered that RBC with En(-α) phenotype has a resistance to *P.falciparum* invasion because the EBA-175 antigen cannot bind to the RBC, which means GPA is a specific receptor for one of the parasite ligands (EBA-175). On the other hand, Mayer *et al*. (2009) have shown that EBL-1 in *P.falciparum* binds to normal RBC but cannot bind with S-s-U-RBC, which means GPB is the erythrocyte receptor for *P.falciparum* EBL-1 ligand (Figure 1.17).

The merozoite carries a protein called merozoite surface protein 1 (MSP1), which coats all malaria species merozoite surfaces, where there is very abundant ligand. It has been suggested that a multi-subunit vaccine against MSP1 be used for malaria because people who are living in malaria endemic regions carry MSP1 antibodies in their blood. Therefore, a very recent study has been done *in vivo* on a genetically modified mouse model with deficient Band-3 and GPA erythrocytes. As a result, the mouse has shown full resistance to the infection of malaria *P.falciparum*. The result indicates that the GPA and Band-3 play an essential role in generating a new complex by binding with MSP1 before the invasion of the RBC by a malaria parasite beside the GPC protein (Baldwin *et al*., 2015).

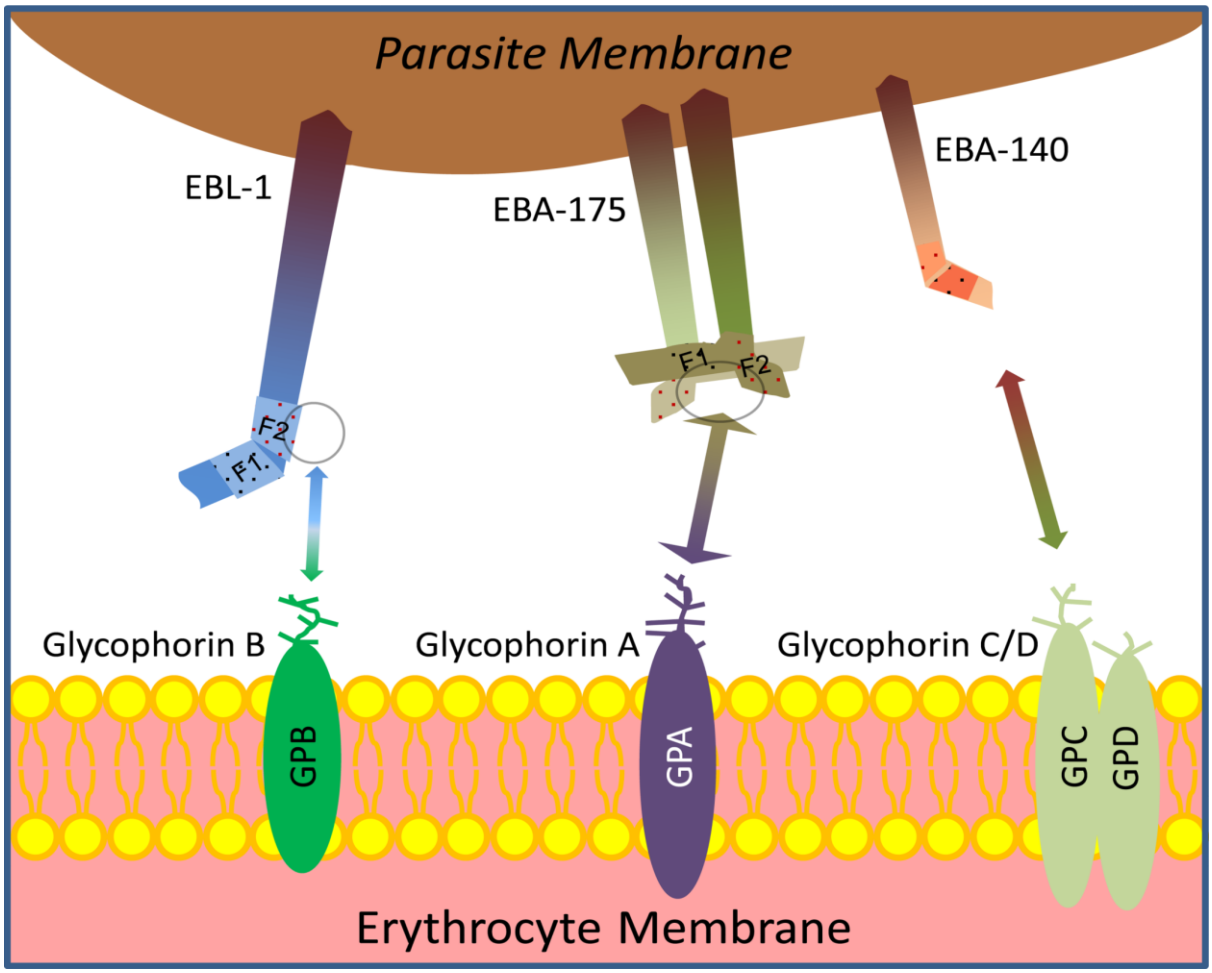



**Figure 1.17: A parasite invasion pathway of a human erythrocyte.** This figure shows red blood cell glycophorin receptors' and malaria parasite ligands' interaction during the invasion of the human erythrocytes. EBL-1, EBA-175 and EBA-140 bind to GPB, GPA GPC respectively (obtained diagram from Li *et al*., 2012).

After the discovery of MNS blood group variant antigens, it has been observed that the Southern African population have a low frequency of Dantu antigen, which is a product of the glycophorin hybrid gene and contains the GPB N-terminal and GPA C-terminal (Table 1.6). Therefore, Field *et al*. (1994) have compared the *P.falciparum* levels between RBCs with different phenotypes and control RBCs. The comparison shows that cells with Dantu and S-s-U- are lower levels of *P.falciparum* than normal RBCs, whereas, cells with Henshaw antigens show similar levels of *P.falciparum* compared with control RBCs. In addition, a study has reported that a GP.Mur (MiII) patient in South East Asia with high Band-3 expressions has a high possibility of malaria survival (Hsu *et al*., 2009).

Copy number variation in glycophorin genes may lead to a partial or full resistance to malaria parasite invasion (Reid, 2009). *P.falciparum* has many alternative pathways to invade the RBC using several different receptors. Therefore, the level of the challenge will be increased

in order to study the exact relationship between GPs and this parasite. Further studies on copy number variants of glycophorins will be necessary to determine the GPs roles and their relationship with malaria infection. Recently, a new SNP (rs186873296) of resistance to severe malaria in a region of ancient balancing selection has been demonstrated by GWAS. The SNP is close to the glycophorin gene family and is particularly located upstream of *GYPE* and downstream of *FREM3* (Malaria Genomic Epidemiology Network *et al*., 2015).

## 1.12 Relationship of CNVs in glycophorin genes to malaria resistance

Genome wide association study (GWAS) has conducted numerical studies in malaria (Band *et al*., 2013; Jallow *et al*., 2009; Timmann *et al*., 2012), and the resent study of GWAS in malaria has presented findings of severe malaria performed in a Tanzanian population (Ravenhall *et al*., 2018). However, GWAS conducted by Band, *et al*. (2015) has identified a novel locus close to the glycophorin gene cluster for severe malaria caused by *P.falciparum*.

The GWAS sample pool was primarily from African countries, and it included 5,633 children with severe malaria, 5,919 population controls and a further 13,946 case controls from sub-Saharan African countries, which were used for the replication phase giving a total sample size of 25,498 individuals. The GWAS research team genotyped all the samples and then conducted further statistical analysis by employing a Bayesian approach. This approach relies on evidence from many models of association and includes probabilities on size and similarity of the genetic effect across populations (Band *et al*., 2015). In order to obtain a statistical summary of the signal of association, the group compared the results to a null model of association. The results of the statistical analysis identified new candidate loci of association, with the strongest signal identified in the region between *FREM3* and the glycophorin gene cluster on chromosome 4. Further analysis of this specific region was conducted using the same methods as before and the results found a particular SNP (rs186873296) corresponding to the *FREM3* gene showing the strongest signal of association (Band *et al*., 2015). Although the signal did not extend fully across the glycophorin gene cluster, the receptor function of the glycophorins may help uncover the biological basis of the association.

A study was conducted by Leffler, *et al*. (2017), has hypothesised that glycophorin gene cluster copy number variation, specifically duplications, provide resistance against severe malaria. This study based its research on findings from the Band, *et al*. (2015). The Leffler *et al*. (2017) used data from GWAS in order to find further variants, which can provide better

understanding of the association signal previously identified between *FREM3* and the glycophorin gene cluster. Initially, they have created a reference panel, which consisted of 2,504 individuals from the 1000 Genomes project data, along with 765 individuals from sub-Saharan African countries (Gambia, Burkina Faso, Cameroon and Tanzania), producing a panel size of 3,269 individuals, of which 1,269 were African. Genome sequencing of the 765 individuals from sub-Saharan African countries was conducted and combined with the 1000 genomes project data set. The data was then tested for association by imputing the panel into published server malaria GWAS data where the signal of association was identified in the same region as in the study by Band *et al*. (2015).

The Leffler *et al.* (2017) has focused on read depth analysis and developed a statistical model known as a hidden Markov model, which imposed a 1600 bp window to identify glycophorin CNVs of all of the panel individuals. The result of the study has disclosed of eight deletions and eight duplications that were confirmed in two or more individuals (Figure 1.13). Variants that were identified included duplications, deletions and hybrid formations such as the hybrid of *GYPB-A* and *GYPE-A*. Further analysis was done to test the association signal, the CNV data and the sever malaria GWAS data were combined. A glycophorin variant classified as duplication 4 (DUP4) has been found to exhibit the strongest signal of association with reduced risk of malaria when compared to the other CNVs in the data set (Leffler *et al*., 2017). Odds ratio was calculated for DUP4 as well and it was equalled OR= 0.59, the odds ratio (OR) refers to "the odds that an outcome will occur given a particular exposure" (Szumilas, 2010). The DUP4 can be described as a *GYPB-A* hybrid, harbouring six copy number variations (Leffler *et al*., 2017). However, at the gene level DUP4 has a duplication of *GYPE*, triplication of the 3' end of *GYPA* and duplication of the 5' end of *GYPB* as well as a deletion of the 3' end of *GYPB*. Because of the complexity of this variant, it has been suggested that several unequal crossing over events were the responsible of DUP4 creation (Figure 1.18). The protein encoded by the glycophorin genes in DUP4 would bind to the extracellular domain of GPB, and the intracellular domain of GPA. However, the association signal found at FREM3 region was explained by DUP4 because of LD with the rs186873296 SNP.

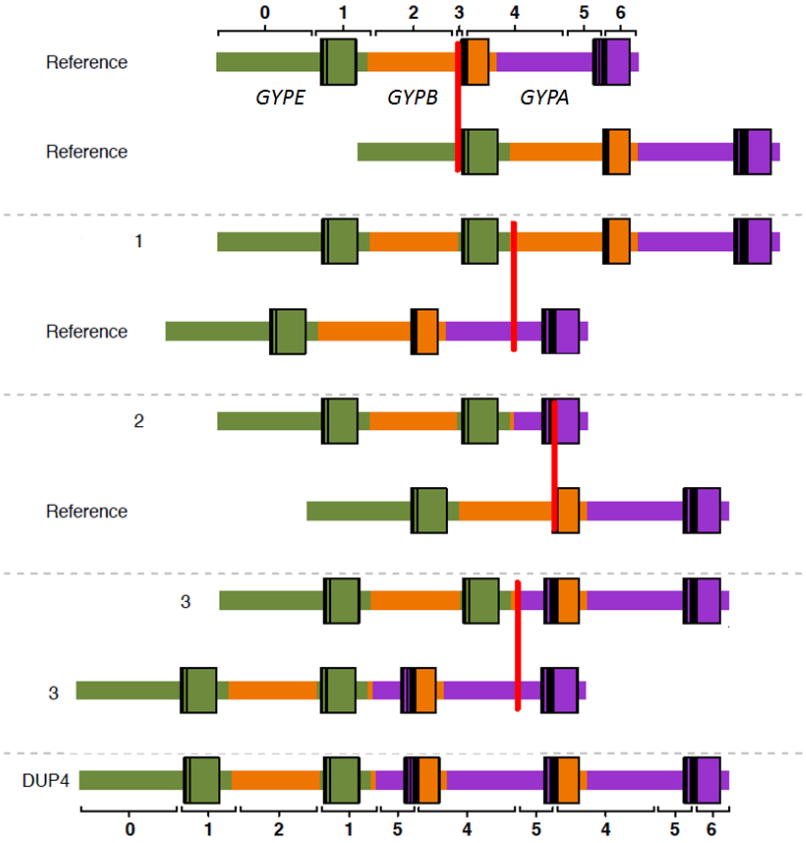



**Figure 1.18: Possible model of unequal crossing over events that may have given rise to DUP4 glycophorin variant.** Each crossing over event is indicated by the red vertical lines and numbers above correspond to segments of the reference sequence. The reference represents the glycophorin gene cluster. Modified figure from Leffler *et al*. (2017).

A recent study was conducted by Ndila *et al*. (2018) has investigated associations between some candidate polymorphisms and risk for severe *P.falciparum* malaria and its specific phenotypes, including cerebral malaria, severe malaria anaemia, and respiratory distress. These polymorphisms are related to the structure or the function of red blood cells. The case control study were in Kilifi County, Kenya. The study has used cases children with severe malaria to the high dependency ward of Kilifi County Hospital. Infants born in the local community between 2006 and 2010, who were part of a genetics study were used as controls in order to test associations between a range of candidate malaria protective genes and risk for severe malaria and its specific phenotypes.

The *GYPA* and *GYPB* are expressed abundantly on the erythrocyte surface and are targeted by parasites during invasion (Satchwell, 2016; Wright and Rayner, 2014). The *P.falciparum* EBA175 binds to the extracellular portion of GPA, which is present in DUP4. However, the *P.falciparum* EBL1 binds to the extracellular portion of GPB which is duplicated in DUP4 but joined to intracellular GPA (Figure 1.19). The duplicated copy of *GYPE* in DUP4 is

uncertain, since *GYPE* is not known to be expressed at the protein level (Cartron and Rouger, 1995). These changes in *GYPA* and *GYPB* copy number could have complex functional effects on the red cell membrane. There are physical interactions between GPA and band 3 (encoded by *SLC4A1*) at the red cell surface (Huang *et al*., 1996) and parasite binding to GPA appears to initiate a signal leading to increased membrane rigidity (Chasis *et al*., 1988). Therefore, the DUP4 GPB-A hybrid proteins could affect both receptor-ligand interactions and the physical characteristics of the red cell membrane (Figure 1.19).

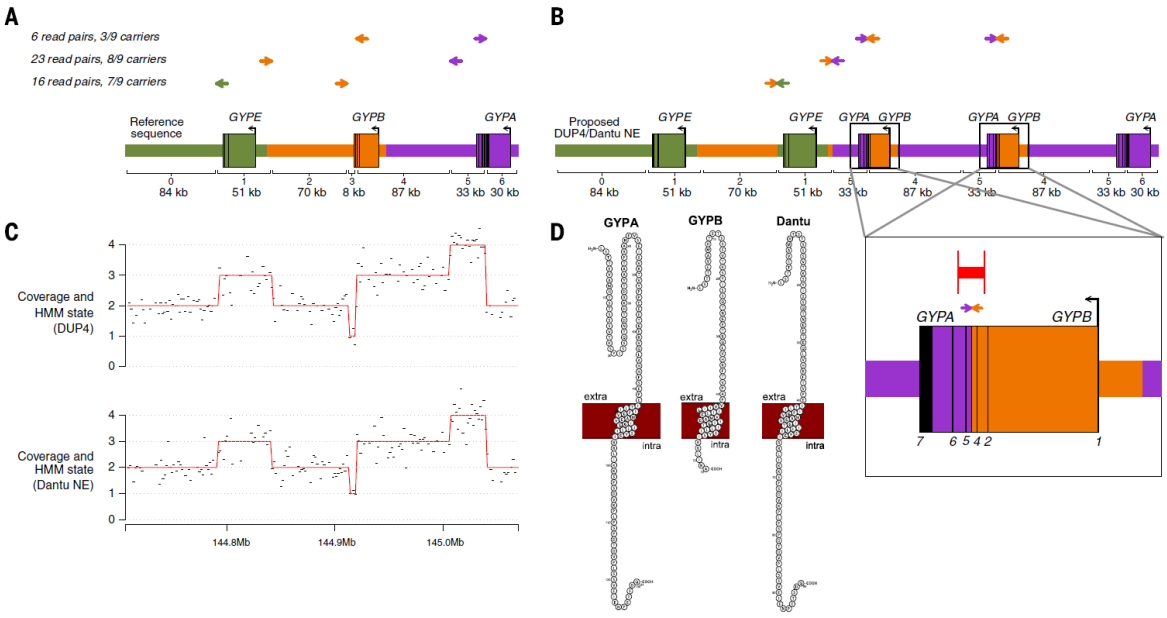



**Figure 1.19: The DUP4 (Dantu NE) protein structure. (A)** Reference sequence. **(B)** The structure of DUP4. **(C)** Normalized coverage in 1600 bp windows (black) and HMM path (red) for a DUP4 carrier (top) and for an individual serotyped as Dantu NE. **(D)** Protein structure of GPA, GPB, and the Dantu hybrid (GPB-A) within the cell membrane illustrating the extracellular, transmembrane, and intracellular domains. Obtained figure from Leffler *et al*. (2017).

## 1.13 Aims of the study

- Characterization and validation of copy number variation of the human glycophorin gene cluster and Identification of the glycophorin variant breakpoints.

- Design robust PRT assays to explore the extent of glycophorin CNV in different global populations.

- Examine the association of the rs186873296 near FREM3 with glycophorin copy number variants.

- Test the association of glycophorin CNV in two epidemiological cohorts with malaria phenotype information.

# Chapter 2: Materials and Methods

## 2.1 DNA samples used

### 2.1.1 HapMap samples

The International HapMap project began in October 2002. This project was studied the genetic and haplotype diversity of four different populations: the Yoruba from Ibadan in Nigeria (YRI-90 individuals), Japanese people from Tokyo (JPT-45 individuals), Han Chinese from Beijing (CHB-45 individuals), and northern and western Europeans from Utah (CEU-90 individuals) (International HapMap Consortium, 2003). The HapMap CEU and YRI samples provided 30 sets of samples from two parents and an adult child (each such set is termed a trio), whereas the total of 90 HapMap Asian samples comprised 45 samples from unrelated Han Chinese from Beijing (CHB), China, and the other 45 from unrelated Japanese from Tokyo (JPT), Japan.

The unrelated individuals from these cohorts were incorporated into the 1000 Genomes Project and sequenced at low coverage (http://www.internationalgenome.org/) and bam files were downloaded from (1000 Genomes Project Consortium *et al*., 2010). The blood samples were converted into lymphoblastoid cell lines using Epstein Barr Virus transformation of peripheral blood lymphocytes by the Coriell Institute (http://ccr.coriell.org). DNA and cell lines from the samples for the research projects were approved by appropriate ethics committees and provided by the Coriell Institute. These samples were typed to estimate the glycophorin copy number using the paralogue ratio test (PRT) method. In general, all chromosome coordinates in this thesis refer to GRCh37 (hg19).

### 2.1.2 ECACC Human Random Control (HRC) samples

The ECACC Human Random Control (HRC) DNA samples represent a control population of 480 original UK Caucasian blood donors. The Public Health England website provides more details about these samples (http://www.hpacultures.org.uk/products/dna/hrcdna/hrcdna.jsp). They were extracted from lymphoblastoid cell lines derived from Epstein Barr Virus (EBV) transformation of peripheral blood lymphocytes from each single-donor blood sample. The stock solutions of these genomic DNAs were provided by Dr Edward Hollox with a 100 ng/µl standard concentration. Generally, the working solutions were between 5-10 ng/µl, and this was used as a reference standard in all assays.

### 2.1.3 CEPH family samples

Lymphoblastoid cell lines of 809 individuals of European origin in 62 three-generation pedigrees were developed by the Centre d'Étude du Polymorphisme Humain (CEPH) (Stevens *et al*., 2012). The CEPH's large collection of reference samples has been extensively used as a benchmark for the analysis of genetic variants to create linkage maps of the human genome and provide samples to the HapMap and 1000 Genomes Projects (NIH/CEPH Collaborative Mapping Group., 1992). However, some of the CEPH samples were used as positive controls for some of the glycophorin variants. These samples were provided by the Coriell Institute.

### 2.1.4 Tori-Bossito malaria cohort (Benin cohort)

Benin is a tropical country located on the west coast of Africa (Figure 2.1). The Tori-Bossito project was carried out between June 2007 and January 2010 and financed by France's Agence Nationale de la Recherche (ANR) (http://www.agence-nationale-recherche.fr/en/). The 93 projects were managed by David Courtin and André Garcia in the Atlantique Department of southern Benin (Institut de Recherche pour le De´veloppement (IRD), UMR216, Me`re et Enfant Face aux Infections Tropicales, Cotonou, Benin). The aim of the project was to investigate the relationship between placental malaria infection and peripheral parasitaemias in infants during the first years of their lives while taking into account environmental and nutritional variables, therefore decreasing the possibility of biases. The infants were actively monitored clinically, parasitologically and immunologically from birth until the age of 18 months. The project was carried out in southern Benin, in the area of Tori-Bossito in the southern part of the Atlantic Department, north of Ouidah Community.

The Tori-Bossito cohort consisted of 656 infants (only 600 infants from the study of Le Port *et al*. (2012) were provided for this study) who received a parasitological (symptomatic and asymptomatic parasitaemia) and nutritional follow-up from birth to 18 months. Ecological, entomological and behavioural data were collected throughout the duration of the study (Le Port *et al*., 2012). The collaborators in this study sent 583 DNA samples collected from a rural area in Benin with two seasonal peaks in malaria transmission.

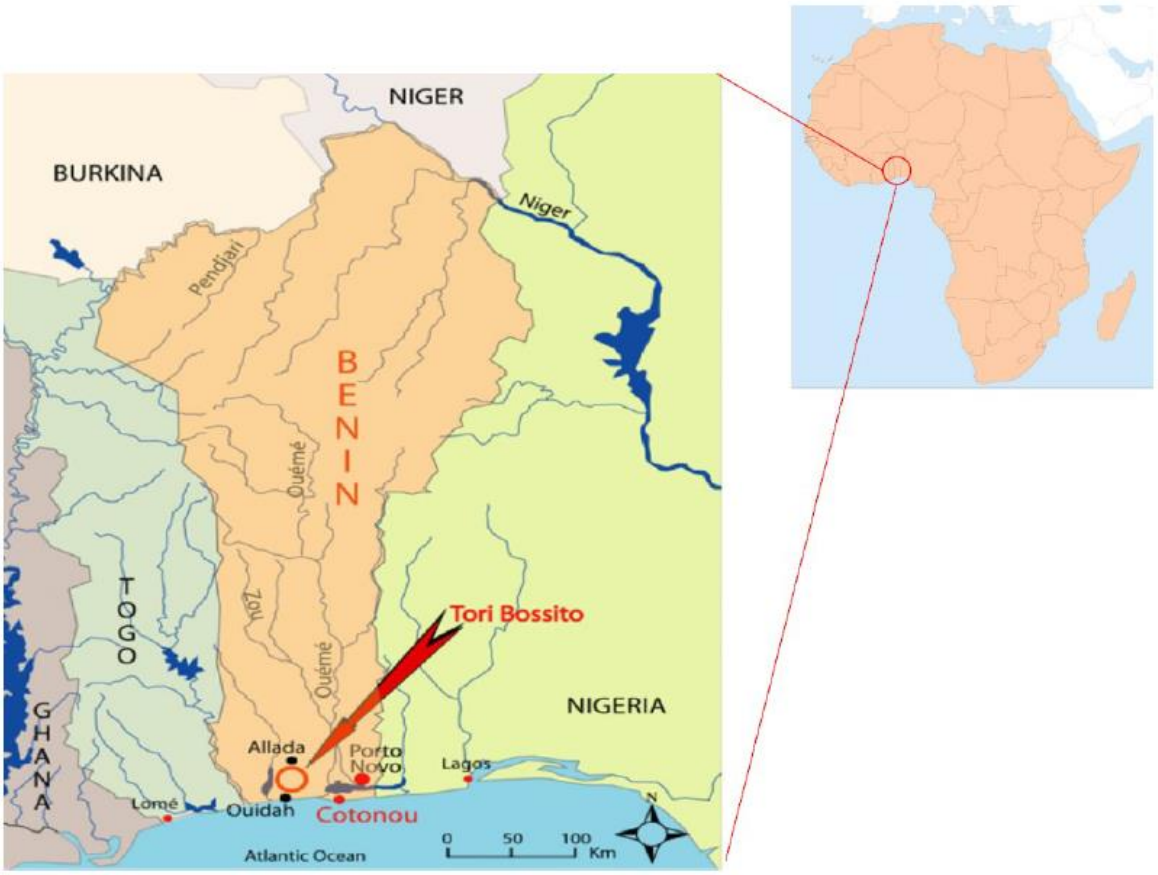



**Figure 2.1: Geographical locations of Tori-Bossito (Benin).** The malaria samples were collected from two locations. The red arrow indicates the location of Tori-Bossito, Benin. Adapted from Le Port *et al*. (2012) and http://www.freeworldmaps.net/.

### 2.1.5 Tanzanian family-based malaria cohort

Tanzania is located on the east coast of Africa (Figure 2.2). The Tanzanian cohort was drawn from a family-based study of the Tanzanian Bantu population (n = 922) in the fishing village of Nyamisati, in the Rufiji river delta, 150km south of Dar-es-Salaam (Carpenter *et al*., 2007, 2009; Rooth, 1992). The region has two rainy seasons (April–June and November–December) and is holoendemic for malaria, with an increase in transmission during the rainy seasons (Rooth, 1992). The prevalence of the predominant malaria parasite of this region (*P. falciparum*) in 1993 was 75%, which had fallen to 48% by 1998, as measured by microscopy in children aged between two and nine years (Carpenter *et al*., 2012). Epidemiological and clinical phenotypes (parasite load, mean number of clinical episodes of malaria and haemoglobin levels) were generated from non-selected 'total' population surveys carried out annually during March and April between 1993 and 1999. The complete annual record of clinical malaria episodes in Nyamisati between 1993 and 1999 were documented by Rooth, who has lived in the village since 1985 (Carpenter *et al*., 2012).


**Figure 2.2: Geographical locations of Nyamisati (Tanzania).** Image **(A)** shows a basic map of Tanzania representing the collection point, Nyamisati village within the Rufiji District. Image **(B)** shows a Google satellite map of Nyamisati in the Rufiji River Delta. Image modified from Färnert *et al*. (2014).

## 2.2 Copy number analysis of glycophorin region

### 2.2.1 Paralogue ratio test (PRT) for glycophorin genes

This project used the paralogue ratio test (PRT) to determine the copy number of the glycophorin gene family in humans. These PRT project primers were designed using PRTPrimer software. PRTPrime is an automated approach to design PRT primers using four computation-based steps (Veal *et al*. 2013). First, a large number of primer pairs are designed in the target interval. Second, the location of potential amplification sites of these primer pairs in the reference human genome is determined. Third, those that are perfect priming matches for only two amplicons in the reference genome are isolated. Fourth, filtering is applied to identify optimal PRT assays for the target region (Figure 2.3). The resource may be used in three ways: the software pre-designed assays can be used across the human genome using the search function, specific assays can be designed using the Online Design function (used for this project), or the software can be downloaded and run locally (Veal *et al*. 2013).

In this project, PRTPrimer online software (http://www.prtprimer.org/) was chosen to design the PRT primers. As the PRT primer is very sensitive and difficult to design, the primer criteria were inserted into the software. For example, the test and the reference primers have

to be located in the glycophorin gene cluster. The project aimed to have between 90 – 450 bp PRT-PCR products in any of the tests or the references in order to visualise them on the agarose gel and on the GeneMapper® software after capillary electrophoresis. However, a size difference of at least three bases between the test and the reference product is needed in order to differentiate and identify them. In general, the optimal primer size is 18 – 27 bp in length, which is inserted in the software (Figure 2.4). Under these criteria, PRTPrimer software provides 915 primer pairs, which were pasted into an Excel spreadsheet for organisation and ease of analysis. However, each pair consists of two tests and one reference product because the glycophorin gene cluster consists of three very close and similar genes. The minimum size difference between the tests and the reference provided by the software was four bases, whereas the maximum size difference was 22 bases (Figure 2.5).

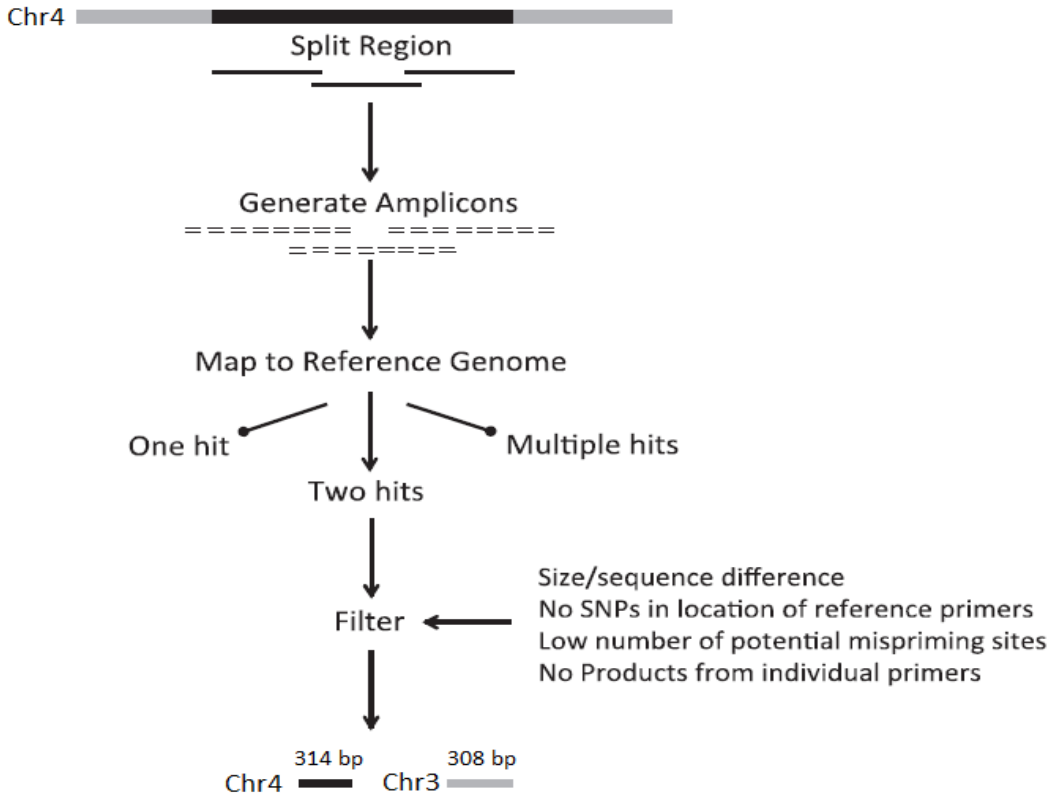



**Figure 2.3: Overview of PRTPrimer.** The target region, shown in black, is required for PRTs on chromosome 4. First, the software splits the region into overlapping segments to ensure an even distribution of PRTs. A large number of amplicons is designed for each of these segments. Following this, each amplicon is aligned to the human genome, allowing for mismatches with the primers. Only amplicons that have exact priming matches twice in the genome are selected for filtering. The final stage filters the results to only those amplicons that meet adjustable criteria, such as size difference between target and reference, no SNPs at primer positions and CNVs spanning the reference amplicon. Here, the target amplicon is 314 bp and the reference amplicon on a different chromosome is 308bp. Modified figure from Veal *et al*., (2013).

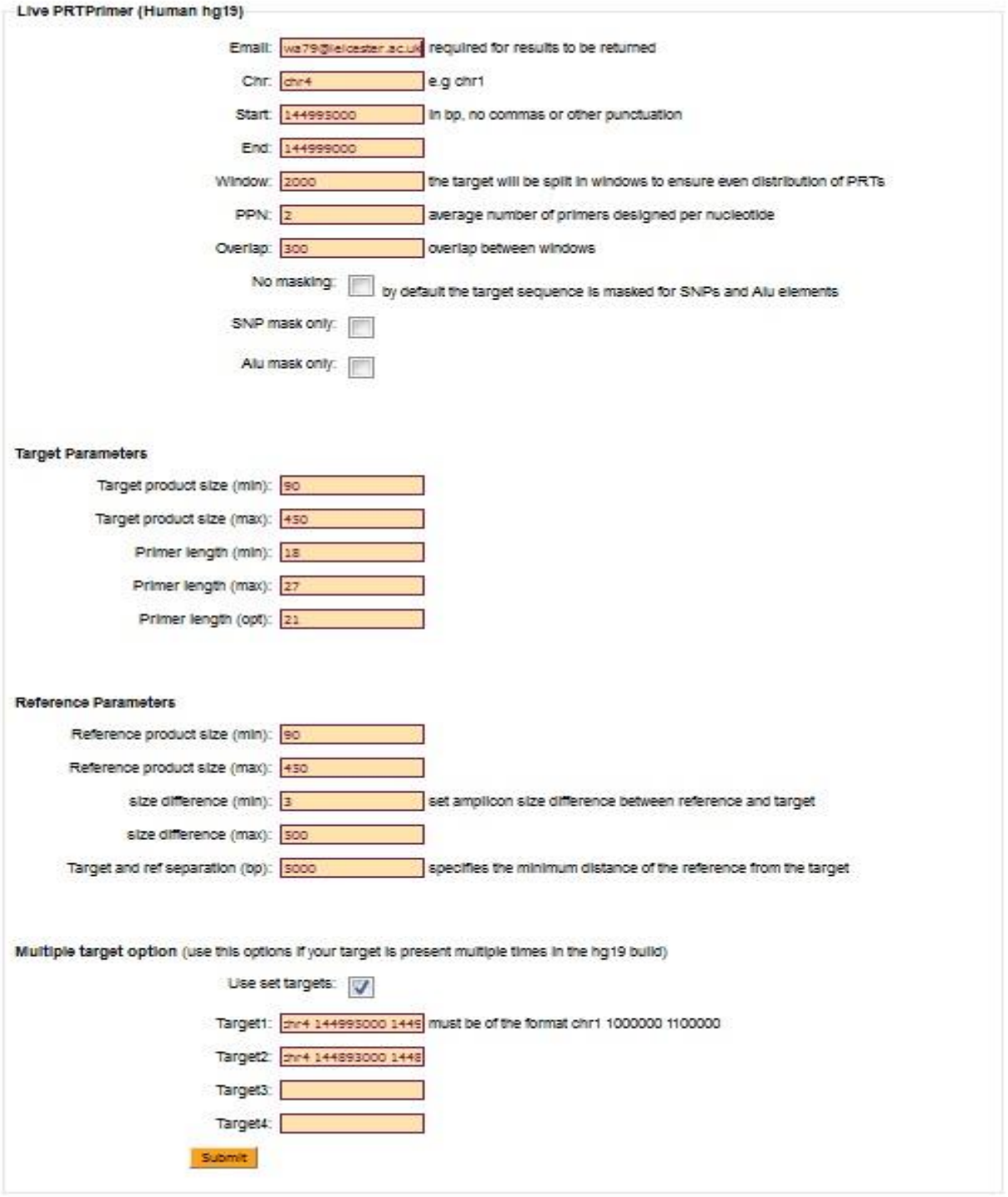
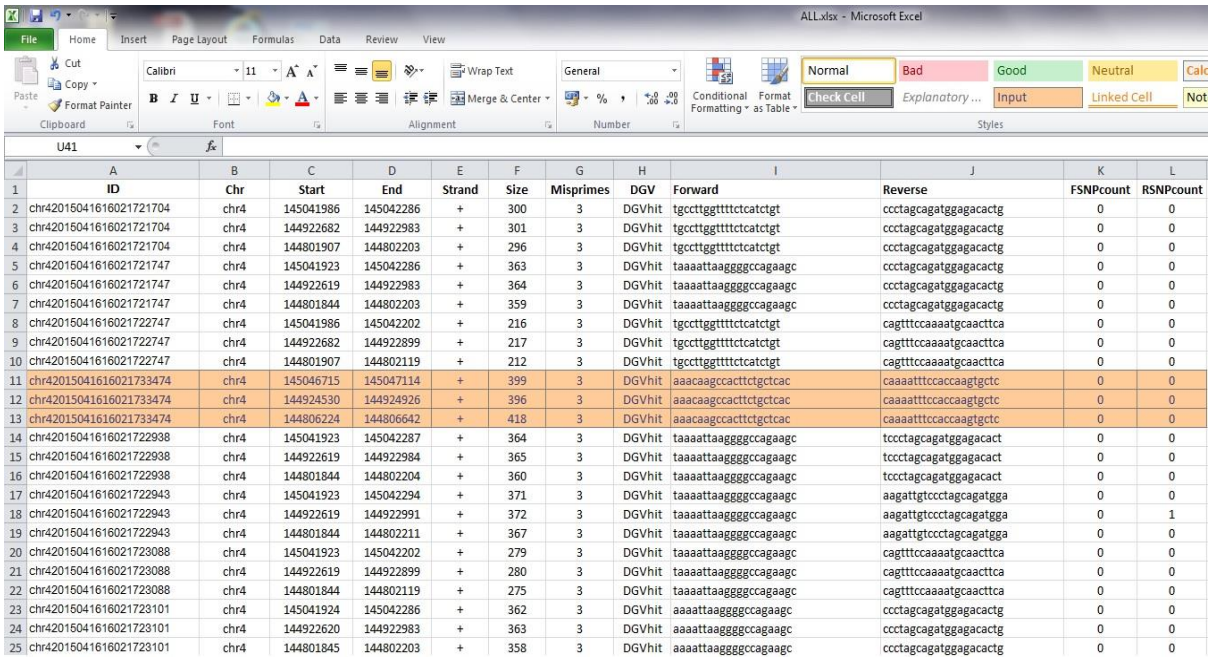



**Figure 2.4: PRTPrimer Online page of website.** The live PRTPrimer section should contain an email address in order to receive the results, the chromosome number, (chr4 in this case), and the specific region of interest on the chromosome. The target parameters section should contain the test primer criteria, while the reference parameters section is for reference primer criteria. Here the minimum size difference was three bases in order to differentiate between test and reference peaks when viewing the results using Genemapper software. However, the last section (multiple target option) was used for designing cis-PRT primers for *GYPA* (target1), *GYPB* (target 2). However, as the software automatically chooses *GYPE* as a reference, in Section 3 the target and reference separation was reduced to 5000 bp.



ALL.xlsx - Microsoft Excel

| | ID | Chr | Start | End | Strand | Size | Misprimes | DGV | Forward | Reverse | FSNPcount | RSNPcount |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | chr4201504161602172l704 | chr4 | 145041986 | 145042286 | + | 300 | 3 | DGVhit | tgccttggttttctcatctgt | ccctagcagatggagacactg | 0 | 0 |
| 3 | chr4201504161602172l704 | chr4 | 144922682 | 144922983 | + | 301 | 3 | DGVhit | tgccttggttttctcatctgt | ccctagcagatggagacactg | 0 | 0 |
| 4 | chr4201504161602172l704 | chr4 | 144801907 | 144802203 | + | 296 | 3 | DGVhit | tgccttggttttctcatctgt | ccctagcagatggagacactg | 0 | 0 |
| 5 | chr4201504161602172l747 | chr4 | 145041923 | 145042286 | + | 363 | 3 | DGVhit | taaaattaaggggccagaagc | ccctagcagatggagacactg | 0 | 0 |
| 6 | chr4201504161602172l747 | chr4 | 144922619 | 144922983 | + | 364 | 3 | DGVhit | taaaattaaggggccagaagc | ccctagcagatggagacactg | 0 | 0 |
| 7 | chr4201504161602172l747 | chr4 | 144801844 | 144802203 | + | 359 | 3 | DGVhit | taaaattaaggggccagaagc | ccctagcagatggagacactg | 0 | 0 |
| 8 | chr4201504161602172l747 | chr4 | 145041986 | 145042202 | + | 216 | 3 | DGVhit | tgccttggttttctcatctgt | cagtttccaaaatgcaacttca | 0 | 0 |
| 9 | chr4201504161602172l747 | chr4 | 144922682 | 144922899 | + | 217 | 3 | DGVhit | tgccttggttttctcatctgt | cagtttccaaaatgcaacttca | 0 | 0 |
| 10 | chr4201504161602172l747 | chr4 | 144801907 | 144802119 | + | 212 | 3 | DGVhit | tgccttggttttctcatctgt | cagtttccaaaatgcaacttca | 0 | 0 |
| 11 | chr4201504161602173347l4 | chr4 | 145046715 | 145047114 | + | 399 | 3 | DGVhit | aaacaagccacttctgctcac | caaaatttccaccaagtgctc | 0 | 0 |
| 12 | chr4201504161602173347l4 | chr4 | 144924530 | 144924926 | + | 396 | 3 | DGVhit | aaacaagccacttctgctcac | caaaatttccaccaagtgctc | 0 | 0 |
| 13 | chr4201504161602173347l4 | chr4 | 144806224 | 144806642 | + | 418 | 3 | DGVhit | aaacaagccacttctgctcac | caaaatttccaccaagtgctc | 0 | 0 |
| 14 | chr4201504161602172293l8 | chr4 | 145041923 | 145042287 | + | 364 | 3 | DGVhit | taaaattaaggggccagaagc | tccctagcagatggagacact | 0 | 0 |
| 15 | chr4201504161602172293l8 | chr4 | 144922619 | 144922984 | + | 365 | 3 | DGVhit | taaaattaaggggccagaagc | tccctagcagatggagacact | 0 | 0 |
| 16 | chr4201504161602172293l8 | chr4 | 144801844 | 144802204 | + | 360 | 3 | DGVhit | taaaattaaggggccagaagc | tccctagcagatggagacact | 0 | 0 |
| 17 | chr4201504161602172294l3 | chr4 | 145041923 | 145042294 | + | 371 | 3 | DGVhit | taaaattaaggggccagaagc | aagattgtccctagcagatgga | 0 | 0 |
| 18 | chr4201504161602172294l3 | chr4 | 144922619 | 144922991 | + | 372 | 3 | DGVhit | taaaattaaggggccagaagc | aagattgtccctagcagatgga | 0 | 1 |
| 19 | chr4201504161602172294l3 | chr4 | 144801844 | 144802211 | + | 367 | 3 | DGVhit | taaaattaaggggccagaagc | aagattgtccctagcagatgga | 0 | 0 |
| 20 | chr4201504161602172308l8 | chr4 | 145041923 | 145042202 | + | 279 | 3 | DGVhit | taaaattaaggggccagaagc | cagtttccaaaatgcaacttca | 0 | 0 |
| 21 | chr4201504161602172308l8 | chr4 | 144922619 | 144922899 | + | 280 | 3 | DGVhit | taaaattaaggggccagaagc | cagtttccaaaatgcaacttca | 0 | 0 |
| 22 | chr4201504161602172308l8 | chr4 | 144801844 | 144802119 | + | 275 | 3 | DGVhit | taaaattaaggggccagaagc | cagtttccaaaatgcaacttca | 0 | 0 |
| 23 | chr4201504161602172310l1 | chr4 | 145041924 | 145042286 | + | 362 | 3 | DGVhit | aaaattaaggggccagaagc | ccctagcagatggagacactg | 0 | 0 |
| 24 | chr4201504161602172310l1 | chr4 | 144922620 | 144922983 | + | 363 | 3 | DGVhit | aaaattaaggggccagaagc | ccctagcagatggagacactg | 0 | 0 |
| 25 | chr4201504161602172310l1 | chr4 | 144801845 | 144802203 | + | 358 | 3 | DGVhit | aaaattaaggggccagaagc | ccctagcagatggagacactg | 0 | 0 |

**Figure 2.5: Part of the primer pairs list. From the 915 primer pairs list, cis-PRT 1 was chosen (highlighted) and optimised.** Misprime means anneal to locations in DNA which might be present other than the intended location.

Seven primer pairs were designed for the determination of CN for the glycophorin genes using PRTPrimers®. This was applied on the HapMap samples to validate the PRT assays and was subsequently applied to two different malaria cohorts (Chapter 5). These primers were grouped in to two types: cis_PRT primers, and trans_PRT primers (Table 2.1). The difference between these types is the location of the reference, which in the case of cis_PRT binds near and on the same chromosome in the test (Figure 2.4), whereas the reference of trans_PRT is located on a different chromosome in the test (Figure 2.5). The primers were between 20 and 22 bp in length, and the forward cis_PRT primers were labelled as FAM, FAM and HEX for cis_PRT 1, 2, and 4 respectively. However, the forward primer of trans_PRT2 was labelled once with HEX and once with FAM, and the forward primer of trans_PRT3 was labelled once with HEX and once with NED in order to distinguish between the trans and cis primer products in the final results.

**Table 2.1: The final PRT primer pairs.**

| Primer name | Forward and Reverse sequences | Dye | Expected Products |
|---|---|---|---|
| **Cis-PRT 1** | 5' AAACAAGCCACTTCTGCTCAC 3'<br>Forward<br>5' CAAAATTTCCACCAAGTGCTC 3'<br>Reverse | FAM<br><br>Unlabelled | 1- *GYPA* (399 bp)<br>2- *GYPB* (396 bp)<br>3- *GYPE* (418 bp) |
| **Cis-PRT 2** | 5' TCACCCGAGTTGTGTACATTG 3'<br>Forward<br>5' TGTCTATTGCAACCCAATTCA 3'<br>Reverse | FAM<br><br>Unlabelled | 1- *GYPA* (366 bp)<br>2- *GYPB* (363 bp)<br>3- *GYPE* (382 bp) |
| **Cis-PRT 4** | 5' GAAAGGTTTTCCAGAGCAAGC 3'<br>Forward<br>5' TCAGACCATTGCTCATGCTG 3'<br>Reverse | HEX<br><br>Unlabelled | 1- *GYPA* (105 bp)<br>2- *GYPB* (102 bp)<br>3- *GYPE* (105 bp) |
| **Trans-PRT2** | 5' TACAGGCATTGGGTAAATGCT 3'<br>Forward<br>5' CCCCAGAATAGTAGGTCCACTG 3'<br>Reverse | HEX+FAM<br><br>Unlabelled | 1- *GYPA,GYPB* and *GYPE* (314 bp)<br>2- Chr.3 (308 bp) |
| **Trans-PRT3** | 5' GGAAACTTACAATCGTGGCAG 3'<br>Forward<br>5' TACAGGCTCATAGGTGGAAGG 3'<br>Reverse | HEX+NED<br><br>Unlabelled | 1- *GYPA,GYPB* and *GYPE* (415 bp)<br>2- Chr.3 (434 bp)<br>3- Chr.9 (444 bp) |

### 2.2.2 PRT assay positive controls

Six DNA samples from the Coriell Institute were used as positive controls (Table 2.2). These samples have known glycophorin CNs from a study by Handsaker *et al*. (2015), which used 849 genomes sequenced by the 1000 Genomes Project to identify most large (>5-kb) multiallelic copy number variations (mCNVs), including 3,878 duplications, of which 1,356 appeared to have three or more segregating alleles. Following this study, a supported initial data resource was created.

**Table 2.2: PRT_PCR positive controls information.** These DNA samples were ordered from Coriell Cell Repositories, 403 Haddon Avenue, Camden, New Jersey, 08103, USA.

| DNA Sample | CN | Concentration mg/ml | Cell Type | Description |
|---|---|---|---|---|
| **NA19190** | 1 copy | 0.342 | B-Lymphocyte | International HapMap Project and 1000 Genomes Project-Yoruba in Ibadan, Nigeria |
| **NA19818** | 1 copy | 0.305 | B-Lymphocyte | International HapMap Project and 1000 Genomes Project-African Ancestry in Southwest USA |
| **NA19085** | 2 copies | 0.34 | B-Lymphocyte | International HapMap Project and 1000 Genomes Project-Japanese in Tokyo, Japan |
| **NA19777** | 2 copies | 0.342 | B-Lymphocyte | International HapMap Project and 1000 Genomes Project-Mexican Ancestry in Los Angeles, California, USA |
| **NA19084** | 3 copies | 0.351 | B-Lymphocyte | International HapMap Project and 1000 Genomes Project-Japanese in Tokyo, Japan |
| **NA19371** | 3 copies | 0.357 | B-Lymphocyte | International HapMap Project and 1000 Genomes Project-Luhya in Webuye, Kenya |

### 2.2.3 Optimisation stage for the assays

#### 2.2.3.1 Gradient PCR

The gradient PCR was applied using a Veriti® Thermal Cycler PCR machine in order to investigate the optimum annealing temperature for each primer pair (Tables 2.3 and 2.4).

**Table 2.3: Gradient PRC mixture.**

| Reagent | Concentration | Volume per reaction |
|---|---|---|
| **Genomic DNA** | 10 ng/µl | 1.0 µl |
| **Forward Primer** | 10 µM | 0.5 µl |
| **Reverse Primer** | 10 µM | 0.5 µl |
| **X10 Low dNTPs Buffer** | - | 1.0 µl |
| **X10 Kapa Buffer** | - | 1.0 µl |
| **Taq Polymerase** | 5U/µl | 0.1 µl |
| **$H_2O$** | - | 5.9 µl |
| **Total** | - | 10 µl |

**Table 2.4: Gradient PCR conditions.**

| Cycling Condition | | Temperature | Time |
|---|---|---|---|
| **Step 1: Initial denaturation** | | 94°C | 2 min |
| **Step 2: Denaturation** | 35 Cycles | 94°C | 30 sec |
| **Step 3: Annealing (Gradient)** | | (60-63-65-67-68-70)°C | 30 sec |
| **Step 4: Extension** | | 70°C | 30 sec |
| **Step 5: Extension** | | 70°C | 5 min |
| **Step 6: Hold** | | 10°C | ∞ |

#### 2.2.3.2 Agarose gel electrophoresis

In order to choose the best annealing temperature for each assay, the PCR products for each primer were run through 2% agarose gel electrophoresis. 100 ml of 1xTBE buffer (Appendix 1) was added to 2g of agarose and heated in a microwave until the agarose completely dissolved; 0.5 µg/ml of ethidium bromide was then added to the solution. After the solution had cooled down, it was moved into a casting tray; and after it had solidified, the comb was removed and the gel casting tray was flooded in the electrophoresis tank containing a 1xTBE buffer. 10 µl of DNA was mixed with 2 µl of 5x loading dye and loaded into the wells alongside 5µl of HyperLadder$^{TM}$V (Bioline DNA marker). Electrophoresis was performed at 100 V for 60 minutes. The gel was then observed under a UV transilluminator.

### 2.2.3.3 PRT-PCR conditions

Three multiplex PCRs were designed. One multiplex PCR was used for the cis_PRT primers (Tables 2.5 and 2.6) and two multiplex PCRs for the trans_PRT primers (Tables 2.7, 2.8 and 2.9):

**Table 2.5: Cis_PRTs multiplex PCR mixture.**

| Reagent | Concentration | Volume per reaction |
|---|---|---|
| **Genomic DNA** | 10 ng/µl | 1.0 µl |
| **Cis-PRT 1 Forward (FAM)** | 10 µM | 0.5 µl |
| **Cis-PRT 1 Reverse** | 10 µM | 0.5 µl |
| **Cis-PRT 2 Forward (FAM)** | 10 µM | 0.5 µl |
| **Cis-PRT 2 Reverse** | 10 µM | 0.5 µl |
| **Cis-PRT 4 Forward (HEX)** | 10 µM | 0.5 µl |
| **Cis-PRT 4 Reverse** | 10 µM | 0.5 µl |
| **×10 Low dNTPs Buffer** | - | 1.0 µl |
| **×10 Kapa Buffer** | - | 1.0 µl |
| **Taq Polymerase** | 5 U/µl | 0.1 µl |
| **Water** | - | 3.9 µl |
| **Total** | - | 10.0 µl |

**Table 2.6: Cis_PRTs multiplex PCR conditions.**

| Cycling Condition | | Temperature | Time |
|---|---|---|---|
| **Step 1: Initial denaturation** | | 94°C | 2 min |
| **Step 2: Denaturation** | 25 Cycles | 94°C | 30 sec |
| **Step 3: Annealing** | | 65°C | 30 sec |
| **Step 4: Extension** | | 70°C | 30 sec |
| **Step 5: Extension** | | 70°C | 5 min |
| **Step 6: Hold** | | 10°C | ∞ |

**Table 2.7: Trans_PRTs multiplex (1) PCR mixture.**

| Reagent | Concentration | Volume per reaction |
|---|---|---|
| **Genomic DNA** | 10 ng/µl | 1.0 µl |
| **Trans-PRT 2 Forward (HEX)** | 10 µM | 0.5 µl |
| **Trans-PRT 2 Reverse** | 10 µM | 0.5 µl |
| **Trans-PRT 3 Forward (NED)** | 10 µM | 0.5 µl |
| **Trans-PRT 3 Reverse** | 10 µM | 0.5 µl |
| **×10 Low dNTPs Buffer** | - | 1.0 µl |
| **×10 Kapa Buffer** | - | 1.0 µl |
| **Taq Polymerase** | 5 U/µl | 0.1 µl |
| **Water** | - | 4.9 µl |
| **Total** | - | 10.0 µl |

**Table 2.8: Trans_PRTs multiplex (2) PCR mixture.**

| Reagent | Concentration | Volume per reaction |
|---|---|---|
| **Genomic DNA** | 10 ng/µl | 1.0 µl |
| **Trans-PRT 2 Forward (FAM)** | 10 µM | 0.5 µl |
| **Trans-PRT 2 Reverse** | 10 µM | 0.5 µl |
| **Trans-PRT 3 Forward (HEX)** | 10 µM | 0.5 µl |
| **Trans-PRT 3 Reverse** | 10 µM | 0.5 µl |
| **×10 Low dNTPs Buffer** | - | 1.0 µl |
| **×10 Kapa Buffer** | - | 1.0 µl |
| **Taq Polymerase** | 5 U/µl | 0.1 µl |
| **Water** | - | 4.9 µl |
| **Total** | - | 10.0 µl |

**Table 2.9: Trans_PRTs multiplex PCR conditions for both multiplexes.**

| Cycling Condition | | Temperature | Time |
|---|---|---|---|
| **Step 1: Initial denaturation** | | 94°C | 2 min |
| **Step 2: Denaturation** | 25 Cycles | 94°C | 30 sec |
| **Step 3: Annealing** | | 68°C | 30 sec |
| **Step 4: Extension** | | 70°C | 30 sec |
| **Step 5: Extension** | | 70°C | 5 min |
| **Step 6: Hold** | | 10°C | ∞ |

### 2.2.4 Capillary electrophoresis

In order to distinguish size differences between the test and the reference to be used for the copy number calculation, the PCR amplicons were resolved using capillary electrophoresis (ABI 3130xl Genetic Analyser). First, 5 µl of MapMarker® 1000XL Rhodamin (Rox1000XL) 50 – 1000bp (BioVentures, Inc.) was added to 1000 µl HIDI Formamide, and 10 µl of the mixture was aliquoted to a 96-well plate. Next, from each of the three multiplex PCR products, 1 µl was added to each well so that each well contained 13 µl in total. The plate was denatured at 96°C for three minutes on the PCR machine, and then moved immediately to ice for two minutes. Following this, the ABI plate was placed in the Genetic Analyser. Fragment analysis was carried out by electrophoresis on a 36-cm capillary using POP-4 polymer with an injection time of 30 seconds. The analysis was carried out using GeneMapper® software (ABI). However, the GeneMapper® software was used to measure the peak areas.

### 2.2.5 CN estimation calculation

Optimum peak area that can be accepted is between 300 and 60,000. An Excel sheet was exported from GeneMapper® that contains all product peak areas of each positive control and unknown CN samples.

From this Excel sheet the test/reference peak area ratio were calculated for each sample, then on a new Excel file the peak areas of the positive controls were used to create a calibration curve of peak area ratios (X axis) versus known copy numbers (Y axis) that provides the slope and the intercept. A specific equation was used the slope and the intercept with peak area ratio of the samples to calculate the estimated CN (`slope x peak area ratio + intercept`).

The predicted copy numbers data were normalized to be fit with the (CN/ref) NGS data, after that, the final Excel file was converted to text file in order to generate a scatterplot. The text file was used in the R program to create a scatterplot (Appendix 6) for each assay representing the estimated glycophorin genes copy number data for any set of PRT assay.

## 2.3 SNP genotyping using an Allele-Specific PCR based method

In order to type (rs186873296) SNP, an allele-specific assay was designed and used for this project.

### 2.3.1 Amplification refractory mutation system (ARMS)

The amplification refractory mutation system (ARMS) is a molecular technique generally used to detect known SNPs. This allele-specific PCR is carried out using a common right primer and two left primers (a wild-type primer and a mutant-type primer) which have different lengths and fluorescent labels. These products can be distinguished using gel electrophoresis or an electropherogram automated sequencer (Huang *et al*., 2013).

### 2.3.2 Allele-Specific primer design

According to the Malaria Genomic Epidemiology Network *et al*. (2015), the rs186873296 SNP is located on chromosome 4:144702474, and its alleles are A or G, of which A is the ancestral allele (UCSC Genome Browser GRCh37/hg19, 2009; http://www.ensembl.org/index.html). Based on this information about the SNP, a primer pair was manually designed (Table 2.10). However, the forward primer was altered by an additional nucleic acid analogue, called Locked Nucleic Acid (LNA), at the 3' end (SNP location) of the forward EUROGENTEC® primer. Incorporating LNA into oligonucleotides improves sensitivity and specificity and increases the stability of the primer (Ugozzoli *et al*., 2004; Johnson *et al*., 2004).

Previous research has shown that 3' LNA residues improved the specificity of allele-specific PCR primers compared to native DNA primers (Latorra *et al*., 2003a; Latorra *et al*., 2003b). In addition, Ballantyne *et al*. (2008) suggest that using LNA primers or probes significantly enhances real-time PCR and DNA sequencing. Moreover, many other technologies can use LNA, such as microarray and *in situ* hybridization (Kauppinen *et al*., 2003). The difference between LNA and common DNA nucleic acid is the modified ribose moiety of the LNA, which comprises 2'-O, 4'-C-methylene bicyclonucleoside monomers, as shown in Figure 2.6 (Koshkin *et al*., 1998). This covalent bridge effectively ‘locks’ the ribose in the N-type (3'-endo) conformation that is dominant in A-form DNA and RNA (You *et al*., 2006). However, the addition of single LNA will increase the annealing temperature by up to 8°C per LNA, which had to be taken into consideration during primer optimisation (Koshkin *et al*., 1998).

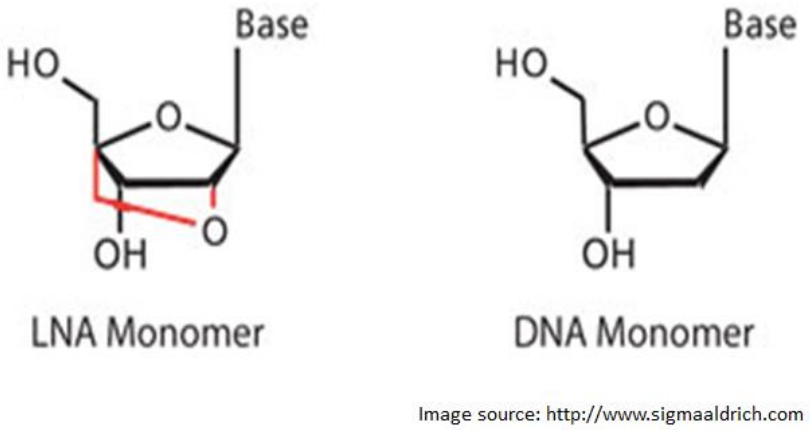



**Figure 2.6: Structure of Single Locked Nucleic Acid (LNA™) and common DNA nucleic acid.** An additional bridge connecting the 2' and 4' carbons is highlighted in red (Kauppinen *et al*., 2005).

However, only a positive control known to be heterozygous for the G allele according to the NGS published data (NA19190) was found; no homozygous for the rare allele was found. Therefore, the forward primer will bind if the rare allele (G) of the SNP is present and gives a product with 200 bp (Appendix 5), but will not work if the common allele (A) is present at the SNP location, and amplification will not occur. As a consequence a primer pair acts as a positive control for successful PCR amplification was used. This positive primer pair (Table 2.10) was provided by Dr Razan Abujaber and amplifies a 301bp DNA sequence from a gene called *DEFB*105 on chromosome 8.

**Table 2.10: Primers for sequencing the positive samples.** The LNA base is shown between brackets.

| Primer name | Sequence |
|---|---|
| **rs186873296_ARMS_Forward** | 5' AGAAGCTGGGAACCCTGTC**[G]** 3' |
| **rs186873296_ARMS_Reverse** | 5' ATTGTAAGACAAGCAAACAGTACTGG 3' |
| **DEFB105_E1_Forward** | 5' ATTGATTCACTTCACGGATCAAG 3' |
| **DEFB105_E1_Reverse** | 5' TCTAAGAAATTCCCATGACAGGT 3' |

### 2.3.3 ARMS-PCR conditions

The PCR conditions were optimised using a Veriti® Thermal Cycler PCR machine, and a multiplex PCR was designed (Tables 2.11 and 2.12). In order to differentiate between the samples that carry the G allele and those that do not, the PCR products were run through 2% agarose gel electrophoresis.

**Table 2.11: ARMS assay PCR mixture.**

| Reagent | Concentration | Volume per reaction |
|---|---|---|
| **Genomic DNA** | 10 ng/µl | 1.0 µl |
| **Forward Primer** | 10 µM | 0.5 µl |
| **Reverse Primer** | 10 µM | 0.5 µl |
| **Positive Forward Primer** | 10 µM | 0.2 µl |
| **Positive Reverse Primer** | 10 µM | 0.2 µl |
| **×10 Kapa A Buffer** | - | 1.0 µl |
| **dNTPs** | 2.5 µM | 0.8 µl |
| **Taq Polymerase** | 5 U/µl | 0.2 µl |
| **$H_2O$** | - | 5.6 µl |
| **Total** | - | 10.0 µl |

**Table 2.12: ARMS assay PCR conditions.**

| Cycling Condition | | Temperature | Time |
|---|---|---|---|
| **Step 1: Initial denaturation** | | 95°C | 2 min |
| **Step 2: Denaturation** | 35 Cycles | 95°C | 30 sec |
| **Step 3: Annealing** | | 65°C | 30 sec |
| **Step 4: Extension** | | 70°C | 30 sec |
| **Step 5: Extension** | | 70°C | 5 min |
| **Step 6: Hold** | | 10°C | ∞ |

### 2.3.4 Homozygosity and Heterozygosity analysis

Because of the absence of the homozygous positive control for the G allele, it was difficult to determine whether a particular sample was homozygous or heterozygous. Therefore, a new primer pair was designed for the samples that were positive for the rare allele in order to sequence their PCR products. The forward primer was designed 148 bases downstream of the SNP while the reverse primer was designed 248 bases upstream of the SNP (Table 2.13) (Appendix 5). The expected PCR product size was 397 bp (Figure 2.7), which was sequenced for each sample that was positive for the G allele in order to determine if it was a homozygous or heterozygous sample.

**Table 2.13: Primers for sequencing the positive samples.**

| Primer name | Sequence |
|---|---|
| **Seq_SNP_WA_Forward** | 5' GAATTCGTTCTTGAAGGCAATT 3' |
| **Seq_SNP_WA_Reverse** | 5' GGAGCTTAAGCTTAAATGCCAAGT 3' |

The expected PCR product size was 397 bp (Figure 2.7), which was sequenced for each sample that was positive for the G allele in order to determine if it was a homozygous or heterozygous sample.

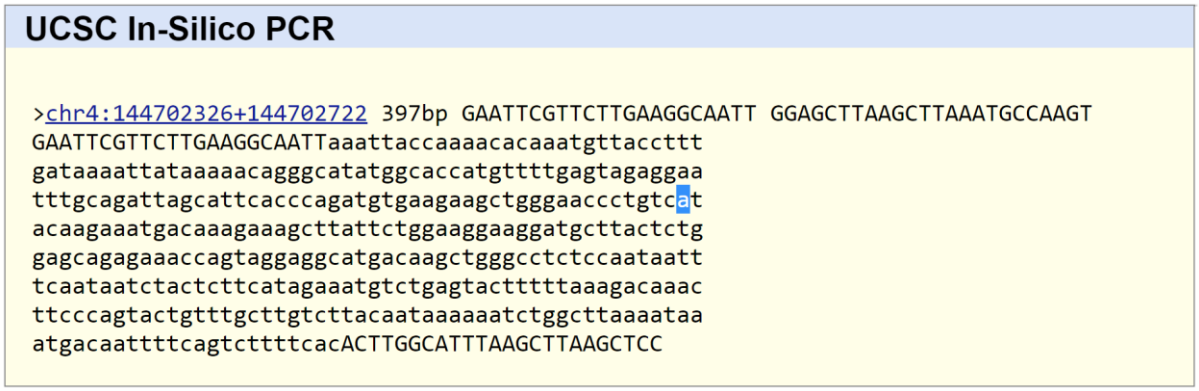



**Figure 2.7: Expected product of the ARMS sequencing primer pair.** The expected product size used for the sequencing was 397 bp. The blue highlighted base is the SNP (A or G). This figure is a screenshot from the UCSC genome browser (GRCh37/hg19) assembly.

Following this, the PCR conditions for the new primers were optimised using a Veriti® Thermal Cycler PCR machine, and a single PCR was designed (Tables 2.14 and 2.15).

**Table 2.14: PCR mixture.**

| Reagent | Concentration | Volume per reaction |
|---|---|---|
| **Genomic DNA** | 10 ng/µl | 1.0 µl |
| **Forward Primer** | 10 µM | 1.0 µl |
| **Reverse Primer** | 10 µM | 1.0 µl |
| **×10 Kapa A Buffer** | - | 2.0 µl |
| **dNTPs** | 2.5 µM | 1.0 µl |
| **Taq Polymerase** | 5 U/µl | 0.2 µl |
| **$H_2O$** | - | 13.8 µl |
| **Total** | - | 20.0 µl |

**Table 2.15: Standard PCR conditions.**

| Cycling Condition | | Temperature | Time |
|---|---|---|---|
| **Step 1: Initial denaturation** | | 94°C | 2 min |
| **Step 2: Denaturation** | 35 Cycles | 94°C | 30 sec |
| **Step 3: Annealing** | | 55°C | 30 sec |
| **Step 4: Extension** | | 70°C | 30 sec |
| **Step 5: Extension** | | 70°C | 5 min |
| **Step 6: Hold** | | 10°C | ∞ |

The PCR product for each sample was extracted from a window gel (see Section 2.4) before being sequenced (using the same protocol as in Section 2.4) in order to identify whether each sample was homozygous or heterozygous for the rare allele.

## 2.4 Determining structural variations of glycophorin genes

### 2.4.1 Detecting CNV of glycophorin genes using high throughput sequencing data

The binary alignment map (BAM) files from the 1000 Genomes Project data website (ftp://ftp.1000genomes.ebi.ac.uk/) for most of the 1000 Genomes Project samples were downloaded. Following this, a sequence read depth of 5 kb windows was calculated from the sample BAM files using Samtools software, which is a set of utilities that manipulates alignments in the BAM format (Appendix 7). Samtools software that can import from and export to the SAM (Sequence Alignment Map) format. It also enables sorting, merging and indexing, and allows reads in any regions to be retrieved (Li *et al*., 2009; Li, 2011).

However, the ratio between the read and the reference (CN/Ref) was taken for each window. Thus, a 5 kb window reads plot for these samples was created from the generated sequence read depth of a 5 kb windows file using R program (Appendix 8). R is a computing language that also works as powerful statistical software. It is used to produce graphs and tables, read and save text files, and to assign and manipulate variables (Ihaka and Gentleman, 1996; Tabelow *et al*., 2011). The R program was installed (version 3.3.1) on one of the Linux-based powerful high performance computing clusters called SPECTRE (Special Computational Teaching and Research Environment) provided by the University of Leicester for staff and students (https://www2.le.ac.uk/offices/itservices/ithelp/services/hpc/spectre/about).

### 2.4.2 Confirmation of detected CNVs of glycophorin genes using fibre-Fluorescence *In Situ* Hybridization

Fibre-Fluorescence *in situ* hybridisation (fibre-FISH) is a laboratory technique used to detect and physically map specific DNA sequences on naked DNA fibres (Ersfeld, 2004). The technique involves the use of a small DNA sequence called a probe, which has a fluorescent molecule attached to it. The fluorescently labelled probe will only hybridise to the complementary DNA sequences for a specific region of interest. A particular advantage of this technique is the option to use probes, each labelled with a different fluorescent dye or a combination of multiple dyes, in one hybridisation solution. This therefore makes the process more efficient as it allows multiple targets to be simultaneously visualized in a single sample (Raap, 1998). This enables DNA rearrangements to be visualised, which it is possible to examine directly using microscopy (Ersfeld, 2004; Raap, 1998). In this study, all fibre-FISH experiments were performed by Dr Sandra Louzada in the laboratory of Dr Fentang Yang at the Wellcome Trust Sanger Institute in order to visualise the evolutionary chromosome rearrangements of the glycophorin variations.

#### 2.4.2.1 Preparation of fibre-FISH probes

All fibre-FISH probes used in this study were either derived from purified fosmid clones or probes generated by long-range PCR amplification. Fosmid clones covering the glycophorin genes were obtained from the Whitehead Institute's random genomic fosmid library (WIBR-2) spanning the region (International Human Genome Sequencing Consortium, 2004). This was based on the availability of fosmid end sequences from the WIBR-2 library on the mapped human reference genome GRCh37 (Church *et al*., 2011). Because of the extensive cross-hybridization of probes that map the tandemly repeated glycophorin regions, the *GYPE* repeat was distinguished using a small *GYPE*-repeat-specific PCR amplified probe. This probe was generated using a specific *GYPE* primer pair (Table 2.16) to amplify a 3,632 bp product from the *GYPE* gene.

**Table 2.16: A specific *GYPE* primer pair.** This primer pair was used to generate a *GYPE* specific probe that was used with fibre-FISH to characterise the structural model of the DUP5 variant. The primer's optimisation and the long-PCR conditions are detailed in subsection 2.4.3.

| Primer name | Sequence |
|---|---|
| **Specific_*GYPE*_ Forward** | 5' TGTATTCCAGTTGGTGGATATTTC 3' |
| **Specific_*GYPE*_ Reverse** | 5' ATGGGCTTTCAAGAACATTTCT 3' |

To convert these products into fibre-FISH probes, approximately 10 ng of purified long-range PCR products were first amplified using a GenomePlex® Complete Whole Genome Amplification Kit (WGA2) (Sigma-Aldrich) following the manufacturer's protocols. The WGA2 amplified products were then labelled using a modified GenomePlex re-amplification kit (WGA3) (Sigma-Aldrich), following the manufacturer's protocols, with a custom-made dNTP mix (Gribble *et al*., 2013). Probes were labelled with Biotin-16-dUTP, Digoxigenin-11-dUTP and DNP-11-dUTP (Jena Bioscience, Jena, Germany) (Table 2.17). On amplification, the WGA-labelled probes were DNase I (Roche) digested in the ratio of 1:15 enzyme to labelling reaction to reveal a size between 200-500 bp. A total of 500 ng of labelled DNA from each probe was ethanol precipitated before being re-suspended in a hybridization buffer containing (2 × SSC, 10% sarkosyl, 2 M NaCl, 10% SDS, and blocking aid) and a 1:1 deionized formamide (final concentration 50%). All labeled dUTPs were purchased from Jena Bioscience.

**Table 2.17: Table shows all of the used fibre-FISH probes and their details.**

| Probe | Probe name | Chromosomal coordinates | Size | Dye |
|---|---|---|---|---|
| F1 | G248P86579F1 | chr4:144,845,018-144,888,868 | 43,851 bp | digoxigenin-11-dUTP |
| *E* PCR | *GYPE* specific | chr4:144,741,272-144,744,903 | 3,632 bp | digoxigenin-11-dUTP |
| G10 | G248P8211G10 | chr4:145,105,022-145,147,081 | 42,060 bp | biotin-16-dUTP |
| F12 | G248P85804F12 | chr4:144,898,743-144,937,429 | 38,687 bp | DNP-11-dUTP |
| F7 | G248P80757F7 | chr4:144,691,547-144,730,867 | 39,321 bp | Cy5-dUTP |

### 2.4.2.2 Pre-fibre-FISH hybridization

Lymphoblastoid cells were grown from the known glycophorin variant samples from the 1000 Genomes Project sample collections, which derived from different ancestries. Single-molecule DNA fibres were prepared using a molecular combing system (MCS) that pulled and stretched DNA molecules into vinylsilane-coated coverslips (Michalet *et al*. 1997) following the manufacturer's instructions. Briefly, one million cells were embedded in agarose plugs. These plugs were vertically treated with ESP solution (0.5 M EDTA, 10 % sarkosyl, proteinase K) and incubated overnight in a 50°C water bath (WB). Following incubation, plugs underwent serial washes with 1X TE solution (10 mM Tris, 1 mM EDTA, PH 8.0). Washed plugs were transferred into a microtube using a spatula, following which proteins were digested with 10 ul of B-agarase enzyme (New England Biolabs, Ipswich, MA, USA). A pre-warmed 0.5 M MES buffer (PH 5.5) was added to the plugs to release the DNA and the tube was gently inverted and placed in a WB overnight to homogenise the solution.

Coverslips were inserted in the MCS holder, dipped in the high molecular weight DNA solution and slowly pulled at a constant speed to ensure the uniform stretching of the DNA onto the coverslips. Coverslips coated with combed DNA fibres were baked at 68°C for four hours.

#### 2.4.2.3 Post-fibre-FISH hybridization

The probe mix was denatured at 65°C for 10 minutes before being applied onto the coverslips. Coverslips coated with combed DNA fibres were dehydrated through a 70%, 90%, and 100% ethanol series for three minutes followed by air-drying at RT. Alkaline denaturation solution (0.5 M NaOH, 1.5 M NaCL) was performed followed by serial washing with 1X PBS, and the hybridization was carried out in a 37°C humid hybridization chamber overnight. The post-hybridization washes consisted of two rounds of washes in 50% formamide, followed by two additional washes in 2 × SSC at 25°C, five minutes each time. Digoxigenin-11-dUTP labelled probes were detected using a 1:100 dilution of monoclonal mouse anti-digoxin antibody (D8156, Sigma-Aldrich) and a 1:100 dilution of goat anti-mouse Texas Red (T6390, Thermo Fisher Scientific). DNP-11-dUTP labelled probes were detected using a 1:100 dilution of rabbit anti-DNP (SP-0603, Vector Laboratories) and donkey anti-rabbit Alexa Fluor 488 (A-21206, Thermo Fisher Scientific). Biotin-16-dUTP labelled probes were detected with 1:100 of Streptavidin Cy3 (S6402, Sigma-Aldrich) and anti-streptavidin conjugated with CF543 (SP-400, Vector Laboratories). After hybridization, slides were mounted using SlowFade Gold antifade mounting solution (Thermo Fisher Scientific) and the slide was sealed with nail vanish. Using an epifluorescence microscope equipped with 40x immersion oil, objective digital images were visualized to capture signals from FITC, Cy3 and Texas Red using SmartCapture software (Digital Scientific, Cambridge, UK) via a cooled CCD camera (Hamamatsu, ORCA-EA).

### 2.4.3 Identification of glycophorin structural variant breakpoints using wet-lab approaches

#### 2.4.3.1 DNA positive controls

The NGS reads of the glycophorin genes were used to design a primer pair for unknown glycophorin variants and to identify a positive control for each variant (Table 2.18).

**Table 2.18: Positive controls information of glycophorin variants.** These DNA samples were ordered from Coriell Cell Repositories, 403 Haddon Avenue, Camden, New Jersey, 08103, USA.

| DNA Sample | Sample project | Sample origin | Variant |
|---|---|---|---|
| **NA19223** | HapMap Project and 1000 Genomes Project | Yoruba in Ibadan, Nigeria | DEL1 |
| **NA19144** | HapMap Project and 1000 Genomes Project | Yoruba in Ibadan, Nigeria | DEL2 |
| **HG04039** | 1000 Genomes Project | Sri Lankan Tamil in the UK | DEL6 |
| **HG02716** | 1000 Genomes Project | Gambian in western division, the Gambia | DEL7 |
| **NA18625** | HapMap Project and 1000 Genomes Project | Han Chinese in Beijing, China | DUP2 |
| **NA19474** | HapMap Project and 1000 Genomes Project | Luhya in Webuye, Kenya | DUP3 |
| **HG02554** | 1000 Genomes Project | African ancestry from Barbados in the Caribbean | DUP4 |
| **HG02585** | 1000 Genomes Project | Gambian in western division, the Gambia | DUP5 |
| **HG02679** | 1000 Genomes Project | Gambian in western division, the Gambia | DUP7 |
| **NA18646** | HapMap Project | Han Chinese in Beijing, China | DUP14 |
| **HG03686** | 1000 Genomes Project | Sri Lankan Tamil in the UK | DUP29 |
| **HG03837** | 1000 Genomes Project | Sri Lankan Tamil in the UK | DUP24 |
| **NA11994, NA12006 and NA12716** | HapMap Project and CEPH/UTAH pedigree 1362, 1420 and 1420 | Utah residents with ancestry from northern and western Europe | *GYPE-B-E* Gene conversion |

### 2.4.3.1 Primer design

According to the 5 kb window NGS reads data and the fibre-FISH results, the expected duplication and deletion and gene conversion area sequences were aligned with the very similar sequences on the glycophorin genes. Based on these expectations and alignments, a specific LNA primer pair for each variant was designed to sequence the PCR products and identify the exact breakpoints for each variant (Table 2.19). However, the expected gene conversion allele is (*GYPE-GYPB-GYPE*), which means that a DNA sequence from *GYPE* gene has been replaced by a DNA sequence from *GYPB* gene.

**Table 2.19: Specific primer pairs for glycophorin variants.** The LNA base is shown between brackets.

| Variant | Primer name | Primer Sequence |
|---|---|---|
| **DEL1** | *GYP*_DEL1_F | 5' CCAGTTGCCTCTAAGTCCAT[C] 3' |
| | *GYP*_DEL1_R | 5' GCAGTGCACACCCTGG[A] 3' |
| **DEL2** | *GYP*_DEL2_F | 5' AGGCAAAAGCTGAGGTCTT[C] 3' |
| | *GYP*_DEL2_R | 5' CAGCCTCTGGTAACCACTGTTA[C] 3' |
| **DEL6** | *GYP*_DEL6_F | 5' GAAGAAAGAGCTAATTCCAT[G] 3' |
| | *GYP*_DEL6_R | 5' AGTTGGAACTTGCAAACTTA[G] 3' |
| **DEL7** | *GYP*_DEL7_F | 5' ATCCTGCACTAGAAATTCCTCCCA[C] 3' |
| | *GYP*_DEL7_R | 5' GATCAGAAAAGCAAAATGGGGC[A] 3' |
| **DUP2** | *GYP*_DUP2_F | 5' GATTTTAAATGCCTTGCTTTTTA[T] 3' |
| | *GYP*_DUP2_R | 5' CTGGATAGTACAGAGCCAGG[A] 3' |
| **DUP3** | *GYP*_DUP3_F | 5' CAAATGAAGTCAAACATCTTC[A] 3' |
| | *GYP*_DUP3_R | 5' CTTGAGACACTCCTTTATATGCTA[C] 3' |
| **DUP5** | *GYP*_DUP5_F | 5' AGCTTGGATGAGATAAATGTCC[T] 3' |
| | *GYP*_DUP5_R | 5' ATTGGATTCTGATGTGCGG[C] 3' |
| **DUP7** | *GYP*_DUP3_F | 5' CAAATGAAGTCAAACATCTTC[A] 3' |
| | *GYP*_DUP3_R | 5' CTTGAGACACTCCTTTATATGCTA[C] 3' |
| **DUP14** | *GYP*_DUP14_F | 5' GTCTTTAAAGTATTGTTTCGTGC[A] 3' |
| | *GYP*_DUP14_R | 5' AGGTTAATCTAAAACTTTAGAGCAA[C] 3' |
| **DUP29** | *GYP*_DUP29_F | 5' GCTGCCAGATCAATAGC[G] 3' |
| | *GYP*_DUP29_R | 5' TAGTAGTATAAACCACAGTGCCTC[A] 3' |
| **DUP24** | *GYP*_DUP8_F | 5' ACTCAGAGGAATAAACCCTC[T] 3' |
| | *GYP*_DUP8_R | 5' AACCCAAATTATTATATGTAAGC[T] 3' |
| ***GYPE-B-E*** | *GYPE*_GC_F | 5' TACATGGGAATACAACTGGAAAA[G] 3' |
| **Gene conversion** | *GYPE*_GC_R | 5' TCTCTCAATGACAACTTACTTGATTC[T] 3' |

### 2.4.3.2 Long-PCR

A very large product (≥ 4 Kb) was expected for each specific variant assay. A single long-PCR conditions were therefore designed and optimised using a Veriti® Thermal Cycler PCR machine on each specific variant assay (Table 2.20 and 2.21).

**Table 2.20: Long-PCR mixture for any glycophorin variant-specific assay except DUP2 variant.**

| Reagent | Concentration | Volume per reaction |
|---|---|---|
| **Genomic DNA** | 10 ng/µl | 1.0 µl |
| **Forward Primer** | 10 µM | 0.5 µl |
| **Reverse Primer** | 10 µM | 0.5 µl |
| **11.1 X Buffer** | - | 2.25 µl |
| **pfu** | 2.5 U/µl | 0.03 µl |
| **Taq Polymerase** | 5 U/µl | 0.125 µl |
| **$H_2O$** | - | 20.6 µl |
| **Total** | - | 25.5 µl |

**Table 2.21: A long-PCR conditions of a glycophorin variant specific assay.** This long-PCR conditions is for the Del6_specific assay. The long-PCR has an additional stage of thermal cycles compared with normal PCR and in Step 6, the [A] means a 15s cycle elongation for each successive cycle.

| Cycling Condition | | Temperature | Time |
|---|---|---|---|
| **Step 1: Initial denaturation** | | 94°C | 1 min |
| **Step 2: Denaturation** | 20 Cycles | 94°C | 15 sec |
| **Step 3: Annealing** | | 63°C | 10 min |
| **Step 4: Denaturation** | 15 Cycles | 94°C | 15 sec |
| **Step 5: Annealing** | | 63°C | 10 min [A] |
| **Step 6: Extension** | | 72°C | 10 min |
| **Step 7: Hold** | | 10°C | ∞ |

The optimised annealing temperature for each primer pair is summarised in (Appendix 2A). However, in the case of DUP2, a Q5 hot start high-fidelity kit was used because it cannot be optimised with the 11.1 X Buffer (Table 2.22). The Q5 hot start high-fidelity kit is manufactured by New England Biolabs (NEB). The kit's master mix includes a modified polymerase enzyme that only works at a high initial denaturation temperature to increase the specificity of the PCR reaction (Table 2.23).

**Table 2.22: Long-PCR mixture for the DUP2-specific assay using Q5 hot start high-fidelity kit.**

| Reagent | Concentration | Volume per reaction |
|---|---|---|
| **Genomic DNA** | 10 ng/µl | 1.0 µl |
| ***GYP*_DUP2_F** | 10 µM | 1.25 µl |
| ***GYP*_DUP2_R** | 10 µM | 1.25 µl |
| **Q5 Hot Start High-Fidelity Master Mix** | - | 12.5 µl |
| **$H_2O$** | - | 9.0 µl |
| **Total** | - | 25 µl |

**Table 2.23: Long-PCR conditions for the DUP2 variant specific assay using a Q5 hot start high-fidelity kit.** This long-PCR condition is different from other conditions. It has a high initial denaturation temperature to increase the specificity of the polymerase enzyme. The [A] means a 15s cycle elongation for each successive cycle.

| Cycling Condition | | Temperature | Time |
|---|---|---|---|
| **Step 1: Initial denaturation** | | 98°C | 1 min |
| **Step 2: Denaturation** | 20 Cycles | 94°C | 15 sec |
| **Step 3: Annealing** | | 64°C | 10 min |
| **Step 4: Denaturation** | 15 Cycles | 94°C | 15 sec |
| **Step 5: Annealing** | | 64°C | 10 min [A] |
| **Step 6: Extension** | | 72°C | 10 min |
| **Step 7: Hold** | | 10°C | ∞ |

### 2.4.3.3 Purification of PCR products from agarose gels

Purification of PCR products from agarose gels was done using Lee *et al*. (2012) protocol of DNA purification for Sanger sequencing. Therefore, a 1% agarose gel was prepared using a 1xTAE buffer (Appendix 1) instead of a 1xTBE buffer (see Section 2.2). Once the gel had solidified, an appropriate size "window" was cut and removed. Following this, the window was filled by a 0.2% agarose gel, also using the 1xTAE buffer, and allowed to set (Figure 2.8).

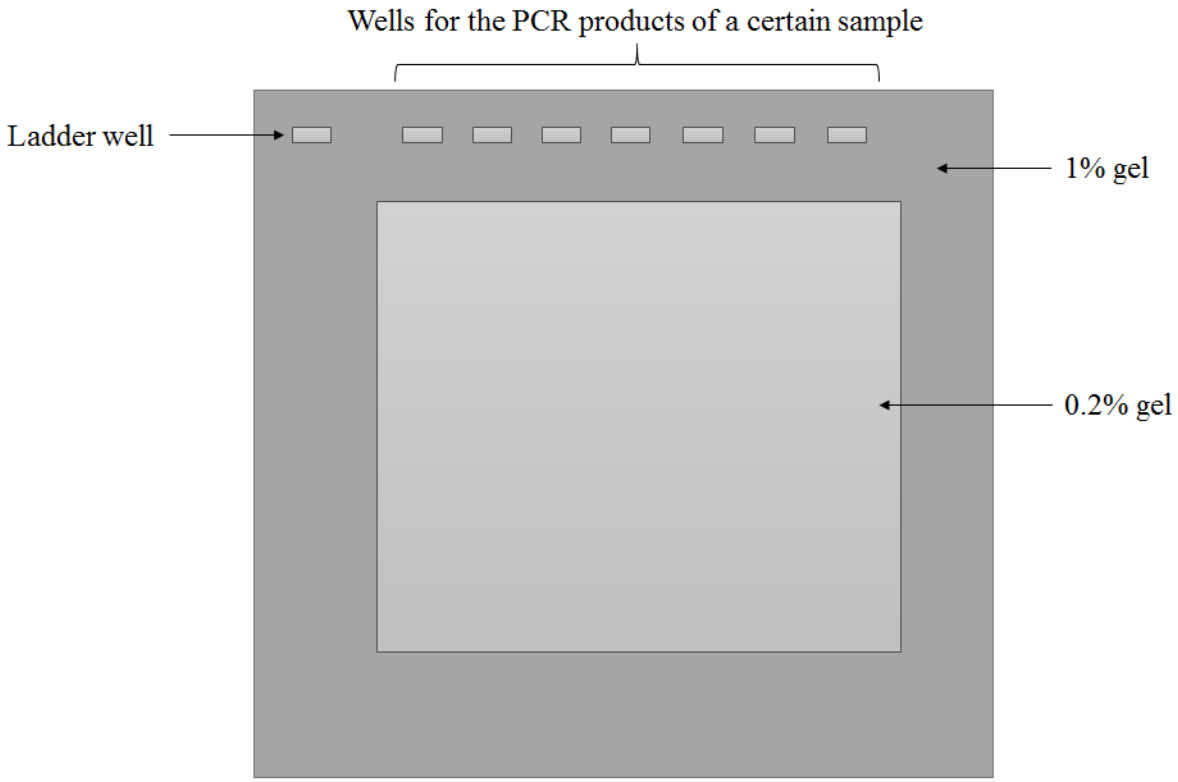



**Figure 2.8: Diagram of a window gel.** The dark section indicates the 1% gel made using the 1xTAE buffer. The lighter section shows where a square was cut out of the 1% gel, and a 0.2% gel poured into this window and left to set. The small top squares show the wells where the samples are placed in the gel.

Once the "window" has dried, the products for a certain sample were run in this window gel until the PCR products were in the 0.2% "window" using the electrophoresis setting at 100V for 90 minutes. Following this, the bands (PCR products) were extracted from the gel using a scalpel under blue light and placed into an Eppendorf tube, which was frozen overnight. The sample was then defrosted at room temperature for 10 minutes, and then centrifuged at 13200rpm for 30 minutes. Once complete, the supernatant was extracted.

#### 2.4.3.4 Concentration stage for extracted PCR products

Different volumes of the extracted supernatant (PCR products) were loaded onto a 1% agarose gel using 1xTBE buffer (Section 2.2), with the same volumes of 1kb hyper ladder. The 1kb hyper ladder provides estimated concentration, which is dependent on the size of the product. The electrophoresis was set to 100V for 90 minutes and an image of the gel was obtained using the UV transilluminator. Visual analysis was conducted to determine which band matched the intensity of the 1kb hyper ladder. This then allowed the DNA concentration (ng/μl) to be estimated.

Concentration of the sample product was carried out using the standard procedures outlined for the Amicon® Ultra-0.5 Centrifugal Filter Devices (Merckmillipore.com, 2016). This kit works based on a centrifugation series in which the DNA is isolated and other components that may still be in the sample from the gel are filtered through. The concentration was calculated based on the previously found concentration of the sample from visual gel analysis and the volume of the sample after the concentration process. Equation: concentration of whole sample ng / volume of sample after concentration = final concentration of sample ng/μl.

However, only 1 μl of the specific *GYPE* product that had been extracted and concentrated was used to assess the specificity of the primer to *GYPE* by Sanger sequencing. The rest of the specific *GYPE* product was placed into a tube and sent off to colleagues in the Sanger Institute to be used as a probe for fibre-FISH analysis in order to characterise DUP5.

#### 2.4.3.5 Sanger sequencing

There were two stages involved in Sanger sequencing. The first stage was cycle sequencing and the second capillary electrophoresis. BigDye v3.1 was used for the cycle sequencing stage. This was conducted using the latest Applied Biosciences (2016) standard procedures. However, guidelines state that for PCR product with a size > 2 Kb, the quantity of DNA should be 20-50 ng; the volume therefore varies dependant on the concentration of the PCR product obtained (Table 2.24). The thermal cycling conditions used for sequencing are described in Table 2.25. The reaction was conducted using a Veriti® Thermal Cycler PCR machine.

**Table 2.24: BigDye sequencing reaction mixture.** The table defines the reagents and volumes required for a single sequencing reaction.

| Reaction components | Concentration | Volume per reaction |
|---|---|---|
| **PCR Product** | 20–50 ng/µl | 1.0 µl |
| **Primer** | 3.2 µM | 1.0 µl |
| **BigDye Terminator Ready reaction Mix** | - | 2.0 µl |
| **5X BigDye Terminator buffer** | - | 1.0 µl |
| **$H_2O$** | - | 5.0 µl |
| **Total** | - | 10.0 µl |

**Table 2.25: Thermal cycle conditions for Sanger sequencing.**

| Cycling Condition | Temperature | Time |
|---|---|---|
| **Step 1: Initial denaturation** | 96°C | 1 min |
| **Step 2** (26 Cycles) | 96°C | 10 sec |
| **Step 3** (26 Cycles) | 50°C | 5 sec |
| **Step 4** (26 Cycles) | 60°C | 4 min |
| **Step 5: Hold** | 4°C | ∞ |

To purify the PCR product, the denaturation method was used. Firstly 10µl of water was added to the reaction along with 2µl of 2.2% SDS, making the total reaction volume 22µl. The reaction was then placed into the Veriti® Thermal Cycler PCR machine, where it was heated at 98ºC for five minutes and then 25°C for 10 minutes. In order to remove the SDS from the reaction, EdgeBio Performa® DTR gel filtration cartridges were used. First, the filtration columns were centrifuged at 3200 rpm for three minutes to remove any liquid in the tube. The sample was then loaded into the column and centrifuged at 3200 rpm for three minutes, after which stage the eluted DNA was present in the tube. The sample was labelled and sent off for Level 3 sequencing at the University of Leicester PNACL service, where capillary electrophoresis was performed and a chromatogram showing the results of the sequencing reaction produced. Each primer is detailed in (Appendix 2B).

### 2.4.4 Sequence alignment tools

1) The UCSC genome browser is an online genome browser created by the University of California, Santa Cruz (UCSC). The UCSC genome browser (GRCh37/hg19) provides users with many useful tools on its website (https://genome.ucsc.edu/index.html), such as the BLAST Like Alignment Tool (BLAT), which rapidly aligns sequences to the genome, and In-Silico PCR, which rapidly aligns PCR primer pairs to the genome (Kent *et al.*, 2002).

2) The European Molecular Biology Open Software Suite (EMBOSS-Needle) is high-quality molecular biology software that can read two input sequences and write their optimal global sequence alignment to file. This open software was accessed on the (EMBL-EBI) website (https://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html). Default pairwise alignment options were applied (Appendix 3).

3) Multiple Alignment using Fast Fourier Transform (MAFFT) is a high-speed multiple sequence alignment program (Katoh and Standley, 2013). This open software can be used by accessing the EMBL-EBI website (https://www.ebi.ac.uk/Tools/msa/mafft/). Default pairwise alignment options were applied, but the output format of the alignment was changed from Pearson/FASTA to Clusatl W (Appendix 3), which is a general purpose DNA or protein multiple sequence alignment program for three or more sequences (Larkin *et al*., 2007).

EMBL-EBI stands for the European Molecular Biology Laboratory (EMBL) - European Bioinformatics Institute (EBI). It is an academic research institute based in the UK, and is part of the European Molecular Biology Laboratory.

## 2.5 Analysis of glycophorin gene conversion allele

### 2.5.1 Restriction fragment length polymorphism (RFLP)

Restriction Fragment Length Polymorphism (RFLP) is a technique that can differentiate between genomic DNAs by analysing patterns derived from the cleavage of their sequences (Bhardwaj *et al*., 2003). This technique includes several stages: PCR reaction stage, RFLP digestion with a restriction enzyme, incubation and gel electrophoresis (Old, 2003). The restriction enzyme was chosen from New England Biolabs (NEB) after using the expected gene conversion PCR product sequence with online software called NEBcutter V2.0 (http://nc2.neb.com/NEBcutter2/). This tool takes a DNA sequence and identifies the restriction sites within the DNA sequence. Basically, the *HinDIII* cuts at the A/AGCTT sequence. However, after selecting a suitable restriction enzyme and buffer, 1.0 µl of the *HinDIII* restriction enzyme (Conc. 20 U/µl), 2.5 µl of NEBuffer2 B7002S (an enzyme buffer) (50 mM NaCl, 10 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.9), 0.2 µl of BSA (Conc. 10 mg/ml) and 6.3 µl of sterile distilled water were added to 10 µl of each PCR product in a total volume of 20 µl, followed by overnight incubation at 37°C. The digested PCR products were run on a 0.8% agarose gel using a 1xTBE buffer (Section 2.2) to visualise the result of the digestion.

## 2.6 Malaria cohorts and statistical association analysis

### 2.6.1 Copy number analysis of glycophorin genes for Benin malaria cohort

PRT assays (2.2) were applied on the Tori-Bossito (Benin) cohort samples and the scatterplot was created (as described in 2.2) for each assay representing the estimated glycophorin genes copy number data for the Benin cohort. Cluster analysis software was used in the R program to differentiate between homozygous and heterozygous samples and different variants (Appendix 9). This powerful cluster analysis software is able to summarise data by grouping observations (clusters) that share similar values for a number of variables (Everitt, 2004). In order to express the confidence of the PRT assays in glycophorin copy number calling and to measure the integer copy number of glycophorin of the Benin samples, a histogram and a fitted text file were created for each PRT assay using CNVtools in the R program (Appendix 10). CNVtools is a software programme that can be used to perform robust case-control and quantitative trait association testing of CNVs (Barnes *et al*., 2008). This malaria cohort data and information were used with the data from the PRT assays to carry out statistical association analysis.

### 2.6.2 Analysis of glycophorin variants and malaria clinical traits

1) Analysis of glycophorin variants and parasite density:

A linear mixed model was constructed using SPSS 20.0 (IBM) to analyse the association in the Tori-Bossito cohort between parasite density following infection and the covariates (mother's age, suspected malaria during pregnancy, sex, birth term, mosquito net, ethnic group, sickle cell and chloroquine intake during pregnancy).

2) Analysis of glycophorin variants and number of malarial infections:

A Poisson regression was constructed using SPSS 20.0 (IBM) to analyse the association in the Tori-Bossito cohort between the number of breakthrough infections and the covariates (mother's age, suspected malaria during pregnancy, sex, birth term, mosquito net, ethnic group, sickle cell and chloroquine intake during pregnancy).

3) The effect of glycophorin variants on Time to First Malarial Infection:

Cox regression, also known as proportional hazards regression, is a method used to investigate the effect of several variables on the time a specified event takes to happen. This regression analysis was carried out using SPSS 20.0 (IBM) to assess the effect of the factors (mother's age, suspected malaria during pregnancy, sex, birth term, mosquito net, ethnic

group, sickle cell and chloroquine intake during pregnancy) on the risk of contracting malaria in the Tori-Bossito cohort.

### 2.6.3 Linkage disequilibrium analysis.

The linkage disequilibrium analysis was completed for different glycophorin variants and the SNP using (http://www.oege.org/software/cubex/).

### 2.6.4 DUP4 genotyping by a specific PCR assay

The structural model of DUP4a and DUP4b variants were confirmed by fibre-FISH using the positive control (HG02554) of DUP4 (Section 2.4). A specific PCR assay for the DUP4 variant was designed using the information described in the Introduction regarding the duplication breakpoints from Leffler *et al*. (2017) and was applied to the Tanzanian malaria cohort. However, the primers were designed by Paulina Brajer, an MSc student in the lab of Dr Edward Hollox (Table 2.26) to be specific for one of the DUP4 (both DUP4a and DUP4b) breakpoints.

**Table 2.26: The DUP4 specific PCR assay primer pair.** The LNA bases are between brackets.

| Primer name | Sequence |
|---|---|
| **DUP4_F2** | 5' GCAAAAGCTGAGGTCTT**[C]** 3' |
| **DUP4_R2** | 5' ATATAATATTCAATGTGCTAAGG**[A]** 3' |

The DUP4 specific assay is a duplex PCR that uses the (Seq_SNP_WA) primer pair as a positive control for PCR success with the main primer pair of the assay. This primer pair was previously designed for another purpose, as described in Section 2.3. The DUP4-specific assay was optimised using a Veriti® Thermal Cycler PCR machine (Tables 2.27 and 2.28) and applied to the genotype of 338 Tanzanian DNA samples.

**Table 2.27: DUP4 specific assay PCR mixture.**

| Reagent | Concentration | Volume per reaction |
|---|---|---|
| **Genomic DNA** | 10 ng/µl | 1.0 µl |
| **DUP4_F2** | 10 µM | 0.8µl |
| **DUP4_R2** | 10 µM | 0.8 µl |
| **Seq_SNP_WA_Forward** | 10 µM | 0.1 µl |
| **Seq_SNP_WA_Reverse** | 10 µM | 0.1 µl |
| **×10 Kapa A Buffer** | - | 1.0 µl |
| **dNTPs** | 2.5 µM | 0.8 µl |
| **Taq Polymerase** | 5 U/µl | 0.2 µl |
| **$H_2O$** | - | 5.2 µl |
| **Total** | - | 10.0 µl |

**Table 2.28: DUP4 specific assay PCR conditions.**

| Cycling Condition | | Temperature | Time |
|---|---|---|---|
| **Step 1: Initial denaturation** | | 95°C | 2 min |
| **Step 2: Denaturation** | 35 Cycles | 95°C | 30 sec |
| **Step 3: Annealing** | | 58°C | 30 sec |
| **Step 4: Extension** | | 70°C | 30 sec |
| **Step 5: Extension** | | 70°C | 5 min |
| **Step 6: Hold** | | 10°C | ∞ |

The PCR products were run on 2% (1x TBE buffer) agarose gel electrophoresis to visualise the result of the assay and identify the positive individuals for DUP4 variant.

### 2.6.5 Homozygous/heterozygous analysis and statistical association analysis of DUP4 in the Tanzanian cohort

#### 2.6.5.1 Measuring intensity of gel images using ImageJ

In terms to check the homozygous and heterozygous state of samples are carrying the DUP4 variant, a quantification of the gel was conducted using computer analysis with image J software. This Java-based image processing software can calculate area and pixel value statistics for user-defined selections and intensity-thresholded objects, measure distances and angles, and create density histograms and line profile plots (Schneider *et al*., 2012).

However, this analysis involved peak area measurement of each band based on the band intensity using the electronic images for the gels. The analysis was and applied to each sample that was positive for DUP4 (Schneider *et al*., 2012). The peak area of each band was

measured and the peak of DUP4 was divided by the peak of the positive primer pair PCR product. The result from each set of samples was plotted on a histogram. The data set provided by Dr Edward Hollox, containing the phenotype information for the Tanzanian samples and the data from the homozygous and heterozygous analysis, was used for additional statistical analysis.

#### 2.6.5.2 Quantitative transmission disequilibrium test (QTDT)

The quantitative transmission disequilibrium test (QTDT) is a family-based analysis that uses parental information to test the association between phenotypes and alleles at a candidate locus. QTDT can include covariates in the analysis, such as sex and age, which are likely to contribute to variation within the quantitative phenotype (Abecasis *et al*., 2000). The QTDT output is a chi-squared value with 1 degree of freedom for each allele tested, when a significant number (≥30) of individuals are present with that allele. The direction of allelic association can be established by observing the parameter W (the observed pattern of allelic association) from the full-model specified in the output file (regress.tbl by default).

Associations between the quantitative epidemiological, parasite and serological phenotypes for malaria and the DUP4 genotype at the glycophorin genes locus were tested using QTDT version 2.6.1 (Abecasis *et al*., 2000). Total evidence of association was modelled including age and sex as covariates and environmental and polygenic heritability as additive major locus variance components (Algady *et al*., 2018). The commands used for the test are shown in appendix 11.

# Chapter 3: Identification and characterization of glycophorin variant breakpoints

## 3.1 Using sequence read depth data of the 1000 Genome Project samples to detect glycophorin variants

The primary aim of this chapter is to analyse the glycophorin variants and identify their breakpoints in order to discriminate and study each variant and investigate if the breakpoint occurred in an actual glycophorin gene and whether it occurred in an exon or intron. The breakpoints allow prediction of the biological consequences of the glycophorin variants from their expected positions. The glycophorin genes are responsible for the MNS blood groups and Miltenberger antigens, some of the glycophorin variants could be related to some of the MNS and Miltenberger antigens, so the breakpoints of a variant can confirm whether it is responsible for any of these blood groups. For example, although the Miltenberger antigens GP.Hil and GP.JL carry a hybrid *GYPA-B* gene and both have neither entire *GYPA* nor *GYPB* gene, the only difference between them is the breakpoint.

The GP.Hil breakpoint is at the 5' end of intron 3 of *GYPA*, while GP.JL breakpoint is at the 5' end of intron 3 of *GYPA* plus 7 nucleotides of exon 4 of *GYPB* (Reid, 2009; Velliquette *et al*., 2008). Knowledge of the breakpoint locations can be used to design a specific PCR assay for a certain glycophorin variant in order to genotype a large cohort and analyse the frequency of a variant or analyse if there is any association of a particular glycophorin variant and susceptibility to malaria.

The breakpoint analysis were done by creating a 5 kb windows plot for each glycophorin variant using the sequence read depth data of the glycophorin regions of the 1000 Genomes Project samples where are there identified as positive controls, then each variant was confirmed using the variant positive controls for fibre-FISH (with thanks to Sandra Louzada and Fengtang Yang from Wellcome Sanger Institute). In addition, the 5 kb windows plot for a set of normal 1000 Genomes Project samples was created (Figure 3.1) in order to compare it with the deletion and duplication plots. This chapter shows how long-PCR and Sanger sequencing were used to identify the glycophorin variant breakpoints.

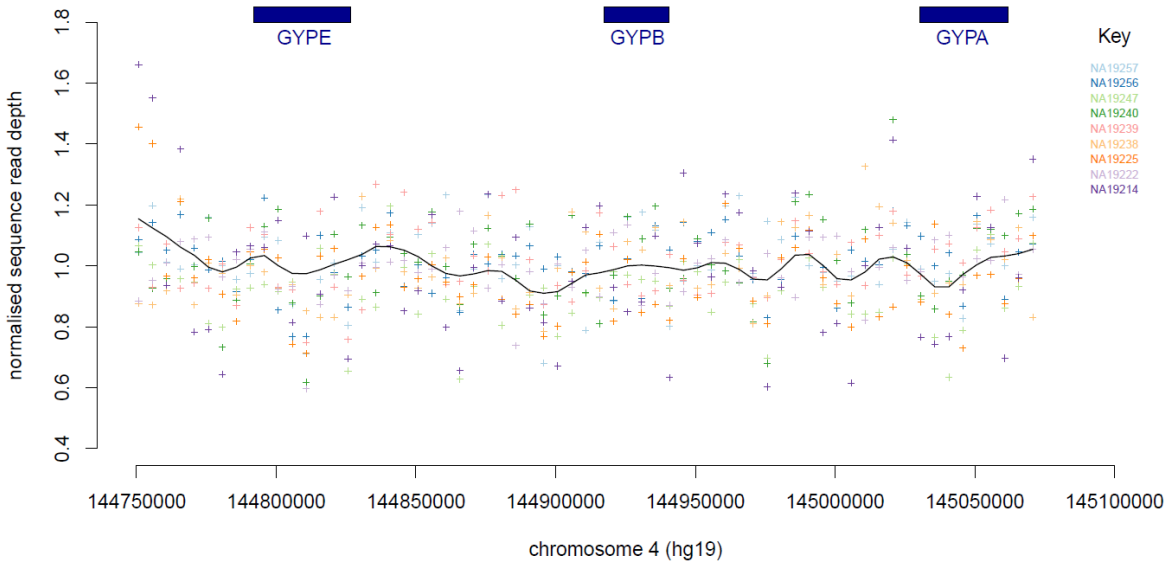



**Figure 3.1: Normal glycophorin genes plot generated using next-generation sequencing data.** The points on the graph indicate the normalised read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

The sequence read depth is the number of sequence reads that uniquely map to a genomic region. The general concept of deep sequencing is to achieve a high number of unique reads for each region of a sequence. The sequence read depth approaches can reflect copy number (Medvedev *et al.*, 2009).

The 1000 Genome Project samples were used because this was the first project to sequence the genomes of a large number of DNA samples from different populations to provide a comprehensive resource for studies of human genetic variation (1000 Genomes Project Consortium *et al.*, 2010). Data from the 1000 Genomes Project are available for the worldwide scientific community via freely accessible public databases. The sequence alignment BAM files from the 1000 Genomes Project samples were downloaded for the candidate glycophorin deletions and duplications, which show low and high normalised NGS read depth data values respectively. The sequence read depth across the glycophorin regions was calculated and divided by a reference region using samtools software version 1.8 (Li *et al.*, 2009) into 65 windows, each with represents 5 kb region (Chapter 2 and Appendices 7 and 8). The reads ratio of each 5 kb window were normalized and each window was corrected to a value of one; therefore, values lower than one indicate a deletion and values higher than one indicate a duplication.

Long-PCR and Sanger sequencing were used because the positive control of each glycophorin variant is known and the sequence read depth data gives an estimation of the breakpoints so the long-PCR primers could be designed. Moreover, the primers were designed to be paralogue-specific and could be confirmed by Sanger sequence of the long-PCR product. Breakpoints of these variants are very important to be exactly known for any further analysis. For example, if all of the glycophorin variant breakpoints are identified, the matched breakpoint of the variants can be determined. In addition, the variant breakpoints can be utilised for designing a standard specific PCR assay for a certain variant in order to apply it in a large amount of DNA samples for any further analysis.

## 3.2 Detection of glycophorin deletion breakpoints

The samples with low value of their normalised NGS data have been categorised according to the glycophorin region that could has been deleted. Thus, same samples were considered as DEL1, DEL2, DEL6 and DEL7.

### 3.2.1 Deletion type 1 (DEL1)

The 5 kb windows plot generated suggests a possible breakpoint for DEL1 and shows the whole *GYPB* gene is deleted in these positive controls (Figure 3.2).

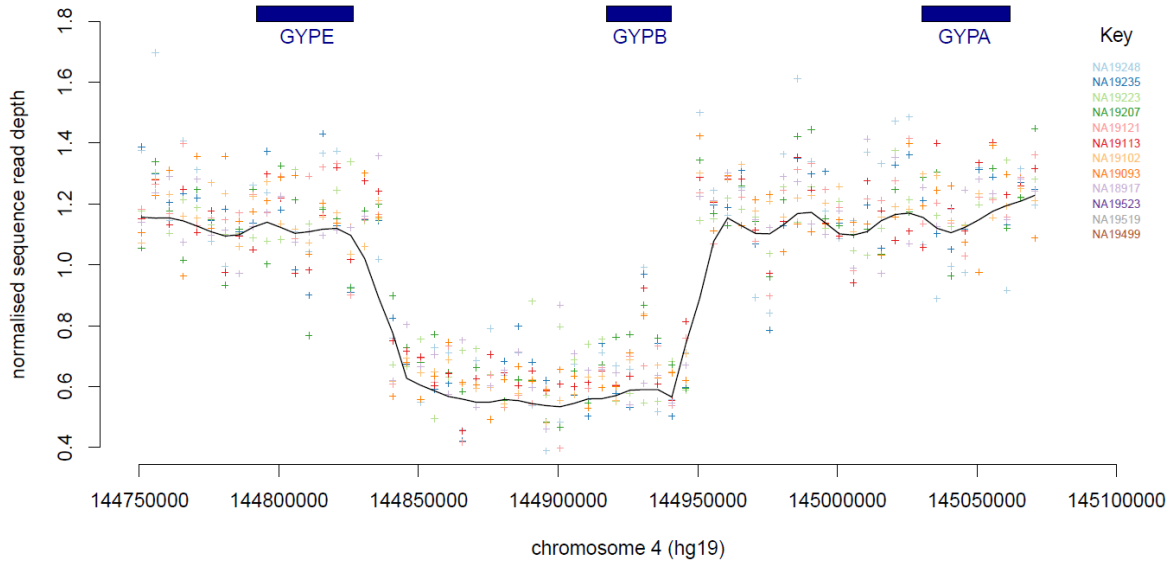



**Figure 3.2: Analysis of DEL1 deletions using next-generation sequencing data.** The points on the graph indicate the normalised read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

Fibre-FISH was applied on the known DEL1 (NA19223) from the 1000 Genomes Project sample collection using four different labelled fosmid probes (F7, F1, F12 and G10) in order to confirm the glycophorin variant (Chapter 2). The fibre-FISH result confirmed the existence of the DEL1 in the positive control used for this analysis (Figure 3.3).

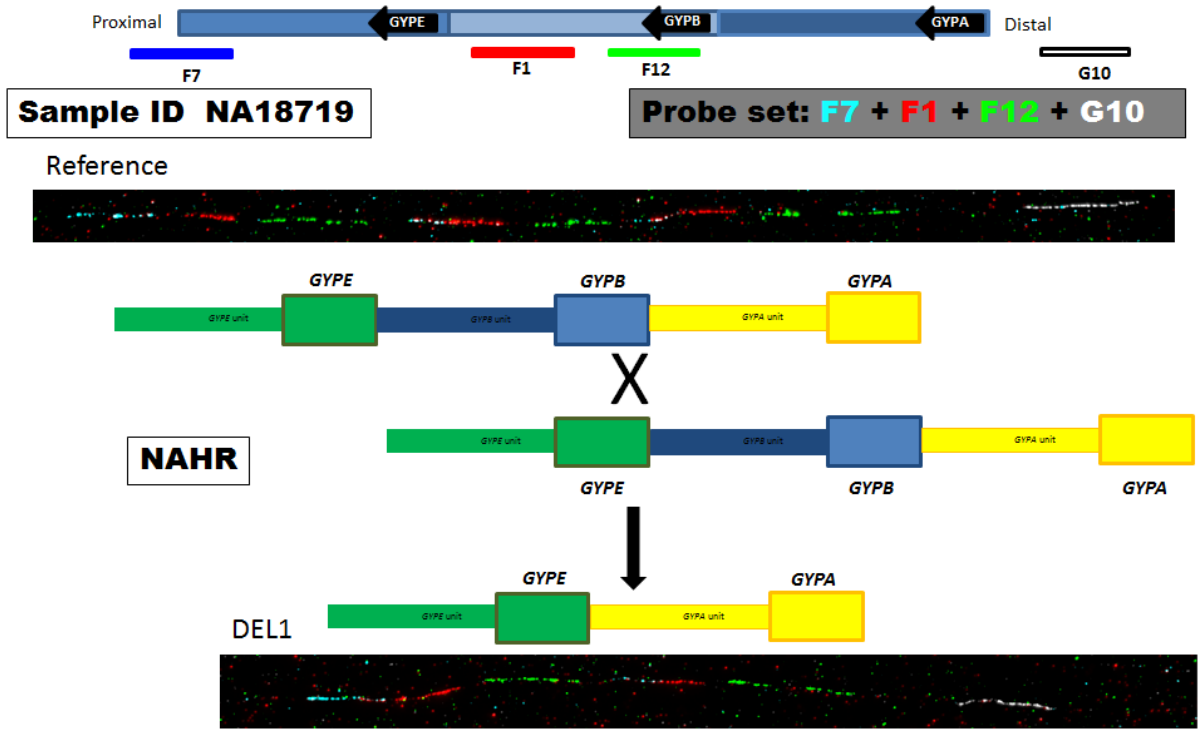



**Figure 3.3: Confirmation of DEL1 deletion using fibre-FISH analysis.** The top of the figure shows the names of the probes used for this analysis and their position on the genome. The two FISH results in the top and the bottom of the figure represent the fibre-FISH result for the reference haplotype compared with the DEL1 positive control fibre-FISH result. The reference fibre-FISH result shows the G10 (white) probe is specific for the distal and F7 (blue) is specific for the proximal, while F12 (green) and F1 (red) are shown three times for *GYPA*, *GYPB* and *GYPE* because they are very similar to each other. However, the DEL1 positive control fibre-FISH result shows the G10 (white) probe is specific for the distal and F7 (blue) is specific for the proximal, while F12 (green) and F1 (red) are shown only two times confirming the deletion of *GYPB* region. The middle of the figure shows a cartoon image of the DEL1 NAHR mechanism.

The specific primer pair for DEL1 was designed using the normalized sequence read depth data, the 5 kb windows plot and the fibre-FISH results (Chapter 2) in order to amplify the deletion breakpoints region and Sanger sequence the PCR product. The primers flank the whole expected region of the DEL1 deletion (Figure 3.4).

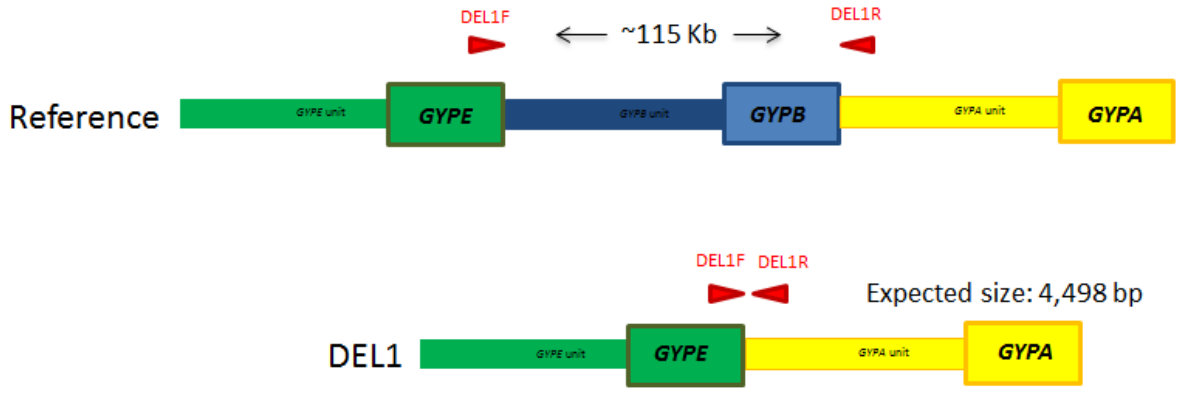



**Figure 3.4: Positions of the DEL1 primer pair used for the Long-PCR.** DEL1F with the red triangle represents the forward primer and DEL1R with the red triangle represents the reverse primer.

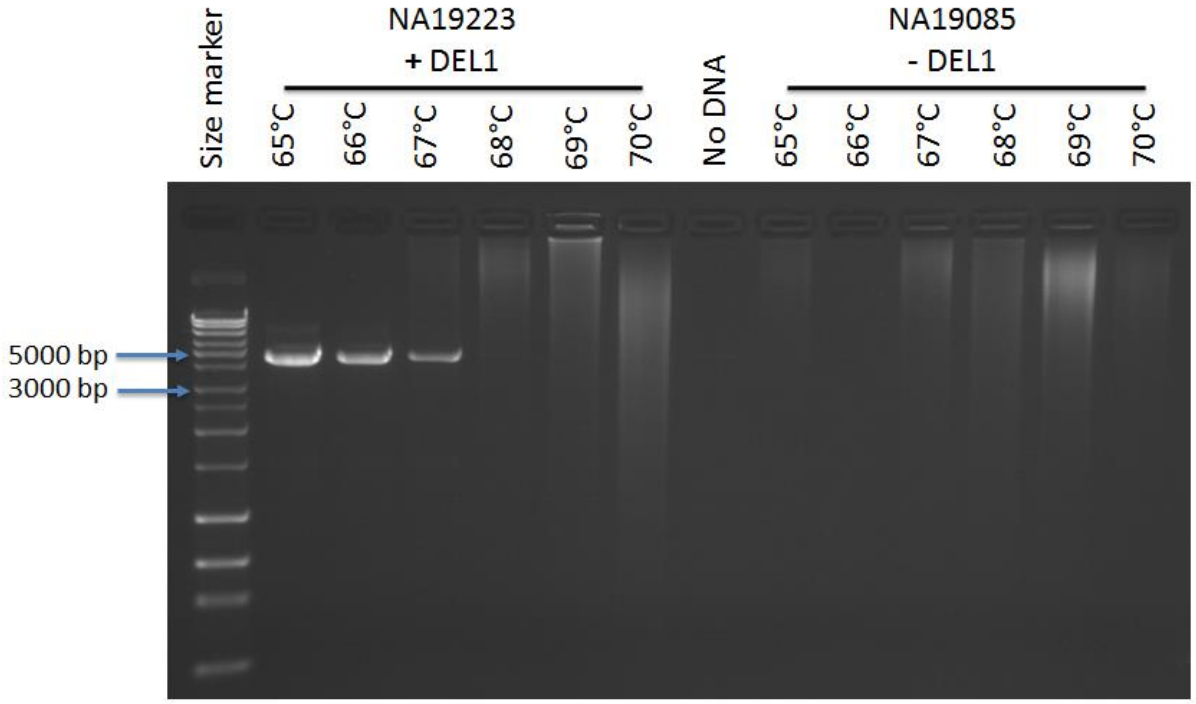



**Figure 3.5: Gradient long-PCR gel result for DEL1 using the primer pair DEL1F and DEL1R.** There is successful amplification only in NA19223 (DEL1 carrier) with the band matching the expected 4,498 bp size for the fragment, indicating that the correct region was amplified. The bands can be seen at 65°C, 66°C and 67°C. The size marker is the Bioline HyperLadder™ 1kb Plus ladder.

After sequencing DEL1 specific PCR product and the multiple sequence alignments, the breakpoints of this variant were identified by the switch from *GYPE* to *GYPA* paralogue specific variants. The results of the alignments show that DEL1 proximal breakpoint is in *GYPE* unit with a range of 137 bp (chr4:144,835,143-144,835,279) (Table 3.1). The proximal breakpoint sequence maps to (CATATA)n (simple repeat) and L1PBa (LINE) repeat elements. The distal breakpoint is in *GYPB* unit with a range of 143 bp (chr4:144,945,375-

144,945,517) (Table 3.1). The distal breakpoint sequence maps to L1ME3 (LINE), (CA)n (simple repeat) and L1PBa (LINE) repeat elements. DEL1 covers the *GYPB* gene and the non-coding region downstream of the *GYPE* gene (see Appendix 2B and 4 for the sequencing stages of the primers and full alignments).

### 3.2.2 Deletion type 2 (DEL2)

The 5kb windows plot generated suggests the possible breakpoint for DEL2 and shows that the entire *GYPB* gene is deleted in these positive controls (Figure 3.6).

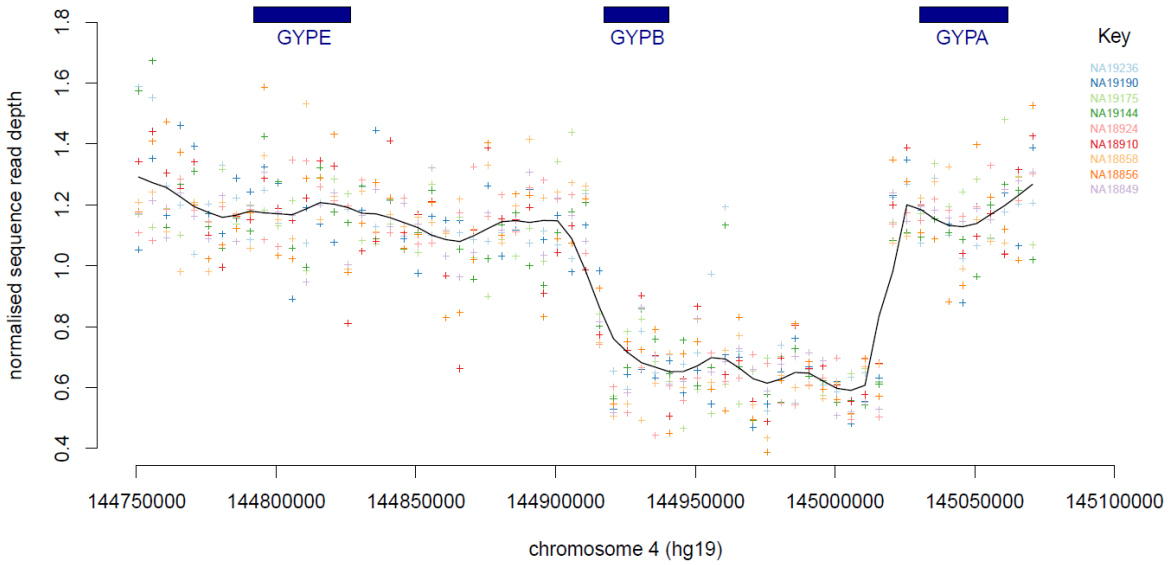



**Figure 3.6: Analysis of DEL2 deletions using next-generation sequencing data.** The points on the graph indicate the value of the sequence read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

Fibre-FISH was applied on the known DEL2 (NA19144) from the 1000 Genomes Project sample collection using four different labelled fosmid probes (F7, F1, F12 and G10) in order to confirm the glycophorin variant (Chapter 2). The fibre-FISH result confirmed the existence of DEL2 in the positive control used for this analysis (Figure 3.7).

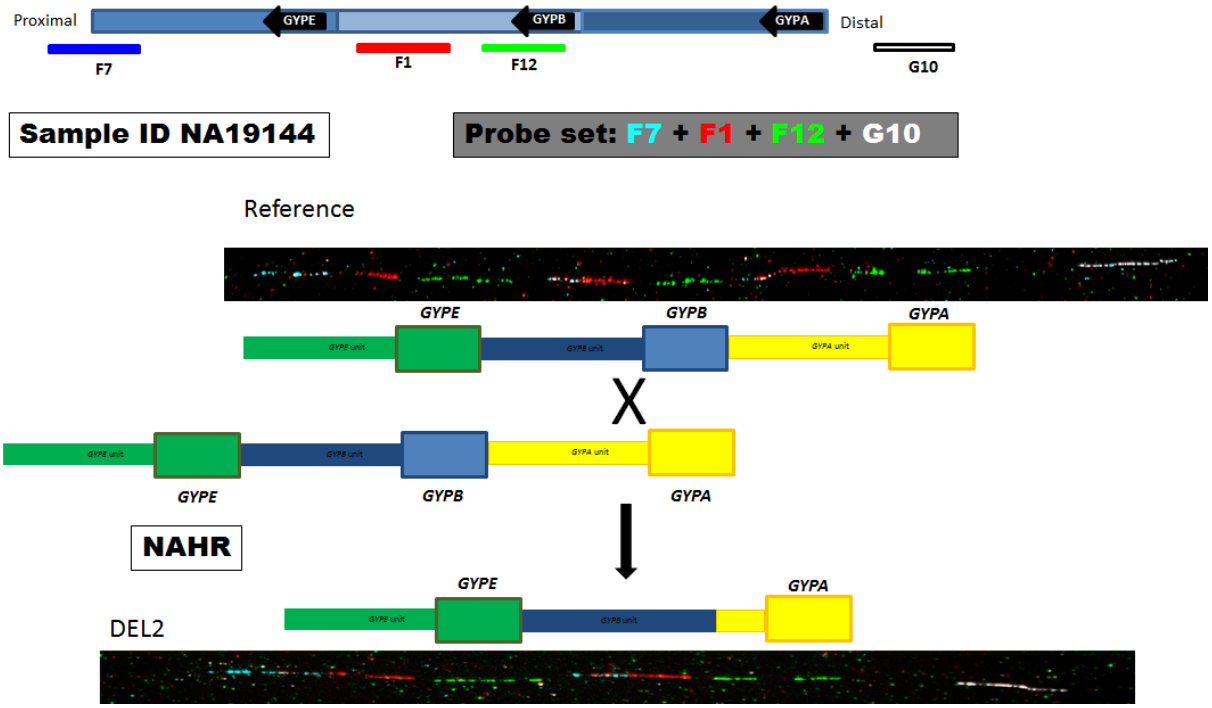



**Figure 3.7: Confirmation of DEL2 deletion using fibre-FISH analysis.** The top of the figure shows the names of the probes used for this analysis and their position on the genome. The two FISH results in the top and the bottom of the figure represent the fibre-FISH result for the reference haplotype compared with the DEL2 positive control fibre-FISH result. The middle of the figure shows a cartoon image of the DEL2 NAHR mechanism.

The specific primer pair for DEL2 was designed using the normalized sequence read depth data, the 5 kb windows plot and the fibre-FISH results (Chapter 2) in order to amplify the deletion breakpoints region and Sanger sequence the PCR product. The primers flank the whole expected region of the DEL2 deletion (Figure 3.8).

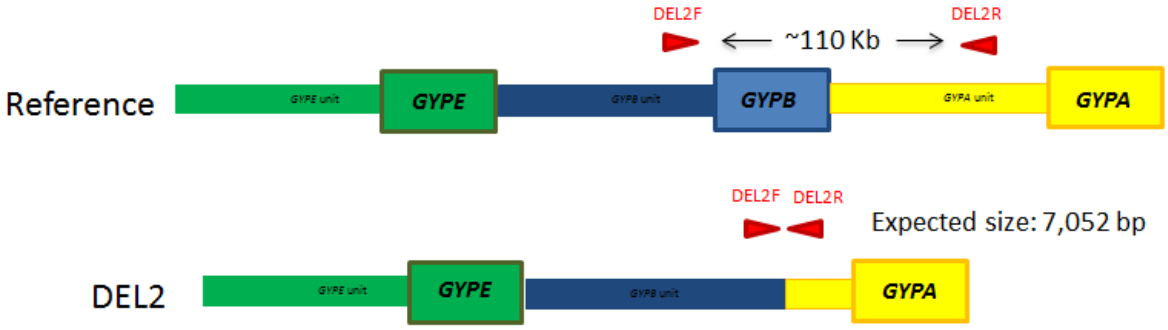



**Figure 3.8: Positions of the DEL2 primer pair used for the Long-PCR.** DEL2F with the red triangle represents the forward primer and DEL2R with the red triangle represents the reverse primer.

Gradient long-PCR was applied using the DEL2 primer pair (Chapter 2) on a DEL2 positive DNA sample (NA19144) and a DEL2 negative DNA sample (NA19085) in order to assess the specificity of the primers and select the best annealing temperature for them. The gel result of the Long-PCR shows the primer pair is specific for the DEL2 variant at 65°C, 66°C,

67°C and 68°C (Figure 3.9); both 65°C and 66°C were good and clear, however, 65°C generates most product, therefore it was used for sequencing.

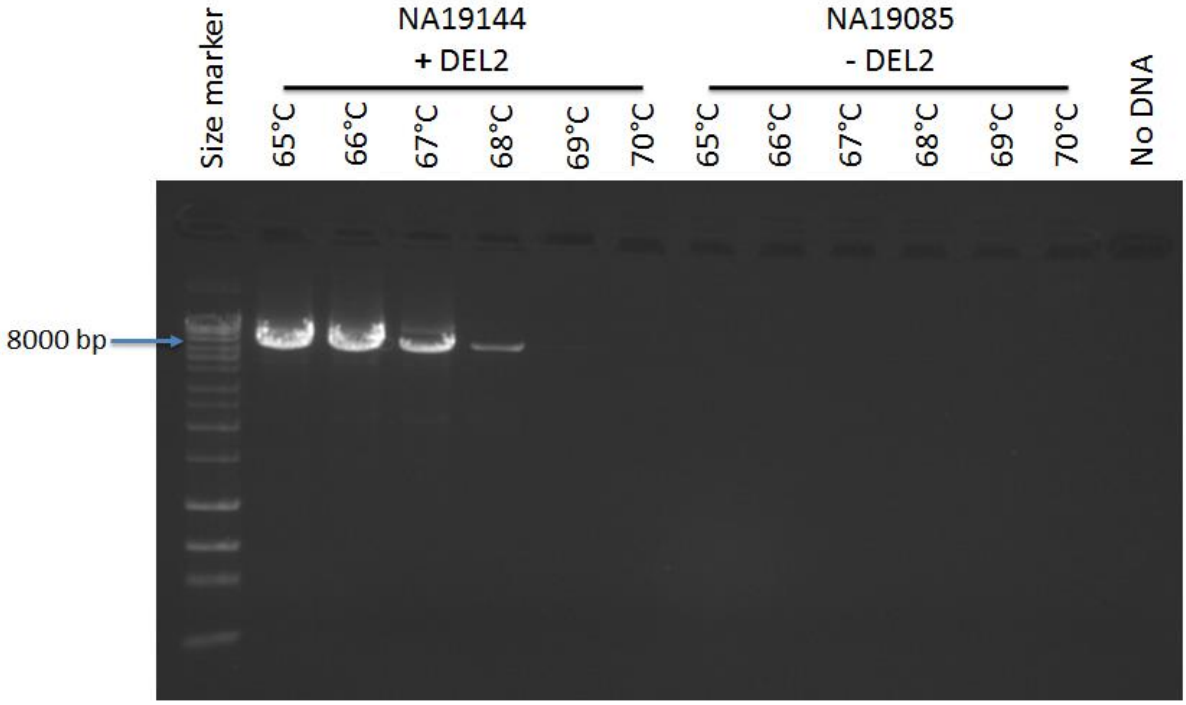



**Figure 3.9: Gradient long-PCR gel result for DEL2 using primer pair DEL2F and DEL2R.** There is successful amplification only in NA19144 (DEL2 carrier) with the band matching the expected 7,052 bp size for the fragment, indicating that the correct region was amplified. The bands can be seen at 65°C, 66°C, 67°C and 68°C. The size marker is the Bioline HyperLadder™ 1kb Plus ladder.

After sequencing DEL2 specific PCR product and the multiple sequence alignments, the breakpoints of this variant were identified by the switch from *GYPB* to *GYPA* paralogue specific variants. The results of the alignments show that DEL2 proximal breakpoint is in *GYPB* unit with a range of 130 bp (chr4:144,912,872-144,913,001) (Table 3.1). The proximal breakpoint sequence maps to THE1C, LTR repeat element. The distal breakpoint is in *GYPA* unit with a range of 130 bp (chr4:145,016,127-145,016,256) (Table 3.1). The distal breakpoint sequence maps to THE1C, LTR repeat element. DEL2 covers *GYPB* gene and the non-coding region upstream of the *GYPA* gene. (See Appendices 2 and 4 for the sequencing stages primers and full alignments).

### 3.2.3 Deletion type 6 (DEL6)

The generated 5kb windows plot suggests the possible breakpoint for DEL6 and shows that the entire *GYPE* and *GYPB* genes are deleted in this positive control (Figure 3.10).

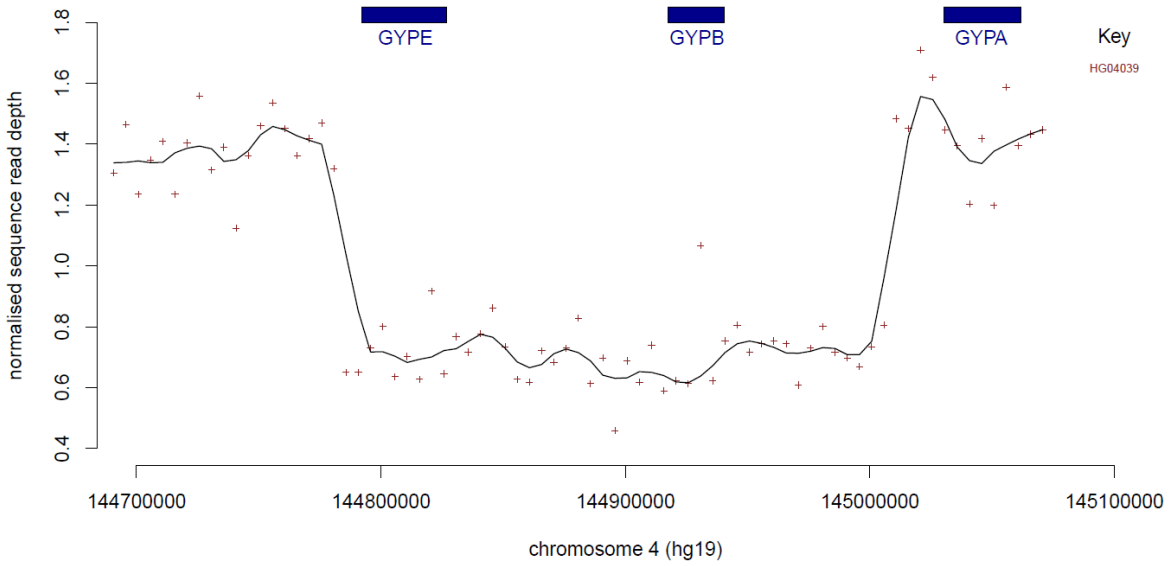



**Figure 3.10: Analysis of DEL6 deletions using next-generation sequencing data.** The points on the graph indicate the normalised read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

Fibre-FISH was applied on the known DEL6 (HG04039) from the 1000 Genomes Project sample collection using four different labelled fosmid probes (F7, F1, F12 and G10) in order to confirm the glycophorin variant (Chapter 2). The fibre-FISH result confirmed the existence of DEL6 in the positive control used for this analysis (Figure 3.11).

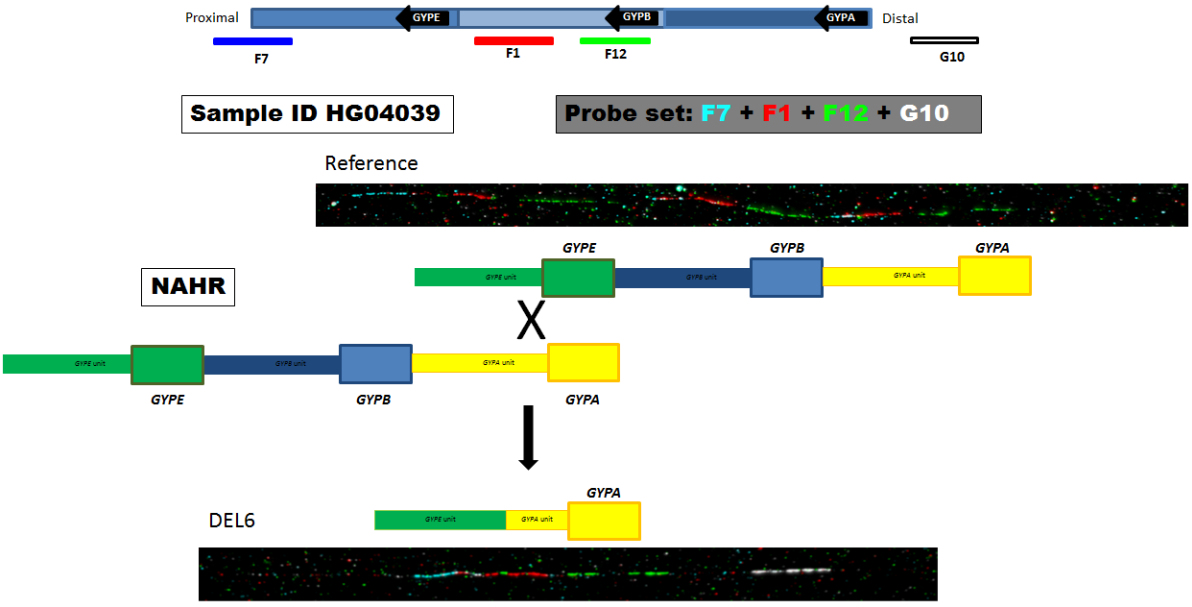



**Figure 3.11: Confirmation of DEL6 deletion using fibre-FISH analysis.** The top of the figure shows the names of the probes used for this analysis and their position on the genome. The two FISH results in the top and the bottom of the figure represent the fibre-FISH result for the reference haplotype compared with the DEL6 positive control fibre-FISH result. The middle of the figure shows a cartoon image of the DEL6 NAHR mechanism.

The specific primer pair for DEL6 was designed using the normalized sequence read depth data, the 5 kb windows plot and the fibre-FISH results (Chapter 2) in order to amplify the deletion breakpoints region and Sanger sequence the PCR product. The primers flank the whole expected region of the DEL6 deletion (Figure 3.12).

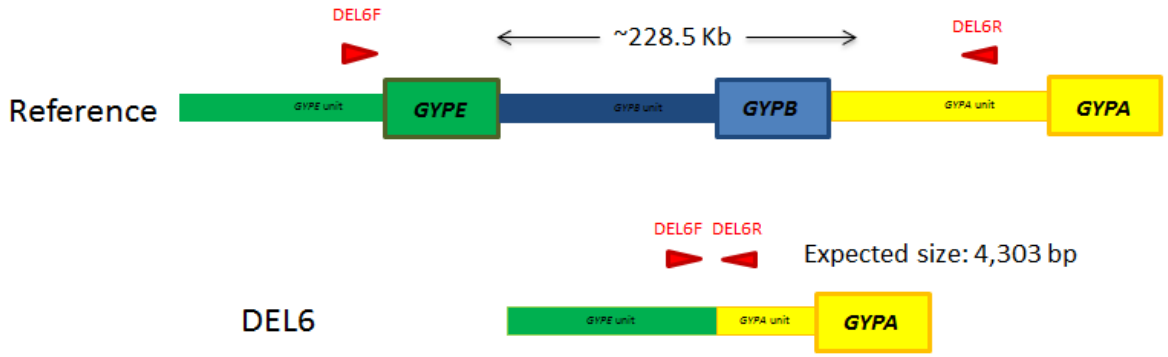



**Figure 3.12: Positions of the DEL6 primer pair used for the Long-PCR.** DEL6F with the red triangle represents the forward primer and DEL6R with the red triangle represents the reverse primer.

Gradient long-PCR was applied using the DEL6 primer pair (Chapter 2) on a DEL6 positive DNA sample (HG04039) and a DEL6 negative DNA sample (Co075) in order to assess the specificity of the primers and select the best annealing temperature for them. The gel result of the Long-PCR shows the primer pair is specific for DEL6 variant at 60°C, 62°C and 63°C (Figure 3.13); however, 63°C generates most product, therefore it was used for sequencing.

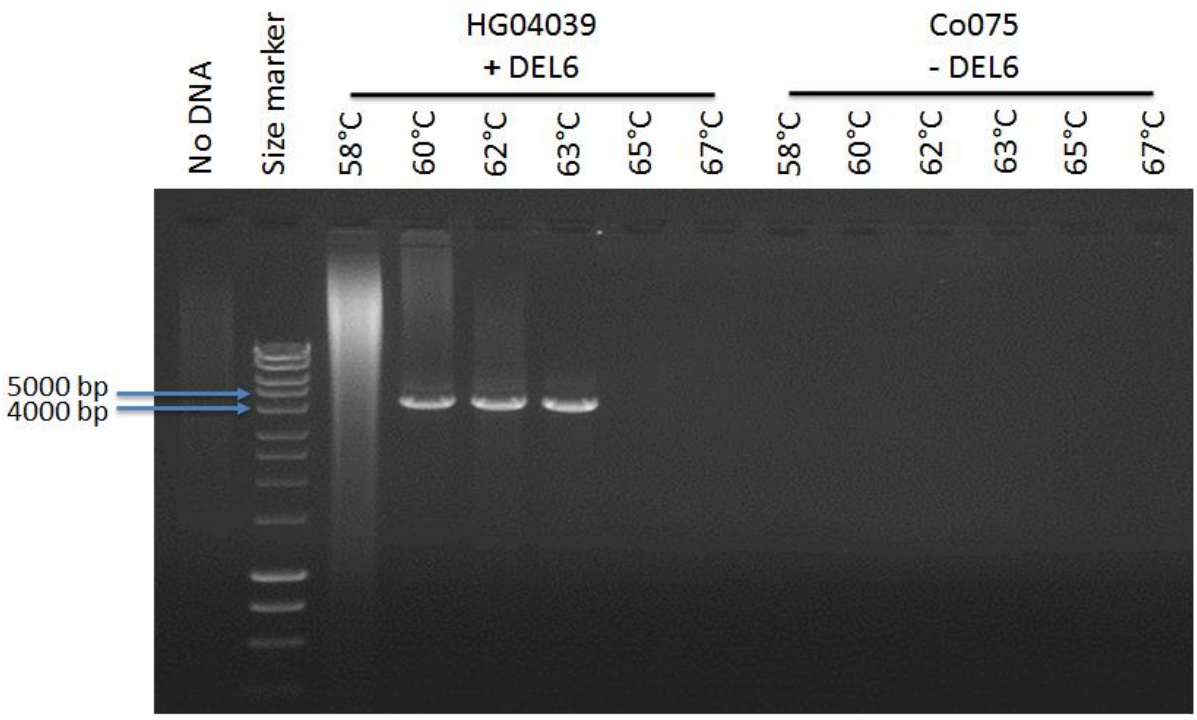



**Figure 3.13: Gradient long-PCR gel result for DEL6 using the primer pair DEL6F and DEL6R.** There is successful amplification only in HG04039 (DEL6 carrier) with the band matching the expected 4,303 bp size for the fragment, indicating that the correct region was amplified. The bands can be seen at 60°C, 62°C and 63°C. The size marker is the Bioline HyperLadder™ 1kb ladder.

After sequencing DEL6 specific PCR product and the multiple sequence alignments, the breakpoints of this variant were identified by the switch from *GYPE* to *GYPA* paralogue specific variants. The results of the alignments show that DEL6 proximal breakpoint is in *GYPE* unit with a range of 93 bp (chr4:144,780,045-144,780,137) (Table 3.1). The proximal breakpoint sequence does not map to any repeat element. The distal breakpoint is in *GYPA* unit with a range of 93 bp (chr4:145,004,120-145,004,212) (Table 3.1). The distal breakpoint sequence maps to L1PA5, LINE repeat element. DEL6 covers the *GYPE* gene, the *GYPB* gene, the non-coding region between the *GYPE* and *GYPB* genes and the non-coding region upstream of the *GYPA* gene (see Appendices 2 and 4 for the sequencing stage primers and full alignments).

### 3.2.4 Deletion type 7 (DEL7)

The 5kb windows plot generated suggests the possible breakpoint for DEL7 and shows that the entire *GYPE* gene is deleted in these positive controls (Figure 3.14).

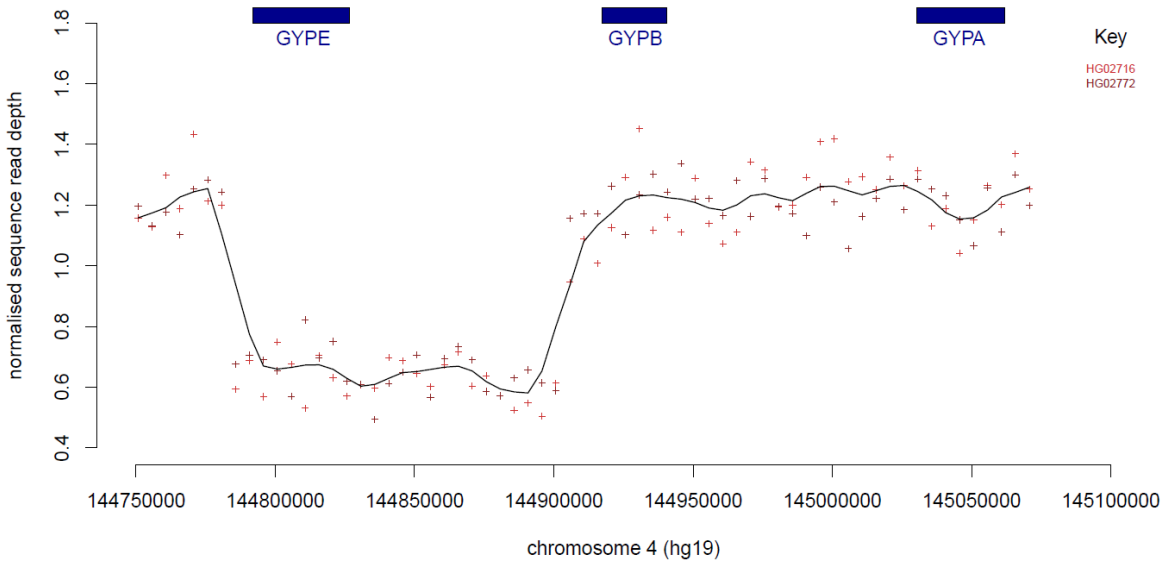



**Figure 3.14: Analysis of DEL7 deletions using next-generation sequencing data.** The points on the graph indicate the value of the sequence read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

Fibre-FISH was applied on the known DEL7 (HG02716) from the 1000 Genomes Project sample collection using four different labelled fosmid probes (F7, F1, F12 and G10) in order to confirm the glycophorin variant (Chapter 2). The fibre-FISH result confirmed the existence of the DEL7 in the positive control used for this analysis (Figure 3.15).

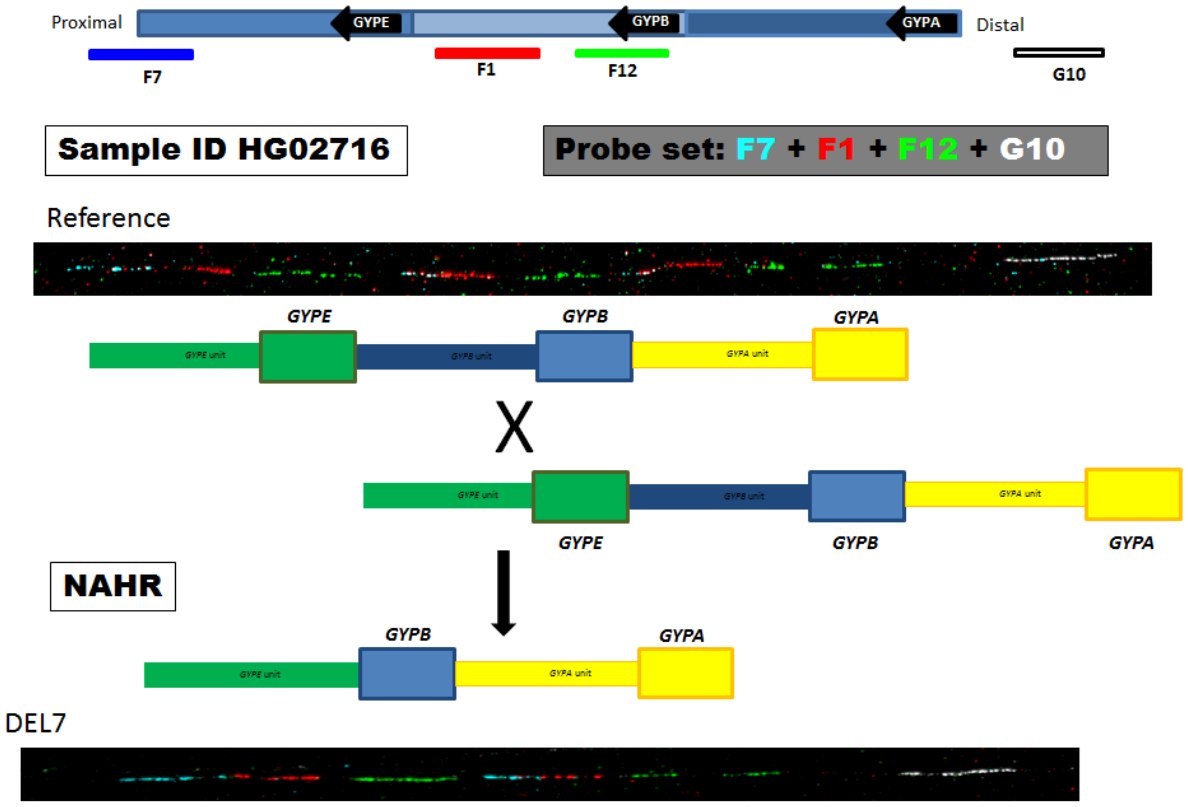



**Figure 3.15: Confirmation of DEL7 deletion using fibre-FISH analysis.** The top of the figure shows the names of the probes used for this analysis and their position on the genome. The two FISH results in the top and the bottom of the figure represent the fibre-FISH result for the reference haplotype compared with the DEL7 positive control fibre-FISH result. The middle of the figure shows a cartoon image of the DEL7 NAHR mechanism.

The specific primer pair for DEL7 was designed using the normalized sequence read depth data, the 5 kb windows plot and the fibre-FISH results (Chapter 2) in order to amplify the deletion breakpoints region and Sanger sequence the PCR product. The primers flank the whole expected region of the DEL7 deletion (Figure 3.16).

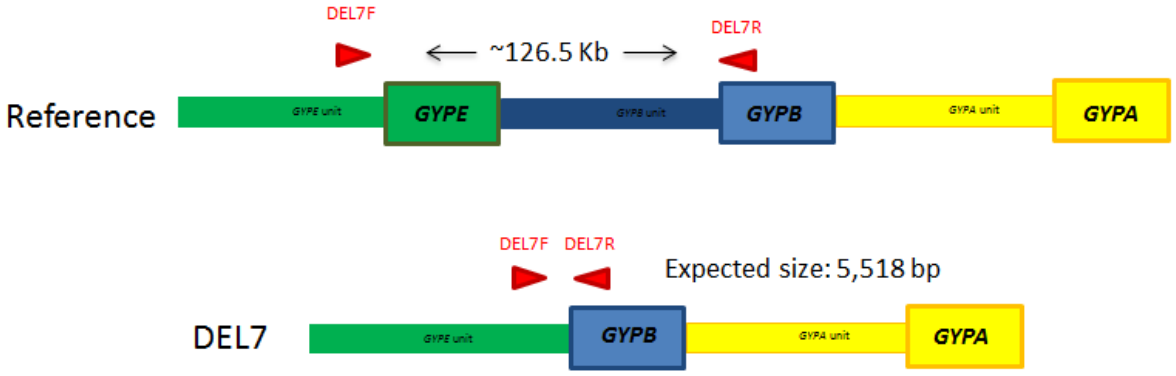



**Figure 3.16: Positions of the DEL7 primer pair used for the Long-PCR.** DEL7F with the red triangle represents the forward primer and DEL7R with the red triangle represents the reverse primer.

Gradient long-PCR was applied using the DEL7 primer pair (Chapter 2) on a DEL7 positive DNA sample (HG02716) and a DEL7 negative DNA sample (Co075) in order to assess the specificity of the primers and select the best annealing temperature for them. The gel result of the Long-PCR shows that the primer pair is specific for the DEL7 variant at 65°C and 67°C (Figure 3.17); however, 65°C generates most product, therefore it was used for sequencing.

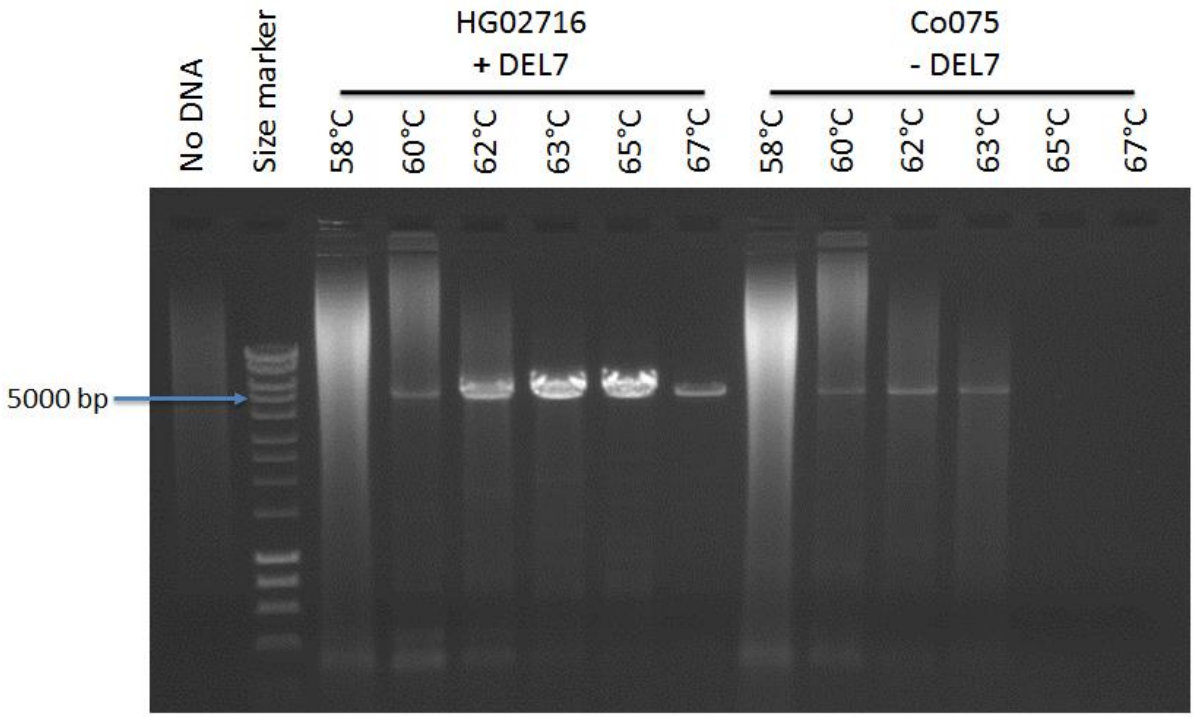



**Figure 3.17: Gradient long-PCR gel result for DEL7 using the primer pair DEL7F and DEL7R.** There is successful amplification in HG02716 (DEL7 carrier) matching the expected 5,518 bp size for the fragment, indicating that the correct region had been amplified. The specific bands of DEL7 can be seen at 65°C and 67°C. The size marker is the Bioline HyperLadder™ 1kb ladder.

After sequencing DEL7 specific PCR product and the multiple sequence alignments, the breakpoints of this variant were identified by the switch from *GYPE* to *GYPB* paralogue specific variants. The results of the alignments show that DEL7 proximal breakpoint is in *GYPE* unit with a range of 48 bp (chr4: 144,780,450 144,780,498) (Table 3.1). The proximal breakpoint sequence does not map to any repeat element. The distal breakpoint is in *GYPB* unit with a range of 48 bp (chr4: 144,901,287 144,901,335) (Table 3.1). The distal breakpoint sequence does not map to any repeat element. DEL7 covers the *GYPE* gene and the non-coding region upstream of the *GYPB* gene (see Appendices 2 and 4 for the sequencing stage primers and full alignments).

## 3.3 Detection of glycophorin duplication breakpoints

The samples with high value of their normalised NGS data have been categorised according to the glycophorin region that could has been duplicated. Thus, same samples were considered as DUP2, DUP3, DUP7, DUP14, DUP29 and DUP24.

### 3.3.1 Duplication type 2 (DUP2)

The 5kb windows plot generated suggests the possible breakpoint for DUP2 and shows that most of the *GYPB* gene and part of the *GYPA* gene are duplicated in these positive controls (Figure 3.18).

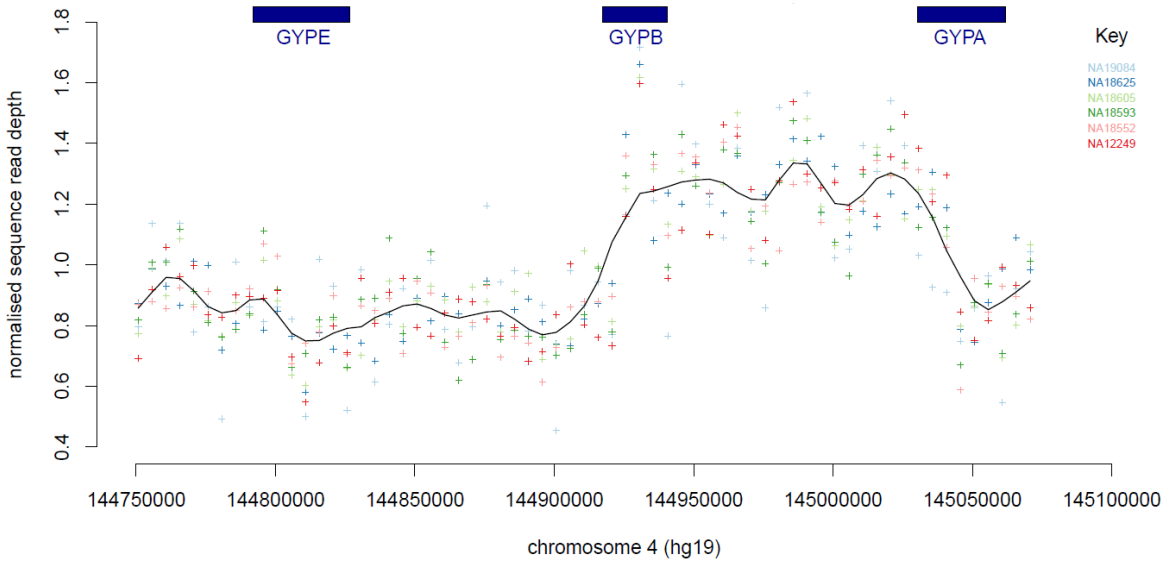



**Figure 3.18: Analysis of DUP2 duplications using next-generation sequencing data.** The points on the graph indicate the value of the sequence read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

Fibre-FISH was applied on the known DUP2 (NA18593) from the 1000 Genomes Project sample collection using four different labelled fosmid probes (F7, F1, F12 and G10) in order to confirm the glycophorin variant (Chapter 2). The fibre-FISH result confirmed the existence of DUP2 in the positive control used for this analysis (Figure 3.19).

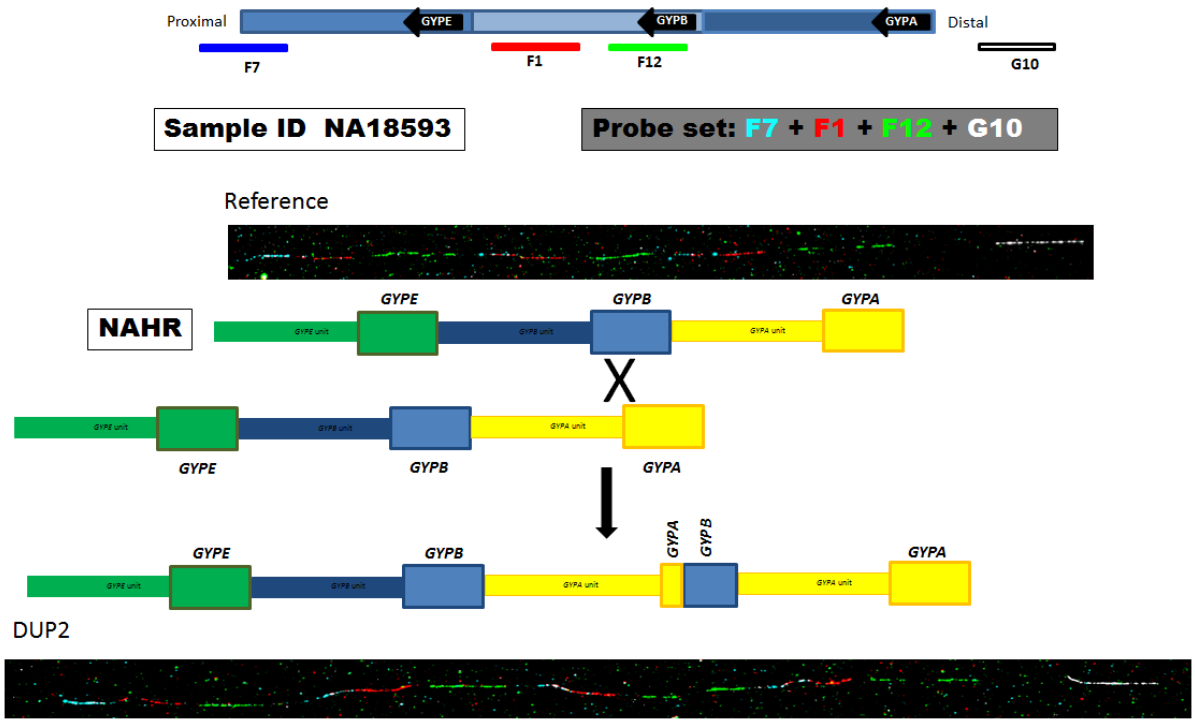



**Figure 3.19: Confirmation of DUP2 duplication using fibre-FISH analysis.** The top of the figure shows the names of the probes used for this analysis and their position on the genome. The two FISH results in the top and the bottom of the figure represent the fibre-FISH result for the reference haplotype compared with the DUP2 positive control fibre-FISH result. The middle of the figure shows a cartoon image of the DUP2 NAHR mechanism.

The specific primer pair for DUP2 was designed using the normalized sequence read depth data, the 5 kb windows plot and the fibre-FISH results (Chapter 2) in order to amplify the duplications breakpoints region and Sanger sequence the PCR product. The primers flank the expected breakpoint region of the DUP2 duplication (Figure 3.20).

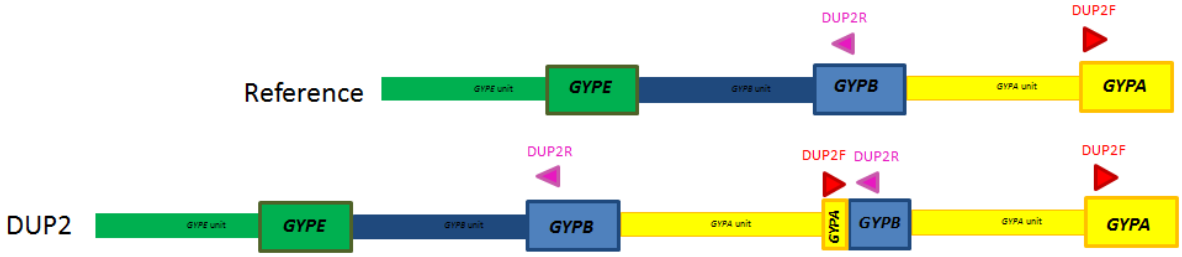



**Figure 3.20: Positions of the DUP2 primer pair used for the Long-PCR.** DUP2F with the red triangle represents the forward primer and DUP2R with the pink triangle represents the reverse primer.

Gradient long-PCR was applied using the DUP2 primer pair (Chapter 2) on a DUP2 positive DNA sample (NA18625) and a DUP2 negative DNA sample (Co185) in order to assess the specificity of the primers and chose the best annealing temperature for them. The gel result of the Long-PCR shows very faint and thin specific bands for the DUP2 variant at 62°C, 63°C

and 64°C (Figure 3.21); 62°C, 63°C and 64°C bands show the same intensity and all of them are faint, however, 64°C was chosen in order to make it more specific for DUP2 as high annealing temperature increases the specificity.

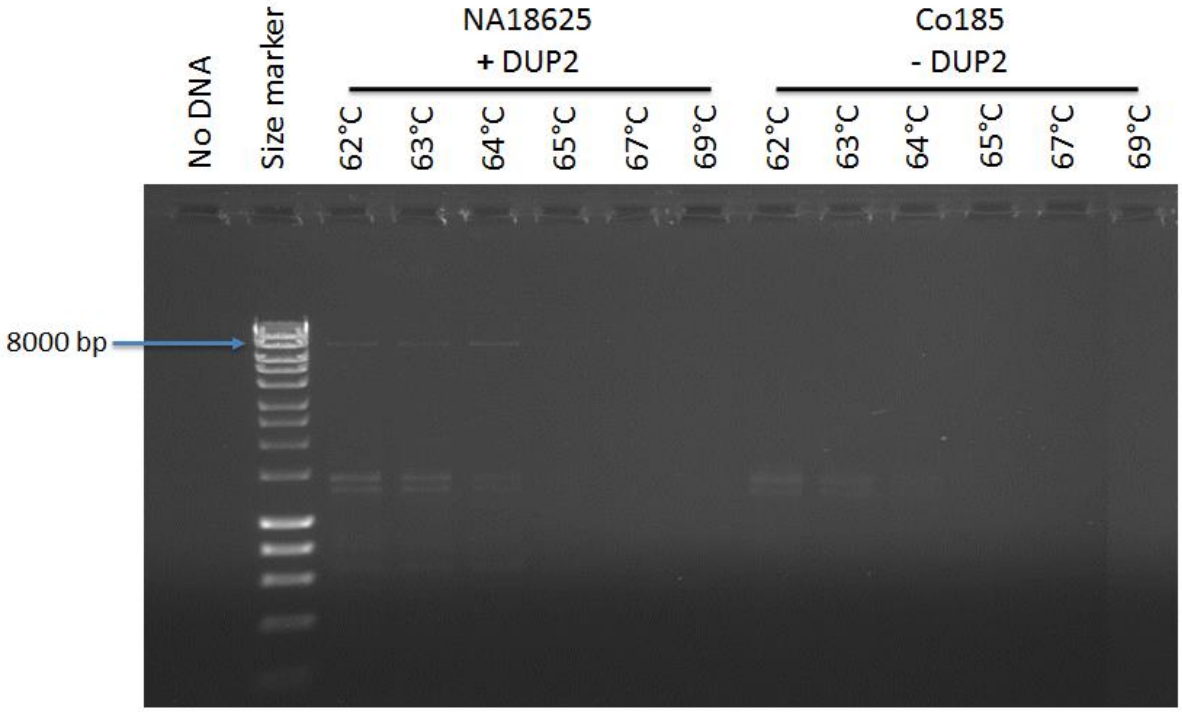



**Figure 3.21: Gradient long-PCR gel result for DUP2 using primer pair DUP2F and DUP2R.** There is successful amplification in NA18625 (DUP2 carrier) matching the expected 9,471 bp size for the fragment, indicating the correct region was amplified. The specific bands of DUP2 can be seen at 62°C, 63°C and 64°C. The size marker is the Bioline HyperLadder™ 1kb ladder.

Although, DUP2 PCR products were generated and extracted from an agarose gel in order to be sequenced, the sequencing cannot be completed because of a sequencing error after the first stage (sequencing the forward and reverse templates). It was thus not possible to design a successful Stage 2 sequencing primer.

### 3.3.2 Duplication type 3 (DUP3)

The 5kb windows plot generated suggests the possible breakpoint for DUP3 and shows that the entire *GYPE* and *GYPB* genes are duplicated in these positive controls (Figure 3.22).

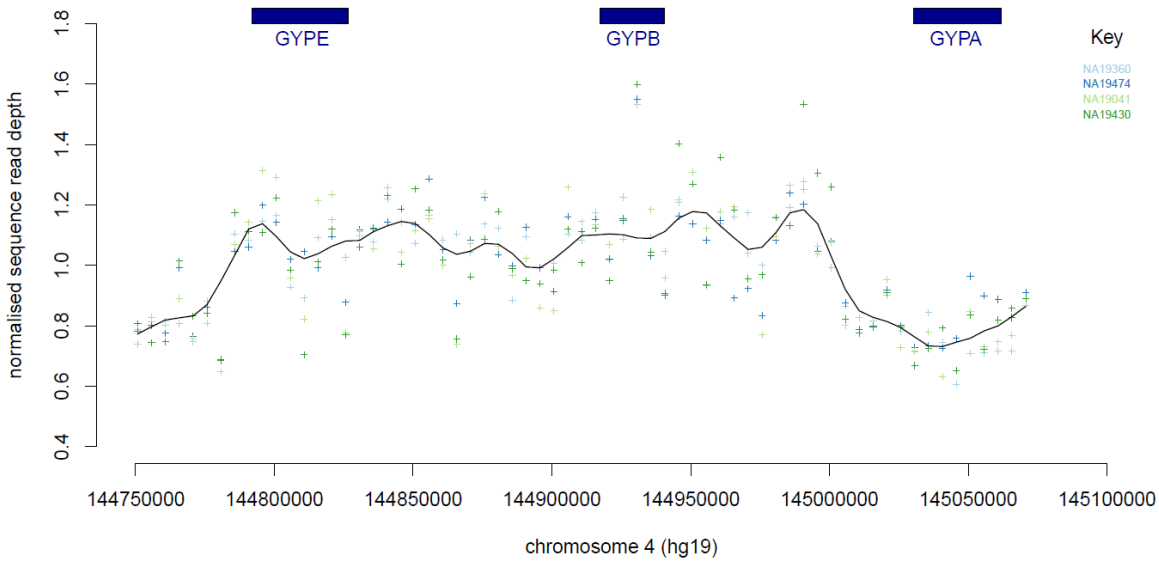



**Figure 3.22: Analysis of DUP3 duplications using next-generation sequencing data.** The points on the graph indicate the normalised read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

Fibre-FISH was applied on the known DUP3 (NA19474) from the 1000 Genomes Project sample collection using four different labelled fosmid probes (F7, F1, F12 and G10) in order to confirm the glycophorin variant (Chapter 2). The fibre-FISH result confirmed the existence of DUP3 in the positive control used for this analysis (Figure 3.23).

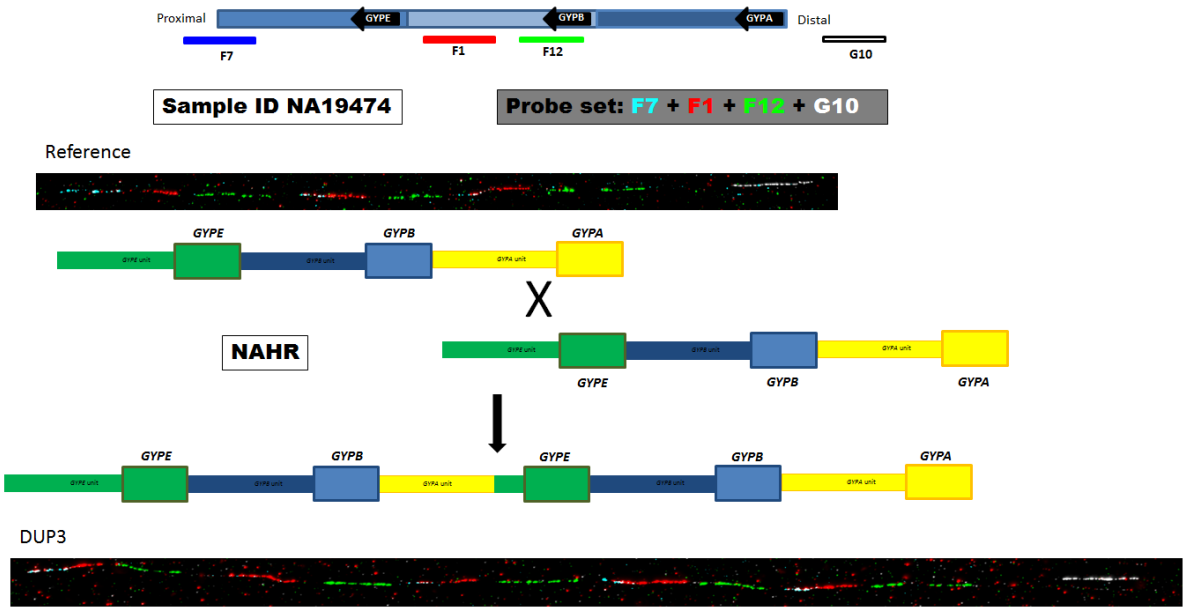



**Figure 3.23: Confirmation of DUP3 duplication using fibre-FISH analysis.** The top of the figure shows the names of the probes used for this analysis and their position on the genome. The two FISH results in the top and the bottom of the figure represent the fibre-FISH result for the reference haplotype compared with the DUP3 positive control fibre-FISH result. The middle of the figure shows a cartoon image of the DUP3 NAHR mechanism.

The specific primer pair for DUP3 was designed using the normalized sequence read depth data, the 5 kb windows plot and the fibre-FISH results (Chapter 2) in order to amplify the duplication breakpoints region and Sanger sequence the PCR product. The primers flank the breakpoints expected region of the DUP3 duplication (Figure 3.24).

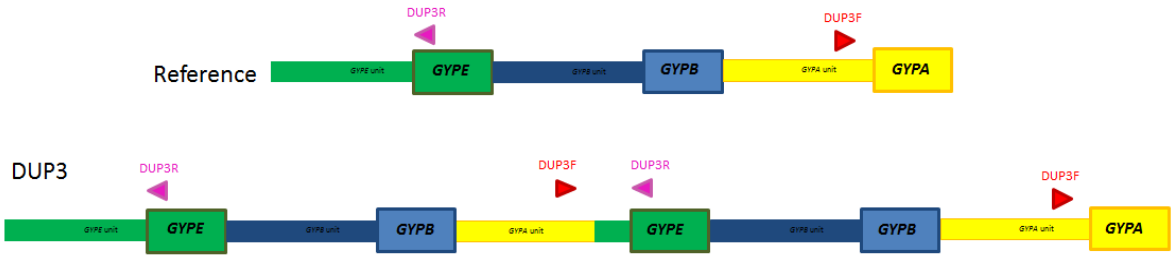



**Figure 3.24: Positions of the DUP3 primer pair used for the Long-PCR.** DUP3F with the red triangle represents the forward primer and DUP3R with the pink triangle represents the reverse primer.

Gradient long-PCR was applied using the DUP3 primer pair (Chapter 2) on a DUP3 positive DNA sample (NA19360) and a DUP3 negative DNA sample (Co906) in order to assess the specificity of the primers and chose the best annealing temperature for them. The gel result of the Long-PCR shows that the primer pair is specific for the DUP3 variant at 63.0°C, 63.5°C and 63.8°C (Figure 3.25); however, 63.5°C generates most product, therefore it was used for sequencing.

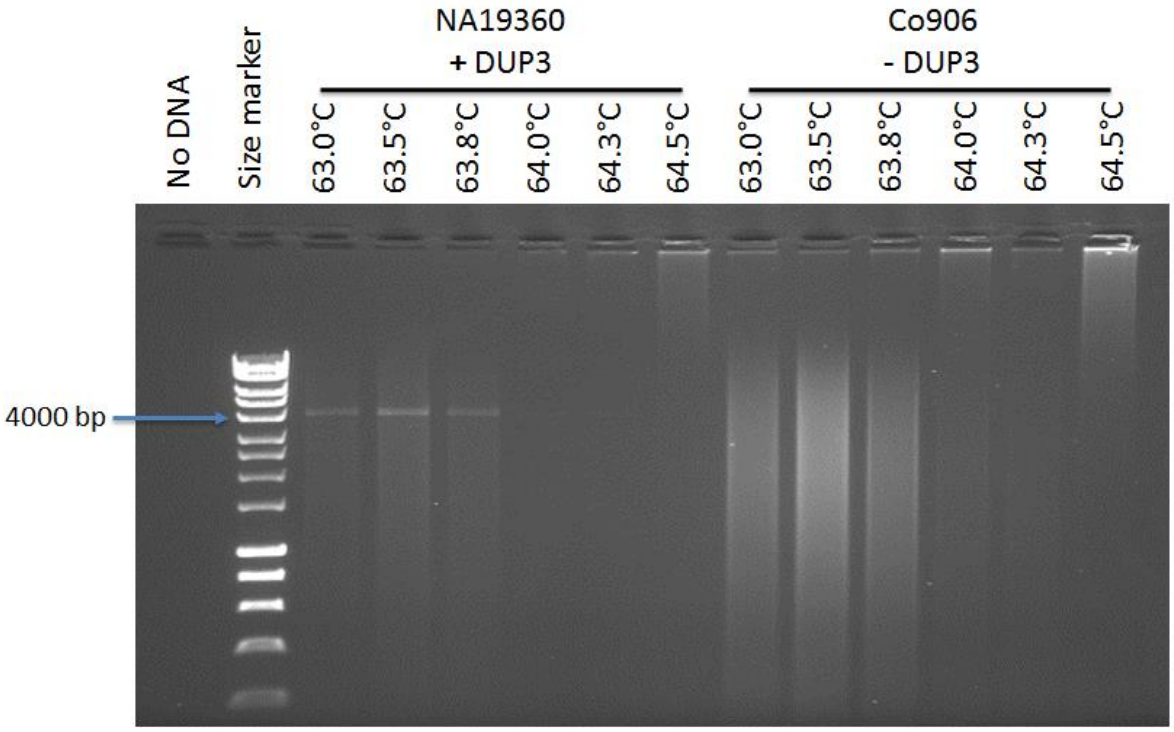



**Figure 3.25: Gradient long-PCR gel result for DUP3 using primer pair DUP3F and DUP3R.** There is successful amplification in NA19360 (DUP3 carrier) matching the expected 4,657 bp size for the fragment, indicating that the correct region was amplified. The specific bands of DUP3 can be seen at 63.0°C, 63.5°C and 63.8°C. The size marker is the Bioline HyperLadder™ 1kb ladder.

Although, DUP3 PCR products were generated and extracted from an agarose gel in order to be sequenced, a lack of time meant that the sequencing could not be completed. However, DUP3 PCR products are ready to be sequenced later by the lab members.

### 3.3.3 Duplication type 7 (DUP7)

The 5kb windows plot generated suggests the possible breakpoint for DUP7 and shows that the entire *GYPE* gene is duplicated in this positive control (Figure 3.26).

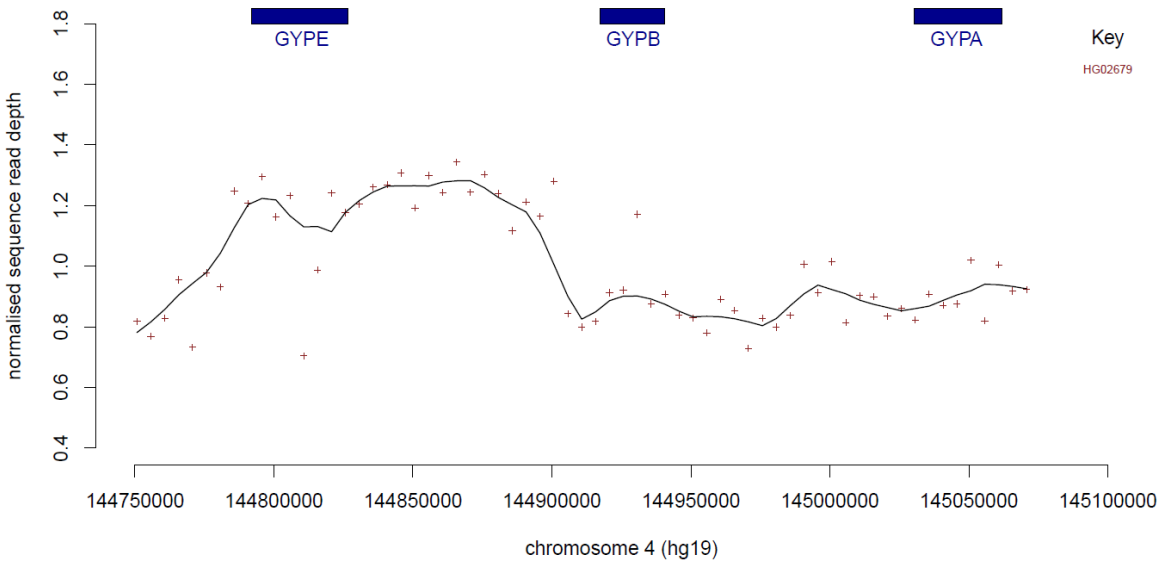



**Figure 3.26: Analysis of DUP7 duplications using next-generation sequencing data.** The points on the graph indicate the normalised read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

Fibre-FISH was applied on the known DUP7 (NA02679) from the 1000 Genomes Project sample collection using four different labelled fosmid probes (F7, F1, F12 and G10) in order to confirm the glycophorin variant (Chapter 2). The fibre-FISH result confirmed the existence of DUP7 in the positive control used for this analysis (Figure 3.27).

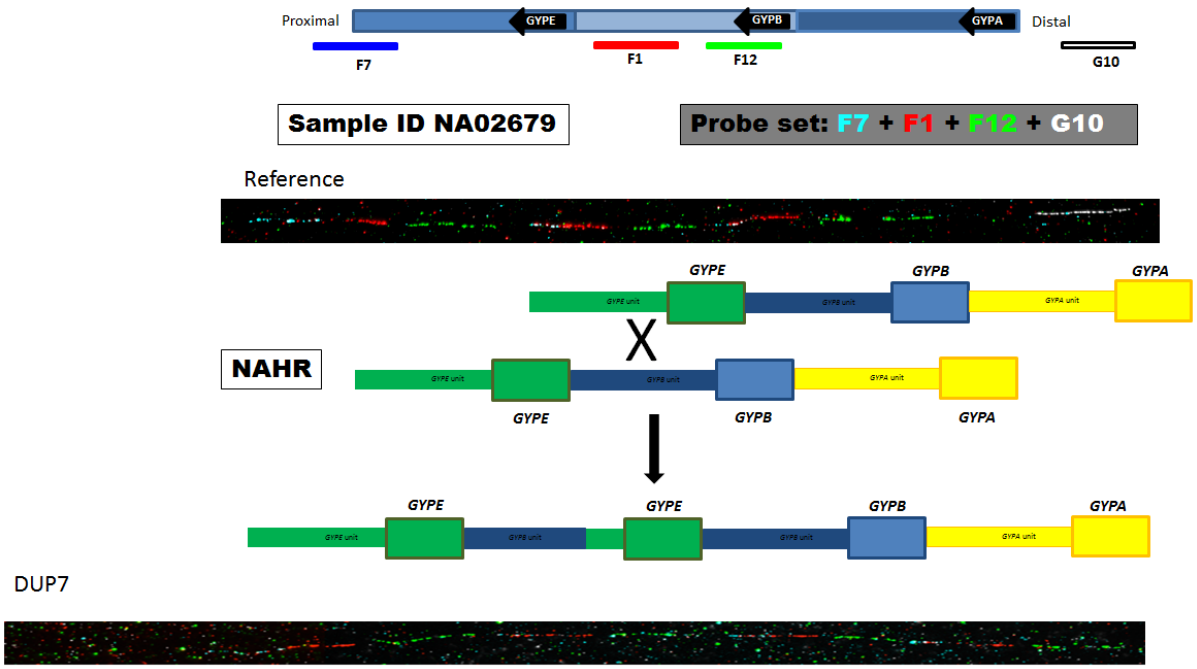



**Figure 3.27: Confirmation of DUP7 duplication using fibre-FISH analysis.** The top of the figure shows the names of the probes used for this analysis and their position on the genome. The two FISH results in the top and the bottom of the figure represent the fibre-FISH result for the reference haplotype compared with the DUP7 positive control fibre-FISH result. The middle of the figure shows a cartoon image of the DUP7 NAHR mechanism.

The specific primer pair for DUP7 was designed using the normalized sequence read depth data, the 5 kb windows plot and the fibre-FISH results (Chapter 2) in order to amplify the duplication breakpoints region and Sanger sequence the PCR product. The primers flank the breakpoints expected region of the DUP7 duplication (Figure 3.28).

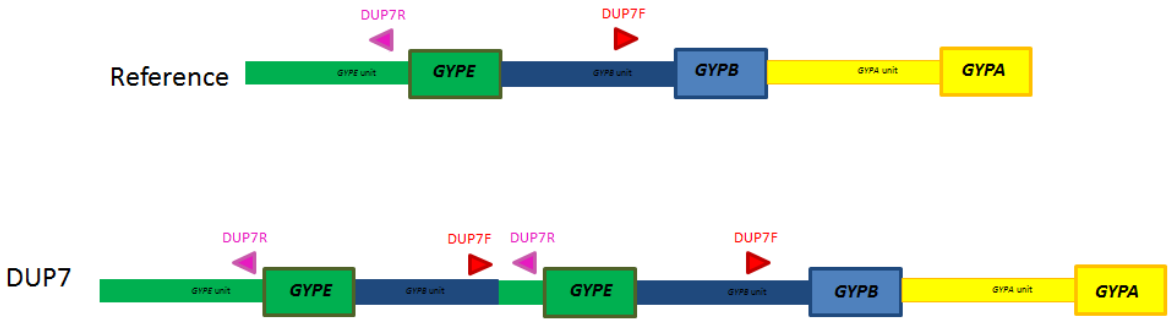



**Figure 3.28: Positions of the DUP7 primer pair used for the Long-PCR.** DUP7F with the red triangle represents the forward primer and DUP7R with the pink triangle represents the reverse primer.

Gradient long-PCR was applied using the DUP7 primer pair (Chapter 2) on a DUP7 positive DNA sample (HG02679) and a DUP7 negative DNA sample (Co935) in order to assess the specificity of the primers and select the best annealing temperature for them. The gel result of the Long-PCR shows the primer pair is specific for the DUP7 variant at all annealing

temperatures (Figure 3.29); the annealing temperatures (62°C - 67°C) show very clear bands and generate most product, however, 64°C was used for sequencing.

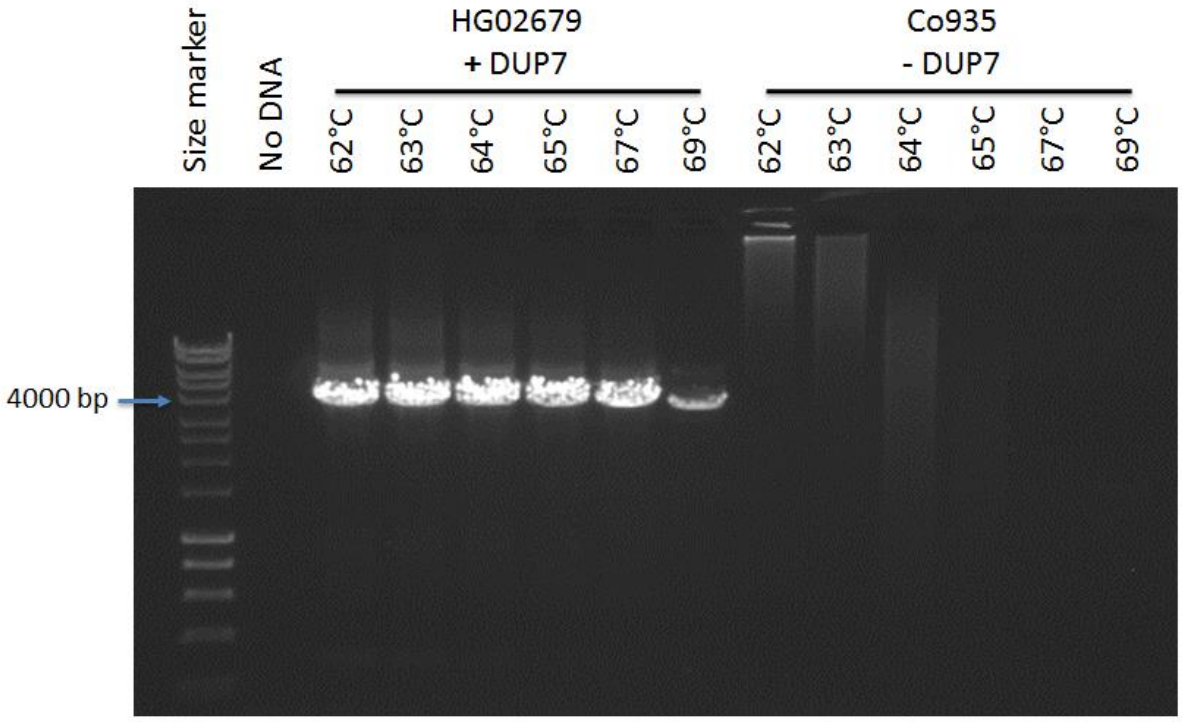



**Figure 3.29: Gradient long-PCR gel result for DUP7 using primer pair DUP7F and DUP7R.** There is successful amplification in HG02679 (DUP7 carrier) matching the expected 3,332 bp size for the fragment, indicating that the correct region was amplified. The specific bands of DUP7 can be seen at all annealing temperatures (62°C - 69°C). The size marker is the Bioline HyperLadder™ 1kb ladder.

Although, DUP7 PCR products were generated and extracted from an agarose gel in order to be sequenced, a lack of time meant that the sequencing could not be completed. However, DUP7 PCR products are ready to be sequenced later by the lab members.

### 3.3.4 Duplication type 14 (DUP14)

The 5kb windows plot generated suggests the possible breakpoint for DUP14 and shows that entire *GYPE* gene is duplicated in this positive control (Figure 3.30).

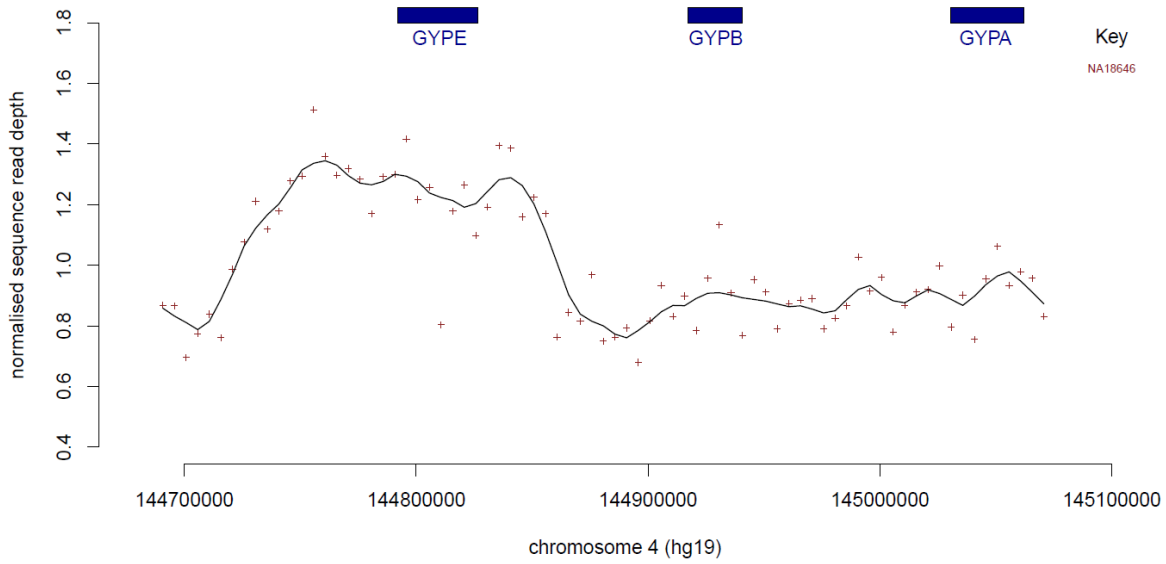



**Figure 3.30: Analysis of DUP14 duplications using next-generation sequencing data.** The points on the graph indicate the normalised read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

Fibre-FISH was applied on the known DUP14 (NA18646) from the 1000 Genomes Project sample collection using four different labelled fosmid probes (F7, F1, F12 and G10) in order to confirm the glycophorin variant (Chapter 2). The fibre-FISH result confirmed the existence of DUP14 in the positive control used for this analysis (Figure 3.31).

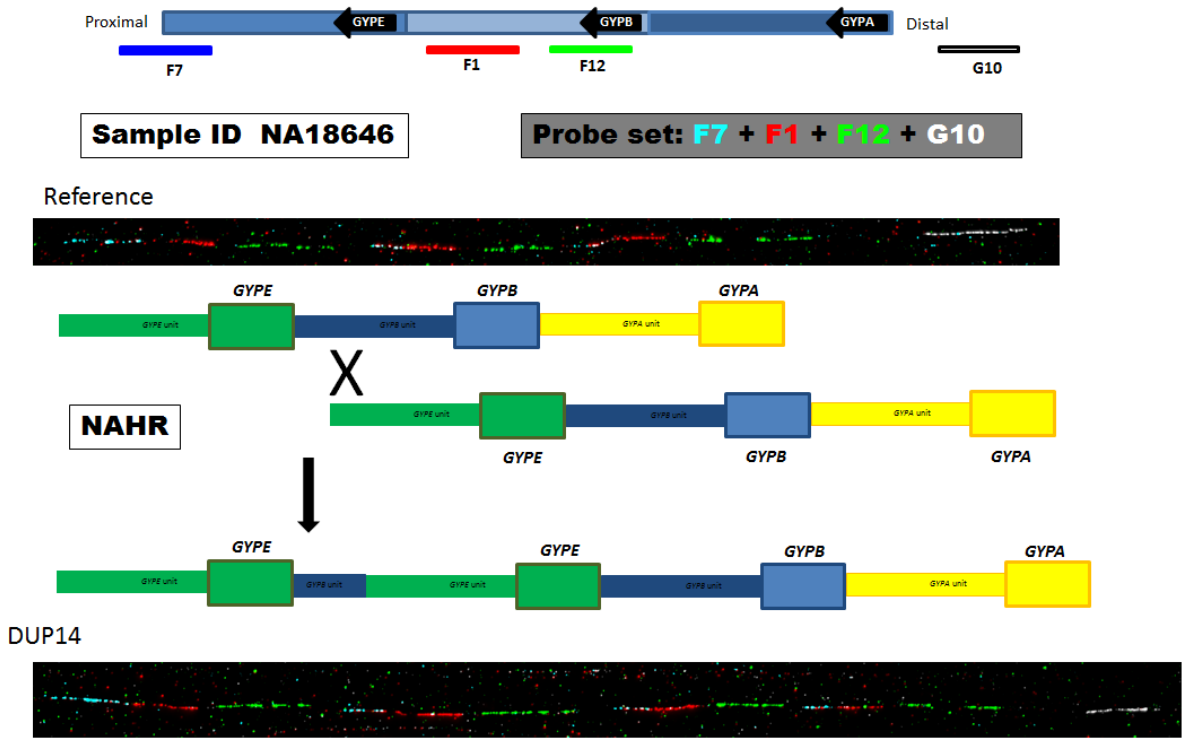



**Figure 3.31: Confirmation of DUP14 duplication using fibre-FISH analysis.** The top of the figure shows the names of the probes used for this analysis and their position on the genome. The two FISH results in the top and the bottom of the figure represent the fibre-FISH result for the reference haplotype compared with the DUP14 positive control fibre-FISH result. The middle of the figure shows a cartoon image of the DUP14 NAHR mechanism.

The specific primer pair for DUP14 was designed using the normalized sequence read depth data, the 5 kb windows plot and the fibre-FISH results (Chapter 2) in order to amplify the duplication breakpoints region and Sanger sequence the PCR product. The primers flank the breakpoints expected region of the DUP14 duplication (Figure 3.32).

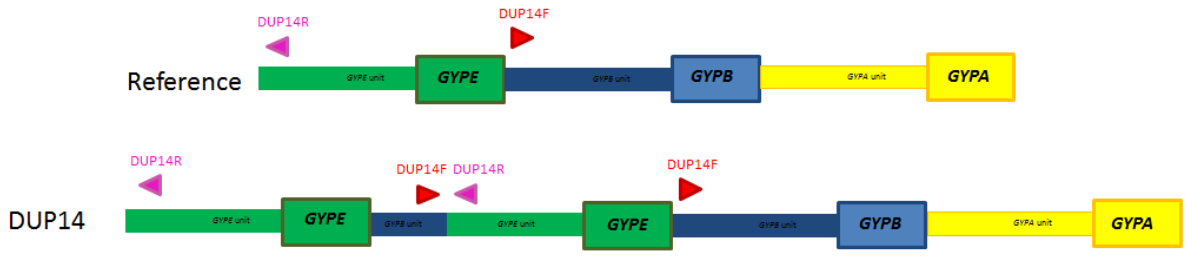



**Figure 3.32: Positions of the DUP14 primer pair used for the Long-PCR.** DUP14F with the red triangle represents the forward primer and DUP14R with the pink triangle represents the reverse primer.

Gradient long-PCR was applied using the DUP14 primer pair (Chapter 2) on a DUP14 positive DNA sample (NA18646) and a DUP14 negative DNA sample (Co075) in order to assess the specificity of the primers and select the best annealing temperature for the primers. The gel result for the Long-PCR shows that the primer pair is specific for DUP14 variant at

most annealing temperatures (Figure 3.33); however, 65°C generates most product, therefore it was used for sequencing.

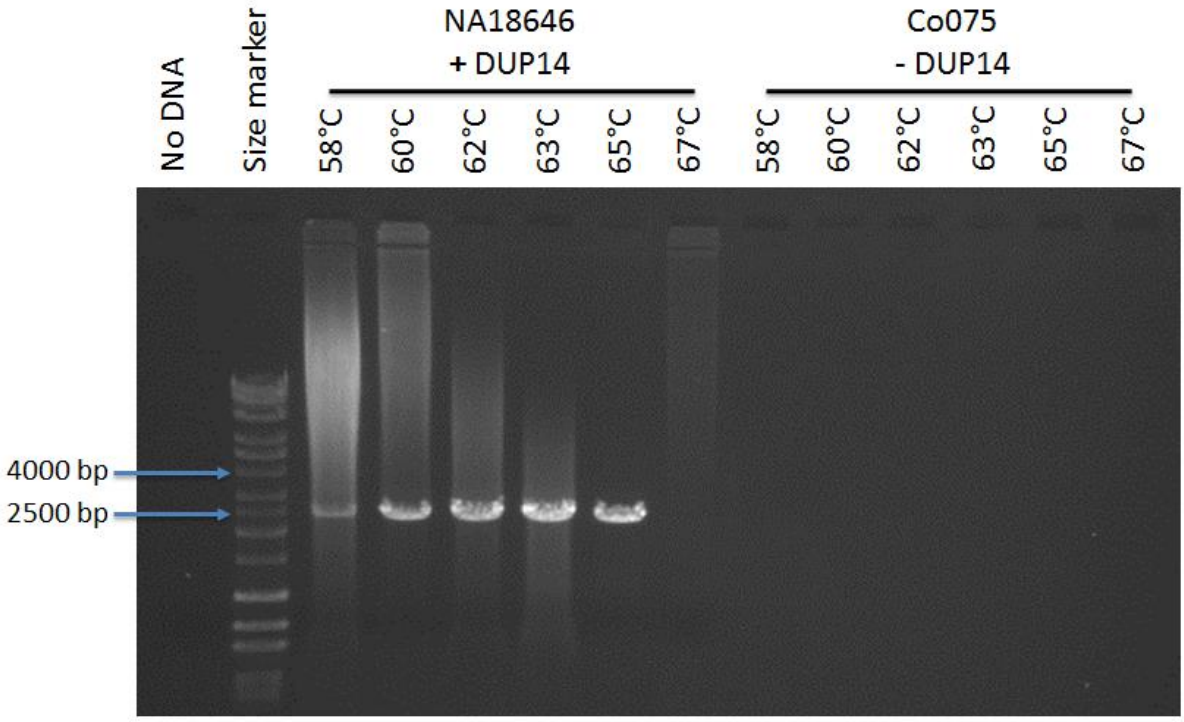



**Figure 3.33: Gradient long-PCR gel result for DUP14 using primer pair DUP14F and DUP14R.** There is successful amplification in NA18646 (DUP14 carrier) matching the expected 2,424 bp size for the fragment, indicating that the correct region was amplified. The specific bands of DUP14 can be seen at 58°C, 60°C, 62°C, 63°C and 65°C. The size marker is the Bioline HyperLadder™ 1kb plus ladder.

After sequencing DUP14 specific PCR product and the multiple sequence alignments, the breakpoints of this variant were identified by the switch from *GYPB* to *GYPE* paralogue specific variants. The alignment results shows that DUP14 proximal breakpoint is in *GYPB* unit with a range of 76 bp (chr4:144,853,613-144,853,688) (Table 3.1). The proximal breakpoint sequence maps to L1PA5, LINE repeat element. The distal breakpoint is in *GYPE* unit with a range of 76 bp (chr4:144,723,019-144,723,094) (Table 3.1). The distal breakpoint sequence maps to L1PA5, LINE repeat element. DUP14 covers the *GYPE* gene with the non-coding region upstream and downstream of the same gene (see Appendices 2 and 4 for sequencing stage primers and full alignments).

### 3.3.5 Duplication type 29 (DUP29)

The 5kb windows plot generated suggests the possible breakpoint for DUP29 and shows that part of the *GYPE* gene and part of *GYPB* gene are duplicated in this positive control (Figure 3.34).

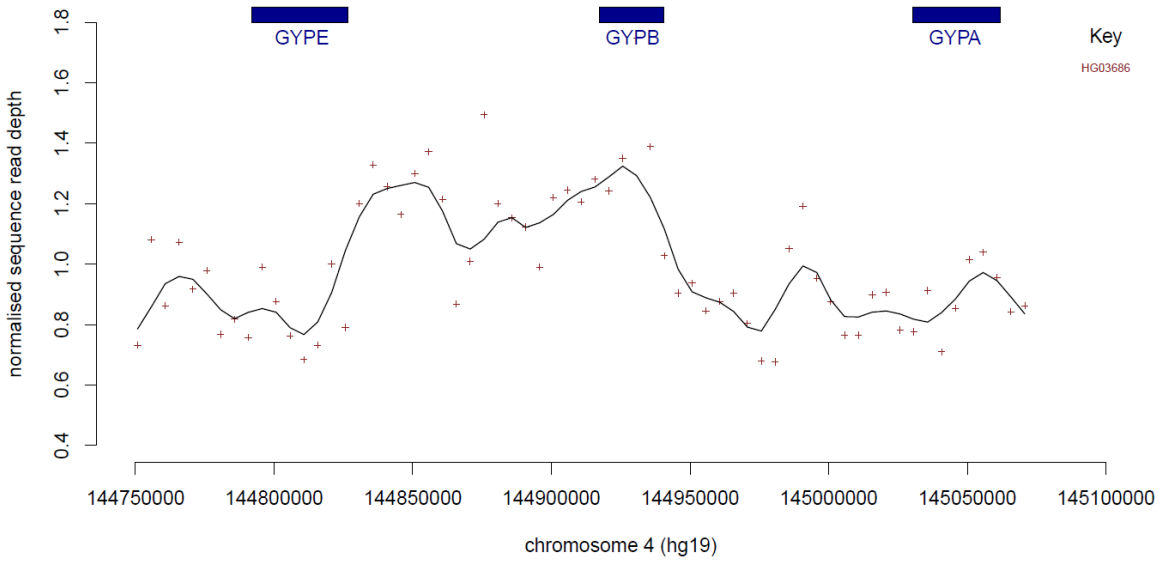



**Figure 3.34: Analysis of DUP29 duplications using next-generation sequencing data.** The points on the graph indicate the normalised read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

Fibre-FISH were applied on the known DUP29 (HG03686) from the 1000 Genomes Project sample collection using four different labelled fosmid probes (F7, F1, F12 and G10) in order to confirm the glycophorin variant (Chapter 2). The fibre-FISH result confirmed the existence of DUP29 in the positive control used for this analysis (Figure 3.35).

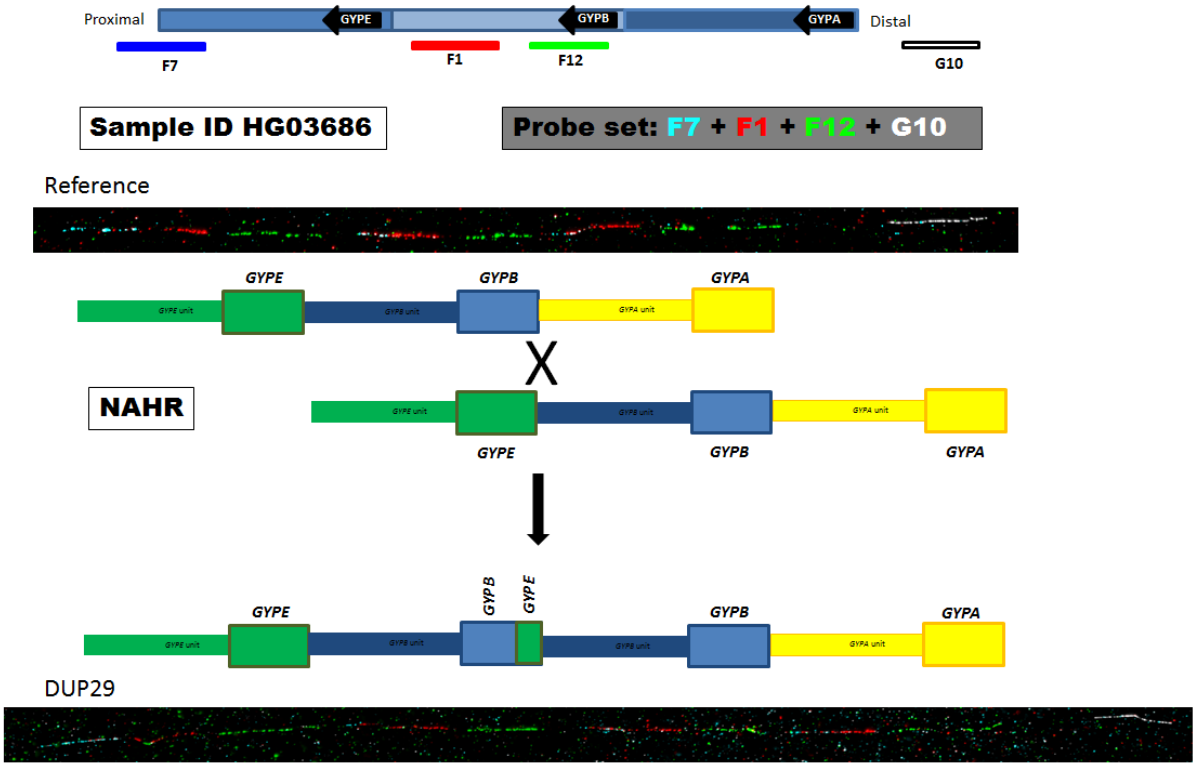



**Figure 3.35: Confirmation of DUP29 duplication using fibre-FISH analysis.** The top of the figure shows the names of the probes used for this analysis and their position on the genome. The two FISH results in the top and the bottom of the figure represent the fibre-FISH result for the reference haplotype compared with the DUP29 positive control fibre-FISH result. The middle of the figure shows a cartoon image of the DUP29 NAHR mechanism.

The specific primer pair for DUP29 was designed using the normalized sequence read depth data, the 5 kb windows plot and the fibre-FISH results (Chapter 2) in order to amplify the duplication breakpoints region and Sanger sequence the PCR product. The primers flank the breakpoints expected region of the DUP29 duplication (Figure 3.36).

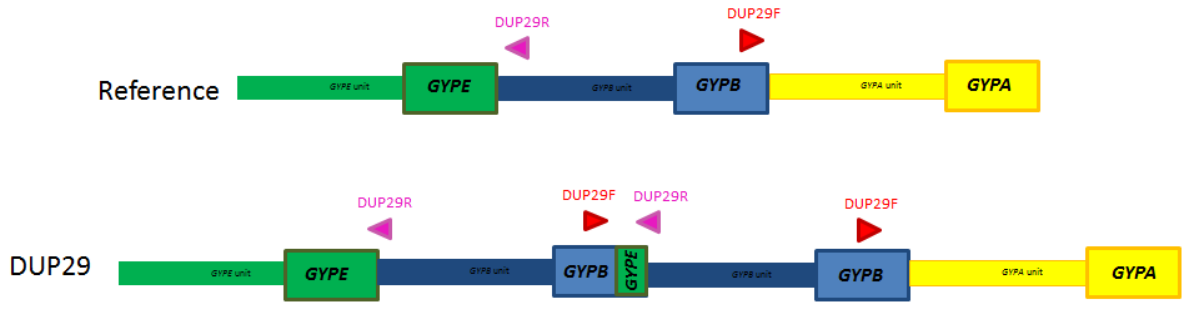



**Figure 3.36: Positions of the DUP29 primer pair used for the Long-PCR.** DUP29F with the red triangle represents the forward primer and DUP29R with the pink triangle represents the reverse primer.

Gradient long-PCR was applied using the DUP29 primer pair (Chapter 2) on a DUP29 positive DNA sample (HG03686) and a DUP29 negative DNA sample (Co075) in order to assess the specificity of the primers and select the best annealing temperature for them. The gel result of the Long-PCR shows that the primer pair is specific for DUP29 variant at 64°C and 65°C (Figure 3.37); however, 64°C generates most product, therefore it was used for sequencing.

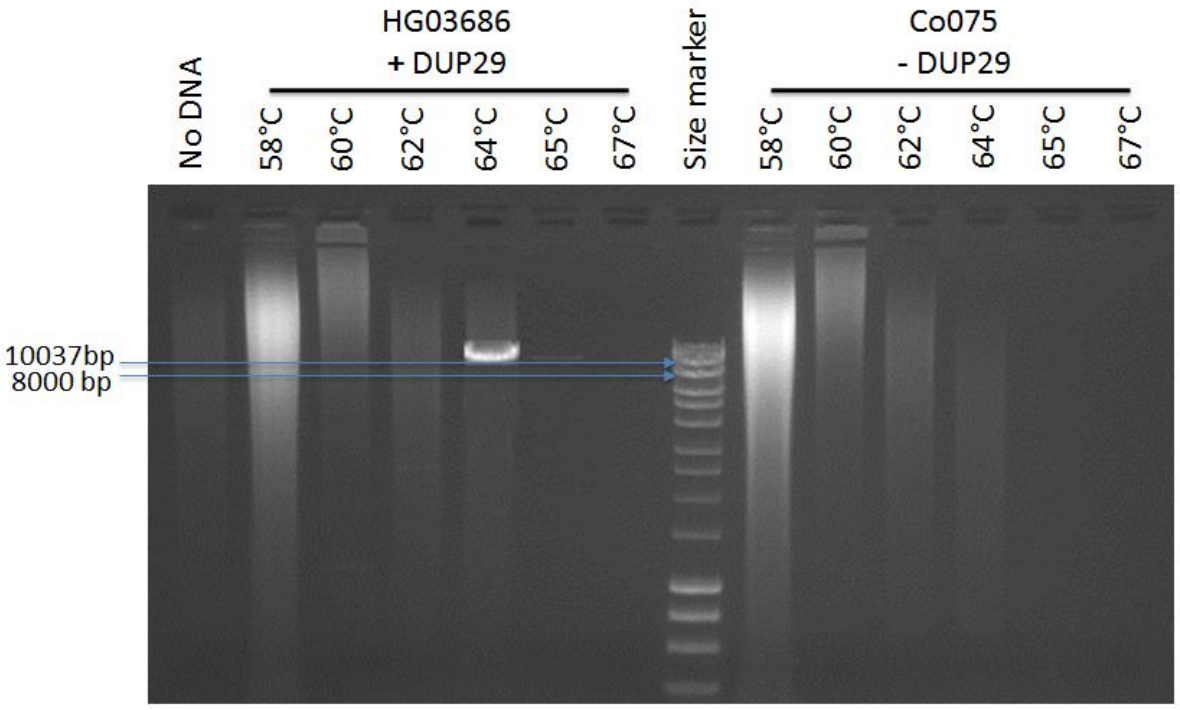



**Figure 3.37: Gradient long-PCR gel result for DUP29 using primer pair DUP29F and DUP29R.** There is successful amplification in HG03686 (DUP29 carrier) matching the expected 11,694 bp size for the fragment, indicating that the correct region was amplified. The specific bands of DUP29 can be seen at 64°C and 65°C. The size marker is the Bioline HyperLadder™ 1kb ladder.

After sequencing DUP29 specific PCR product and the multiple sequence alignments, the breakpoints of this variant were identified by the switch from *GYPB* to *GYPE* paralogue specific variants. DUP29 consists a partial duplication of *GYPE* gene and a partial duplication of *GYPB* gene. The alignment result shows that DUP29 proximal breakpoint is in *GYPB* unit with a range of 61 bp (chr4:144,939,393-144,939,453) (Table 3.1). The proximal breakpoint sequence maps to MIR, SINE repeat element. The distal breakpoint is in *GYPE* unit with a range of 61 bp (chr4:144,825,583-144,825,643) (Table 3.1). The distal breakpoint sequence maps to MIRb, SINE repeat element. DUP29 creates a hybrid gene (*GYPB-E*) covers a small part of the *GYPE* gene (1,072 bp) which includes exon 1 of the *GYPE* gene, the whole of the non-coding region downstream of the *GYPE* gene and most of the *GYPB* gene (22,215 bp)

which includes all exons (6, 5, 4, 3 and 2) of the *GYPB* gene except exon 1 (see Appendices 2 and 4 for sequencing stage primers and full alignments).

### 3.3.6 Duplication type 24 (DUP24)

The 5kb windows plot generated suggests the possible breakpoint for DUP24 and shows that the entire *GYPB* gene is duplicated in these positive controls (Figure 3.38).

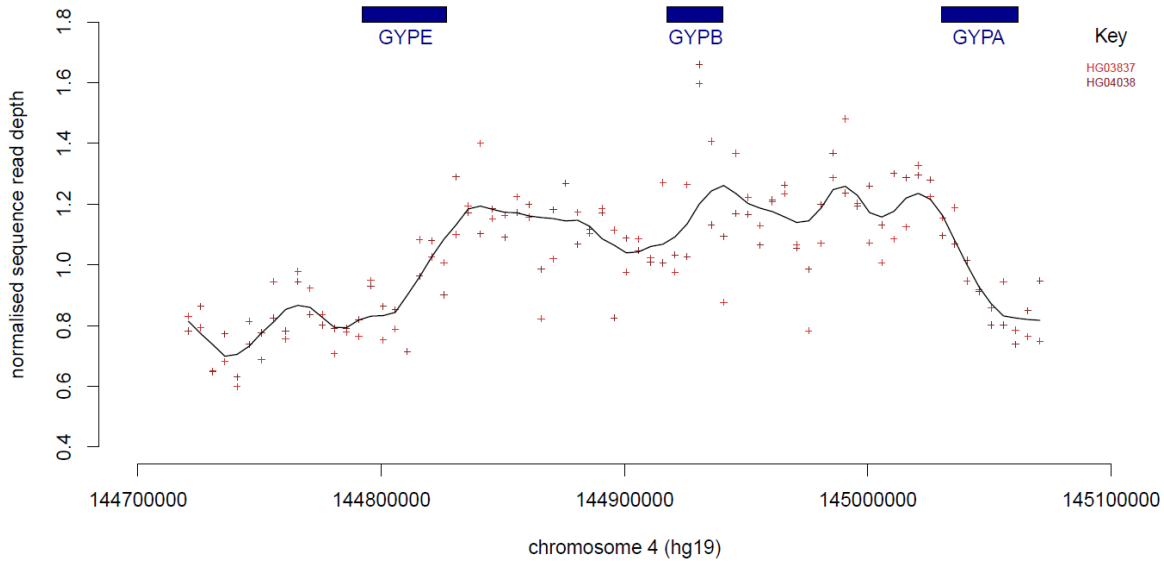



**Figure 3.38: Analysis of DUP24 duplications using next-generation sequencing data.** The points on the graph indicate the normalised read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

Fibre-FISH was applied on the known DUP24 (HG03837) from the 1000 Genomes Project sample collection using four different labelled fosmid probes (F7, F1, F12 and G10) in order to confirm the glycophorin variant (Chapter 2). The fibre-FISH result confirmed the existence of DUP24 in the positive control used for this analysis (Figure 3.39).

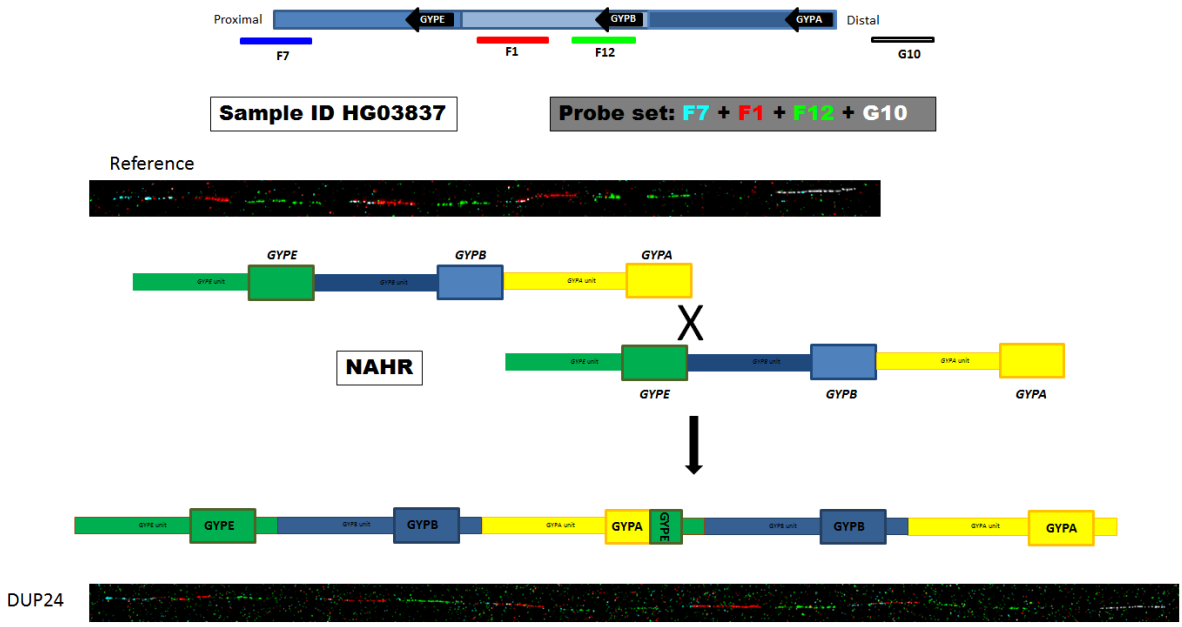



**Figure 3.39: Confirmation of DUP24 duplication using fibre-FISH analysis.** The top of the figure shows the names of the probes used for this analysis and their position on the genome. The two FISH results in the top and the bottom of the figure represent the fibre-FISH result for the reference haplotype compared with the DUP24 positive control fibre-FISH result. The middle of the figure shows a cartoon image of the DUP24 NAHR mechanism.

The specific primer pair for DUP24 was designed using the normalized sequence read depth data, the 5 kb windows plot and the fibre-FISH results (Chapter 2) in order to amplify the duplication breakpoints region and Sanger sequence the PCR product. The primers flank the breakpoints expected region of the DUP24 duplication (Figure 3.40).

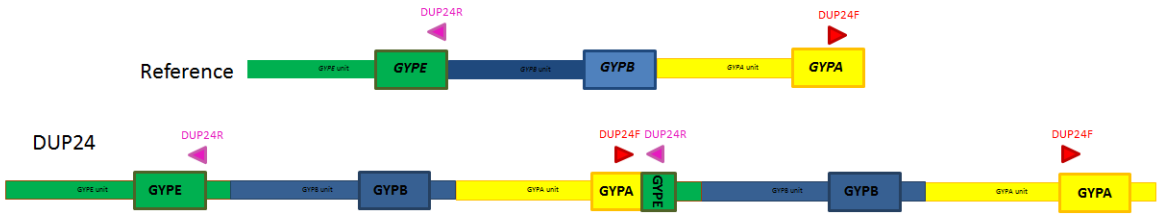



**Figure 3.40: Positions of the DUP24 primer pair used for the Long-PCR.** DUP24F with the red triangle represents the forward primer and DUP24R with the pink triangle represents the reverse primer.

Gradient long-PCR was applied using the DUP24 primer pair (Chapter 2) on a DUP24 positive DNA sample (HG03837) and a DUP24 negative DNA sample (Co075) in order to assess the specificity of the primers and select the best annealing temperature for them. The gel result of the Long-PCR shows that the DUP24 primer pair is not specific to the DUP24 variant at any annealing temperature (Figure 3.41).

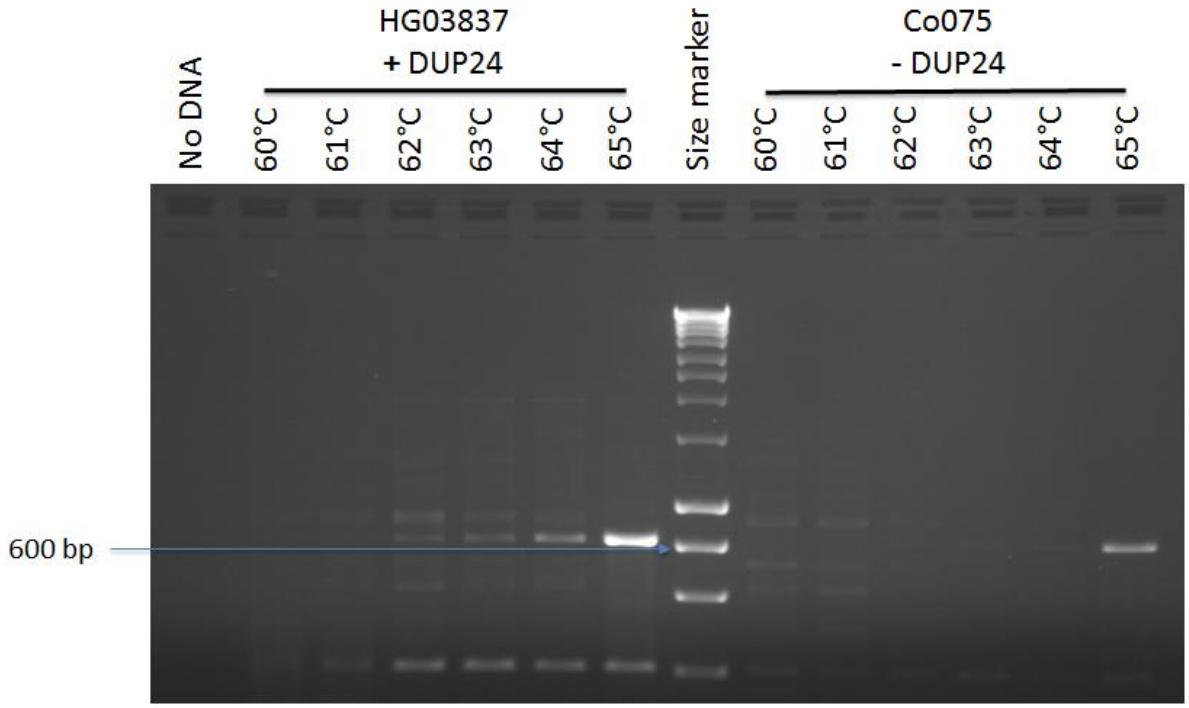



**Figure 3.41: Gradient long-PCR gel result for DUP24 using primer pair DUP24F and DUP24R.** There are amplifications for HG03837 (DUP24 carrier) and Co075 (negative for DUP24), with a high degree of non-specific binding at most of the annealing temperatures, indicating that the target region was not amplified correctly. No specific bands for DUP24 were observed at all the annealing temperatures (60°C - 65°C). The size marker is the Bioline HyperLadder™ 1kb ladder.

More primer pairs were designed for this variant; however, some of them did not show any product and some were not specific to the DUP24 variant. Therefore, it was not possible to optimise any of the primers designed for DUP24. As a consequence, HG03837 was not sequenced and the DUP24 breakpoints remain unknown.

### 3.4 Distinguishing *GYPE*, *GYPB* and *GYPA* repeat units using fibre-FISH

Given the high proportion of sequence identity between the tandemly repeated glycophorin regions, there is extensive cross-hybridization of probes that map to the *GYPB* repeat with the *GYPE* and *GYPA* repeats. As a result, it was difficult to distinguish *GYPE* from *GYPB* and *GYPA*. However, fibre-FISH showed that the reference haplotype generates signals consistent with the genome reference sequence (Figure 3.42), and that the *GYPA* repeat has a specific insertion that can be identified and visualized by a gap in the green fosmid probe signal, caused by 16kb of unique sequencing in the *GYPA* if using a *GYPB* unit probe (Figure 3.42). However, *GYPE* has a smaller specific insertion, which is too small to see as a gap in the probe signal generated from the fosmid. However, a specific probe (3,632 bp) can be generated by a specific *GYPE* long-PCR assay to distinguish the *GYPE* repeat by hybridization of the small *GYPE*-repeat-specific PCR product (Figure 3.42).

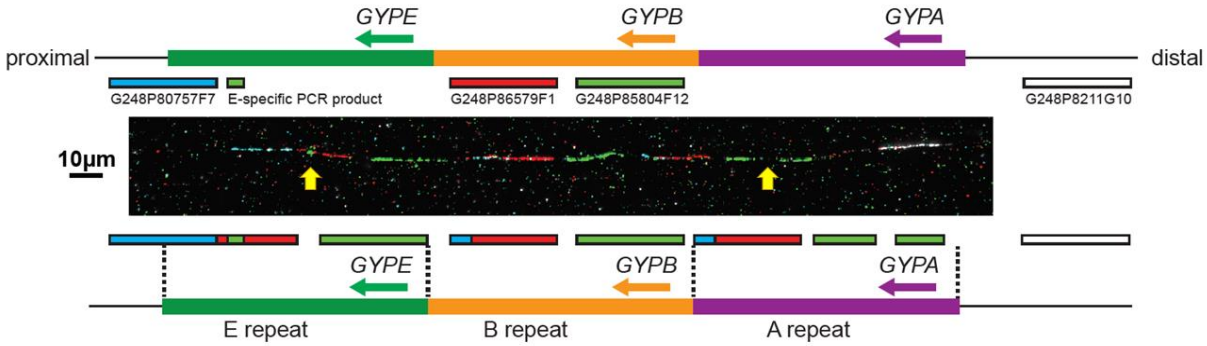



**Figure 3.42: An example DNA fibre from the reference haplotype with the *GYPE*-specific probe.** The position and label colour of the fosmid probes is indicated above the fibre on a representation of the human reference genome, and the interpretation of the FISH signals is shown below the fibre. The left hand yellow arrow points to the *GYPE*-specific probe, while the right hand yellow arrow points to the *GYPA* gap. The figure was obtained from Algady *et al*. (2018).

This specific *GYPE* PCR assay was designed, then Paulina Brajer (MSc) sequenced the PCR products after the long-PCR stage (Chapter 2) to generate a specific *GYPE* probe used for the fibre-FISH analysis in order to characterise the structure of the complex variations in DUP5. However, although the specific *GYPE* fibre-FISH probe was initially generated for DUP5, it was also utilized as an extra confirmation probe for other variants. The gel result for the Long-PCR shows that the optimum annealing temperature for the specific *GYPE* PCR primer pair (Specific_*GYPE*_F and Specific_*GYPE*_R) is 63°C. As expected, amplification occurred for each different reference DNA sample that was used showing the expected size band (Figure 3.43).

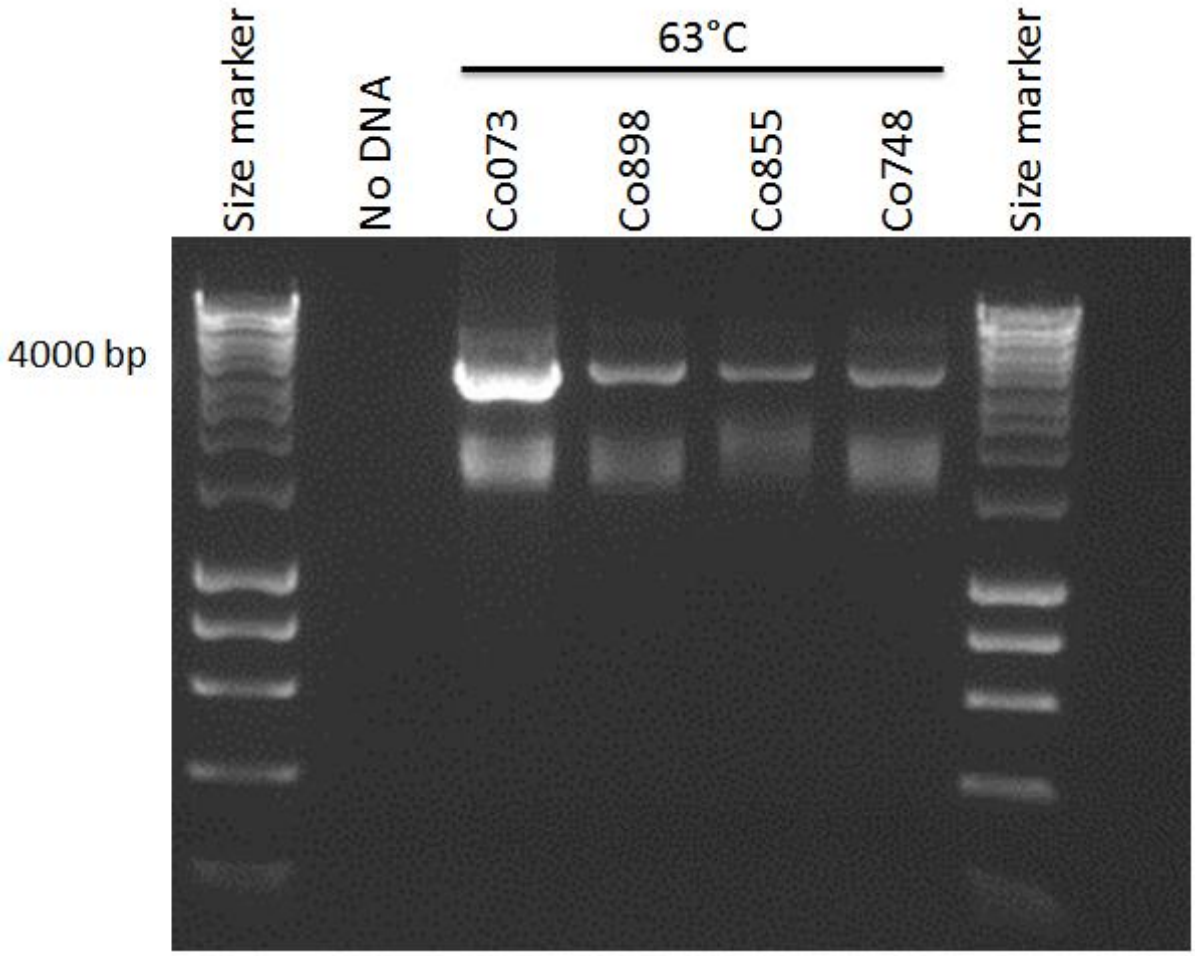



**Figure 3.43: Long-PCR gel result for the specific GYPE PCR primer pair.** There are successful amplifications at the 63°C annealing temperature with the band matching the expected 3,632 bp size for the fragment, indicating that the correct region had been amplified. The bands can be seen for all of the DNA samples (HRC samples) that were used. The size marker is the Bioline HyperLadderTM 1kb Plus ladder.

More PCR products were generated and extracted from an agarose gel in order to confirm the specificity of the PCR assay by sequencing the PCR product (Chapter 2). After sequencing the product from the forward and reverse sides (Chapter 2), the result of the sequencing reaction was provided in the form of a chromatogram. The sequence shows clear peaks and low background noise (Figure 3.44).

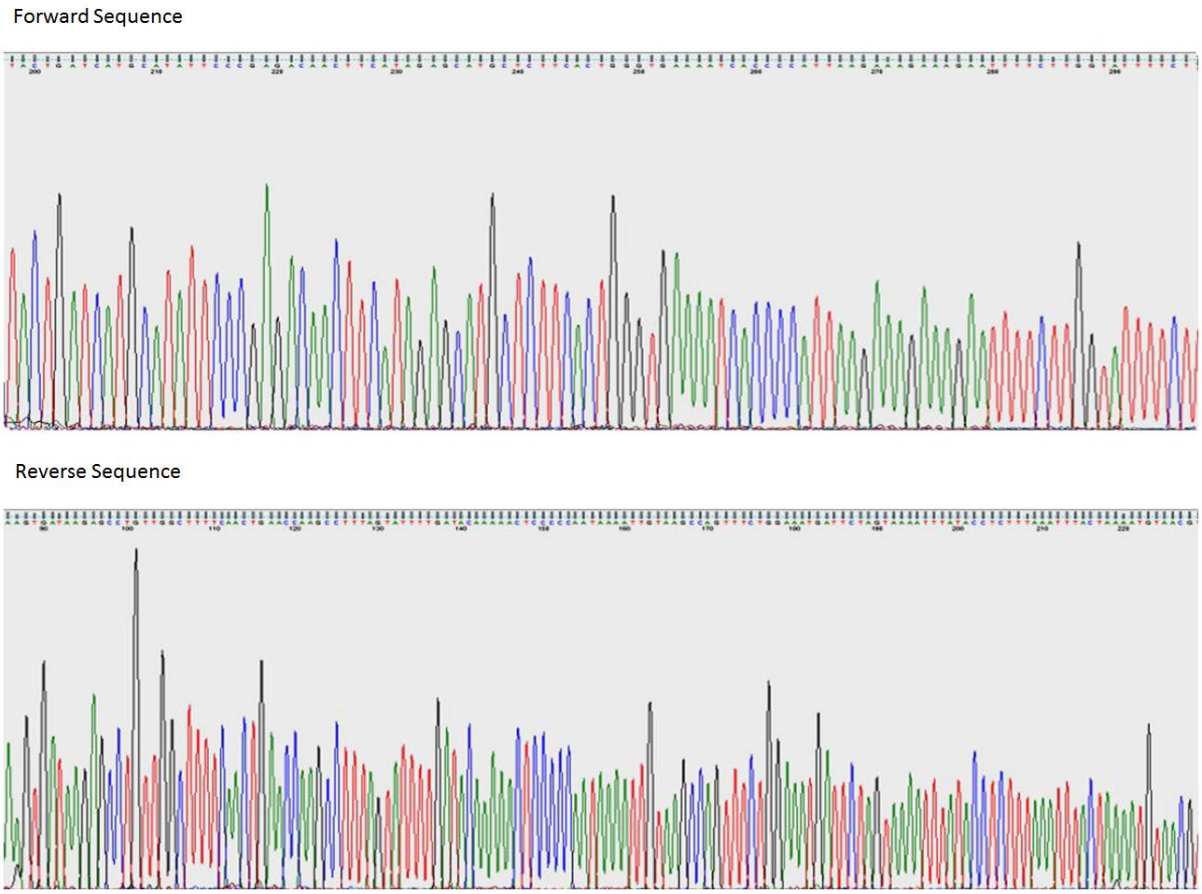



**Figure 3.44: Sequencing result of the specific GYPE primer set.** The figure shows the chromatogram obtained from Sanger sequencing result, detailing the sequence obtained from the reaction.

The sequence obtained from the chromatogram was then blated using the UCSC genome browser (GRCh37/hg19) "BLAT" tool (Chapter 2). The genome browser results indicate that the specific *GYPE* product extracted is specific to *GYPE*. The sequencing result is positioned directly above the gene of interest, indicating that the sequence matches a region within *GYPE* (Figure 3.45).

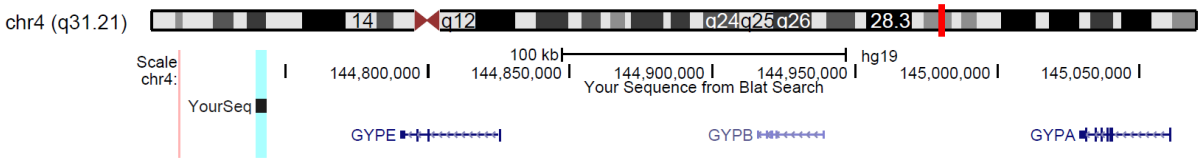



**Figure 3.45: Genome browser blat result for specific GYPE sequencing.** The image above shows the position of the sequencing result, and genes that relate to the sequence. The light blue highlight indicates the specific *GYPE* sequence (yourseq) and the blue various genes are indicated at the bottom of the image.

## 3.5 Characterisation and identification of the *GYPs* complex duplication (DUP5)

A complex duplication was identified by the Leffler *et al*. (2017) study, duplication 5 (DUP5). The frequency data in populations showed that unlike DUP4, which is found primarily in East African populations, DUP5 is mainly found in West African populations. This variant is particularly intriguing as suggestions have been made about the complexity with potential inversions and triplications among the variants present. The study has suggested that DUP5 contains two forms of CNVs, one triplication and two duplications, however, they did not provide specifics of the structure or what genes were subject to these variations.

The 5kb windows plot generated cannot suggest the possible breakpoints for DUP5 as it shows different high and low values (Figure 3.46). Therefore, the specific *GYPE* fibre-FISH probe was generated (Section 3.4).

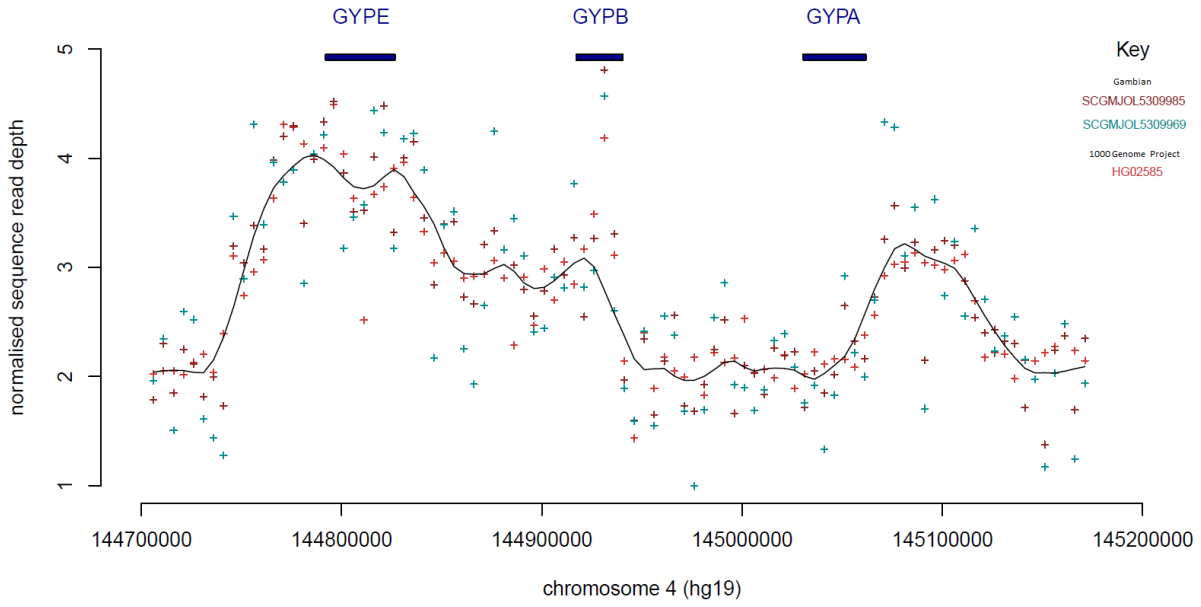



**Figure 3.46: Analysis of DUP5 complex duplication using next-generation sequencing data.** The points on the graph indicate the normalised read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples. The first two samples are from Gambian Genome Diversity Project data, which have been downloaded by an MSc student Eleanor Weyell.

Fibre-FISH was applied on the known DUP5 (HG02585) from the 1000 Genomes Project sample collection using four different labelled fosmid probes (F7, F1, F12, G10) and the specific *GYPE* PCR product in order to confirm the *GYP* complex variant (Chapter 2). The fibre-FISH result confirmed the existence of DUP5 in the positive control used for this analysis (Figure 3.47). In the figure confirms the expected duplication part of *GYPE* by the specific *GYPE* probe.

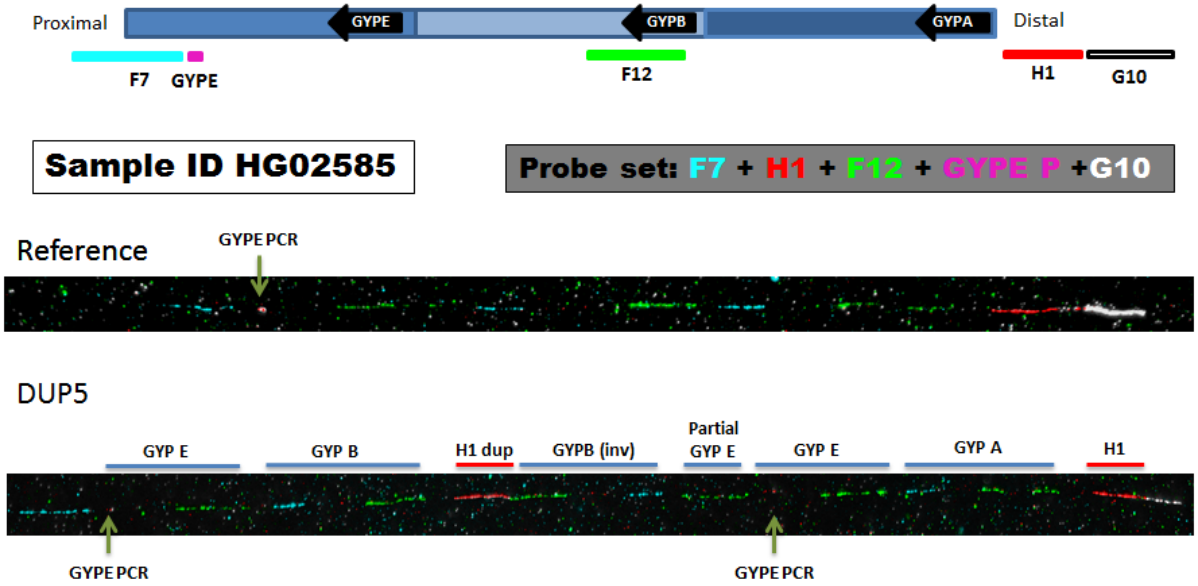



**Figure 3.47: Confirmation of DUP5 complex duplication using fibre-FISH analysis.** The top of the figure shows the names of the probes used for this analysis and their position on the genome, as well as the specific *GYPE* probe that was used. The two FISH results at the bottom of the figure represent the reference haplotype fibre-FISH result compared with the DUP5 positive control fibre-FISH result.

The specific primer pair for DUP5 was designed using the normalized sequence read depth data, the 5 kb windows plot and the fibre-FISH results (Chapter 2) in order to amplify the duplication breakpoints region and Sanger sequence the PCR product. The primers flank the most crucial breakpoints event expected region of the DUP5 complex duplication (Figure 3.48).

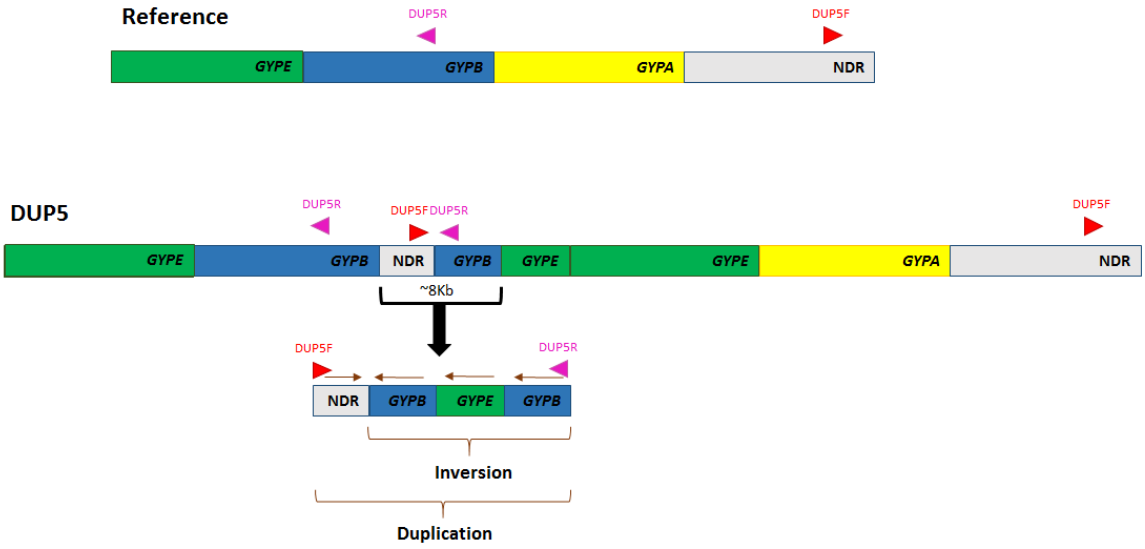



**Figure 3.48: Positions of the DUP5 primer pair used for the Long-PCR.** DUP5F with the red triangle represents the forward primer and DUP5R with the red triangle represents the reverse primer. NDR means non duplicated region.

Gradient long-PCR was applied using the DUP5 primer pair (Chapter 2) on a DUP5 positive DNA sample (HG02585) and a DUP5 negative DNA sample (Co075) in order to assess the specificity of the primers and select the best annealing temperature for them. The gel result of the Long-PCR shows the primer pair is specific for the DUP5 variant at all of the annealing temperatures (65°C - 67°C) (Figure 3.49). Because the gel result showed good bands for all of the gradient annealing temperatures, all of the bands were extracted from an agarose gel in order to be sequenced.

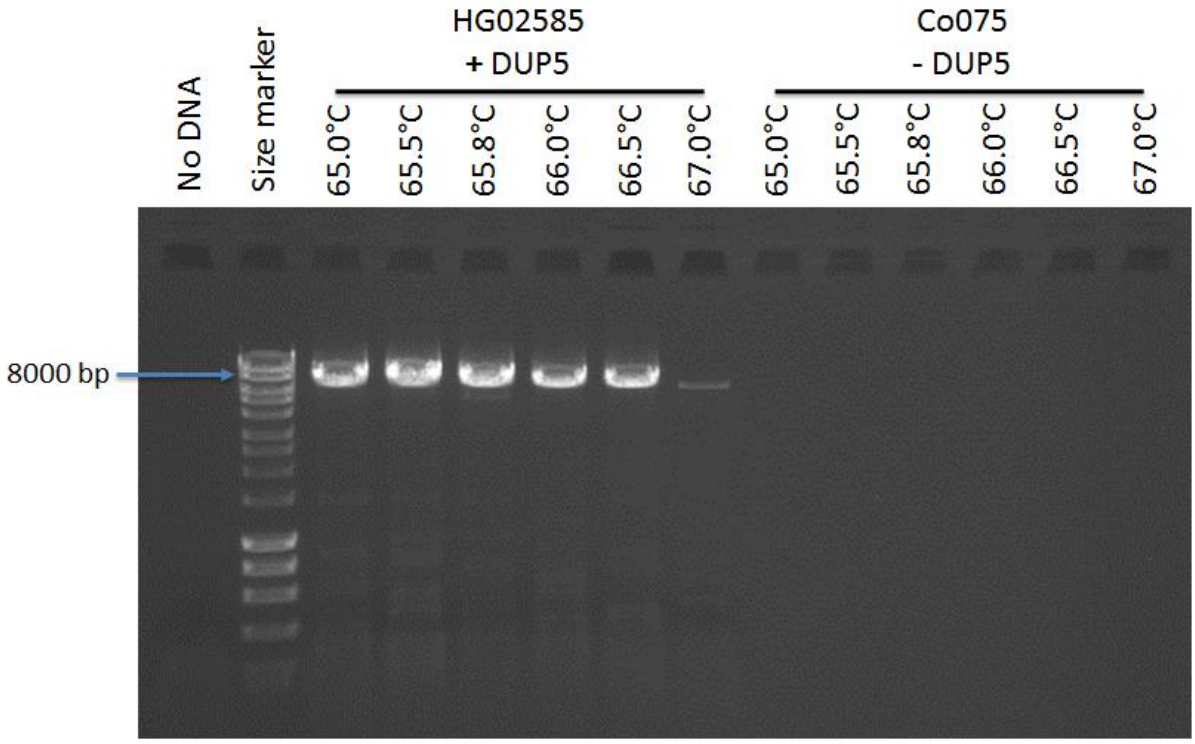



**Figure 3.49: Gradient long-PCR gel result for DUP5 using primer pair DUP5F and DUP5R.** There is successful amplification only in HG02585 (DUP5 carrier) with the band matching the expected 7,854 bp size for the fragment, indicating that the correct region was amplified. The bands can be seen at (65°C - 67°C). The size marker is the Bioline HyperLadder™ 1kb Plus ladder.

After sequencing DUP5 PCR products and the multiple sequence alignments using, the breakpoints of this complex variant were identified by the switches from NDR to *GYPB*, *GYPB* to *GYPE* and *GYPE* to *GYPB* paralogue specific variants. The multiple alignment result shows that DUP5 has three breakpoints; the breakpoint A has an exact proximal breakpoint in the non-duplicated region downstream of the *GYPA* gene (chr4:145,113,700) and an exact distal breakpoint is in *GYPB* unit (chr4:144,936,865) (Table 3.2). The proximal breakpoint sequence maps to MLT1J2, LTR repeat element, while the distal breakpoint sequence does not map to any repeat element. The breakpoint B has a proximal breakpoint in *GYPB* unit with a range of 13 bp (chr4:144,933,933-144,933,945) and a distal breakpoint, which is in *GYPE* unit with a range of 13 bp (chr4: 144,820,933-144,820,945) (Table 3.2). The proximal breakpoint sequence maps to L1PA2, LINE repeat element, while the distal breakpoint sequence maps to L1MB5, LINE repeat element. The breakpoint C has a proximal breakpoint in *GYPE* unit with a range of 32 bp (chr4:144,819,856-144,819,887) and a distal breakpoint, which is in *GYPB* unit with a range of 32 bp (chr4: 144,932,860-144,932,891) (Table 3.2). The proximal breakpoint sequence does not map to any repeat element, while the distal breakpoint sequence does not map to any repeat element.

This complex variant (DUP5) has many events, and the current result shows that DUP5 has a duplication of the non-coding region downstream of the *GYPA* gene given a total of 299 bp, and an inverted (*GYPB-E-B*) gene conversion event (Figure 3.48). This event includes a partial inverted duplication of the *GYPB* gene given a total of 2,932 bp, partial inverted duplication of the *GYPE* gene given a total of 1,088 bp and partial inverted duplication of the *GYPB* gene given a total of 2,848 bp (see Appendices 2 and 4 for sequencing stage primers and full alignments).

## 3.6 Characterisation and identification of a *GYPs* gene conversion (*GYPE-B-E*)

### 3.6.1 Detection and primer optimisation

Three samples of the 1000 Genomes Project were expected to be carriers of a gene conversion allele from the sequence read depth data of these three samples. The 5 kb window plot was created and the resulting plot shows that one of the 5kb read windows on the *GYPB* represents a high value of the normalized sequence read depth for the three samples, while one of the 5kb read windows on the *GYPE* represents a low value of the normalized sequence read depth for the three samples (Figure 3.50).

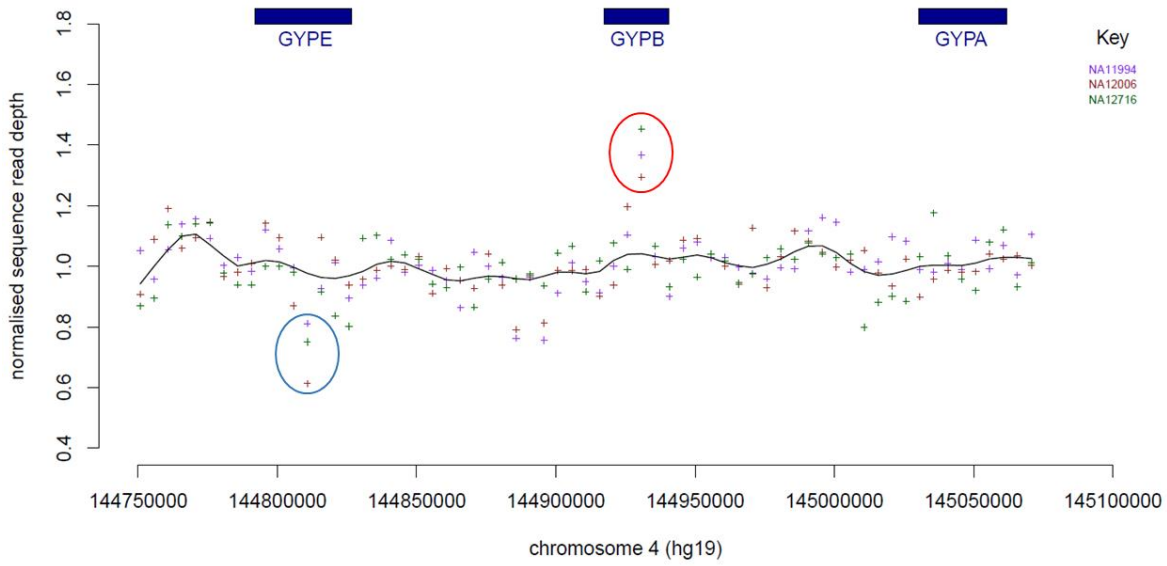



**Figure 3.50: Analysis of GYPE-B-E using next-generation sequencing data.** The points on the graph indicate the normalised read depth at this position on the chromosome. The colour indicates which DNA sample the data corresponds to. The line across the graphs indicates the average overall read depth across samples.

These changes in the read depth values suggest initial evidence of a gene conversion event between *GYPB* and *GYPE* (Figure 3.50). Gene conversion is a homologous recombination

mechanism that includes a unidirectional transfer of genetic fragment from a donor sequence (*GYPB*) to a highly homologous acceptor sequence (*GYPE*). Gene conversion could be happened because of multiple crossovers or as a result of the DNA double-strand break repairs (Ira *et al*., 2003). The *GYPE-B-E* is a non-allelic gene conversion. The non-allelic gene conversion could be occurred in trans between non-allelic gene copies residing on sister chromatids (more likely to be the case of the *GYPE-B-E*), in cis between non-allelic gene copies residing on the same chromatids or in trans between non-allelic gene copies residing on homologous chromosomes (Chen *et al*., 2007).

The normalized sequence read depth data, and the 5 kb windows results, indicate that some of the *GYPB* was added to the *GYPE*; with the result that a *GYPE-B-E* allele appeared. Consequently, a specific primer pair for the *GYPE-B-E* allele was designed using the information available for these samples (Chapter 2) r to amplify the gene conversion breakpoints region and Sanger sequence the PCR product. The primers flank the entire expected gene conversion region (Figure 3.51).

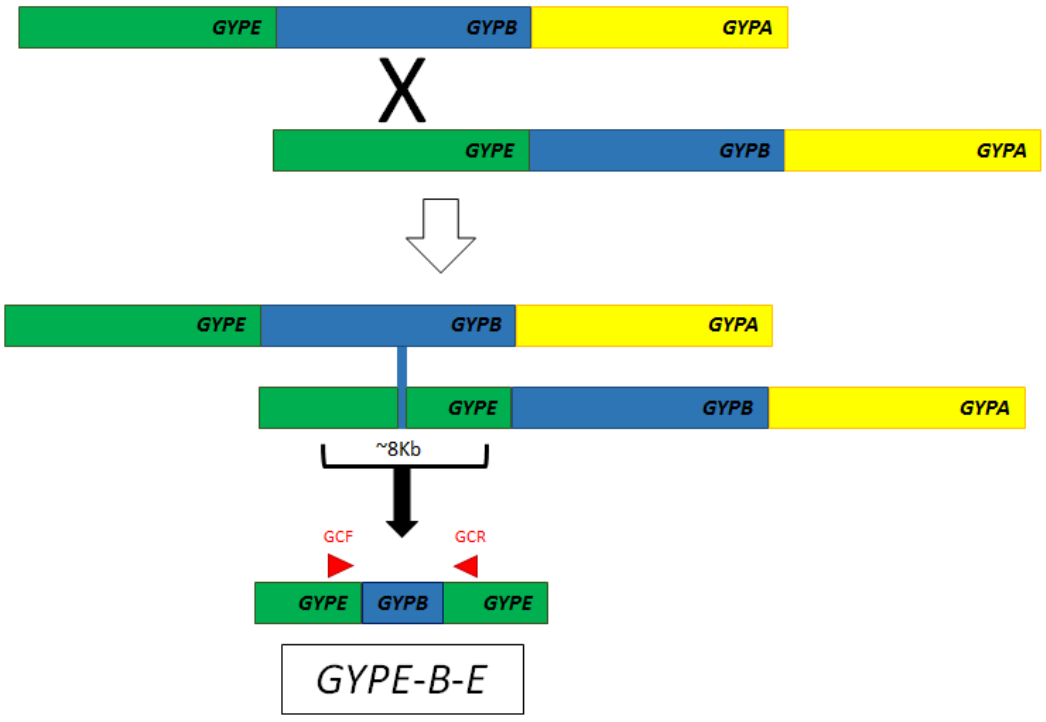



**Figure 3.51: A scheme of the gene conversion event mechanism and its expected structure.** The diagram shows two simple steps in the gene conversion event. The bottom figure shows that a duplicated fragment of the *GYPB* gene has been inserted into the middle of the *GYPE* gene. The two red triangles in the bottom diagram represent the positions of the gene conversion primer pair used for the long-PCR, the GCF is the forward primer and GCR is the reverse primer.

Only two of these three positive control samples were used for the Long-PCR assay. The gel result shows that the optimum annealing temperature for the *GYPB-E-B* gene conversion primer pair (*GYPE*_GC_F and *GYPE*_GC_R) is 65°C. As expected, an amplification occurred for the reference DNA sample and positive controls for the gene conversion (NA12716 and NA11994), producing the expected size band (Figure 3.52). This primer pair amplifies and produces the expected PCR product with any DNA sample as it has been designed not to be specific for any variant; instead, it only covers the expected gene conversion region.

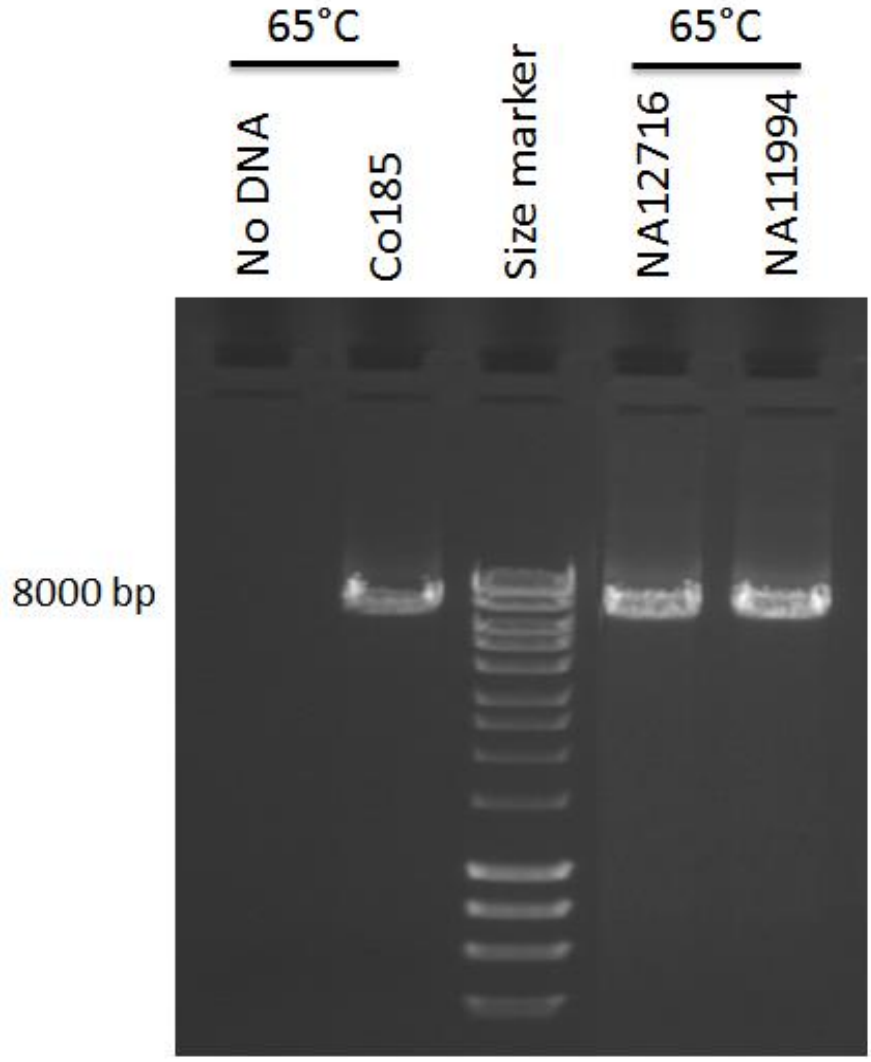



**Figure 3.52: Long-PCR gel result for the *GYPE-B-E* gene conversion primer pair.** There are successful amplifications at the 65°C annealing temperature for the positive and negative controls for the gene conversion with the band matching the expected 7,394 bp size for the fragment, indicating that the correct region was amplified. The size marker is the Bioline HyperLadderTM 1kb ladder.

### 3.6.2 Confirmation of the existence of the *GYPE-B-E* gene conversion by an RFLP assay

In order to differentiate between the gene conversion carrier samples and reference samples, and prior to sequencing the PCR products of the positive controls to determine the gene conversion breakpoints, an RFLP assay was designed to confirm the three candidate samples were carrying the *GYPE-B-E* allele for comparison with normal samples. The assay was adjusted to cut at the AAGCTT restriction site, which occurs once in the PCR product if the entire PCR product is *GYPE*; there are two restriction sites if the PCR product combines the expected *GYPE* and *GYPB*. The expected result for this assay was one cut for the reference samples (non-gene conversion carrier), creating two products, and two cuts for the gene conversion positive controls, creating three products (Figure 3.53).

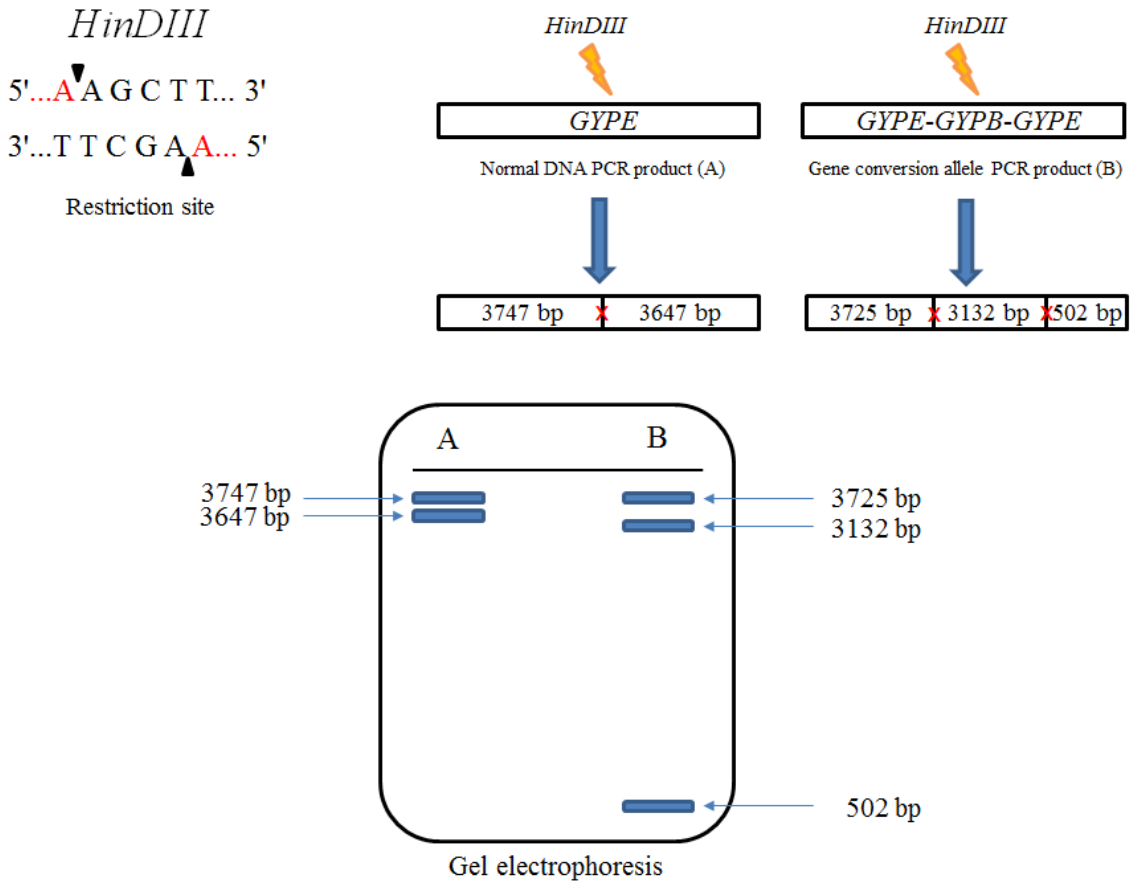



**Figure 3.53: RFLP assay expected gel electrophoresis result.** The restriction enzyme (*Hin*DIII) cuts the PCR product (A) at one site (A/AGCTT) and provides two very similarly sized products, whereas it cuts the PCR product (B) at two sites (A/AGCTT) and provides three products. The cutting sites are represented by red crosses.

The RFLP assay was applied to the remaining *GYPB-E-B* gene conversion primer pair long-PCR products of the positive and negative controls to the gene conversion. The RFLP gel confirms the presence of the *GYPE-B-E* gene conversion allele in the candidate samples

(NA12716 and NA11994) by having three products, while the Co185 (normal DNA sample) has only two products. However, the gel was run for a long time in order to separate the similarly sized bands from each other and to enable the gel to be clearly visualised under the UV light. As a result, the smallest size band (502 bp) ran out of gel, and the gel only shows two bands for each sample. However, the two band sizes for the positive control differ from the two bands obtained for the normal sample, as was expected (Figure 3.54)

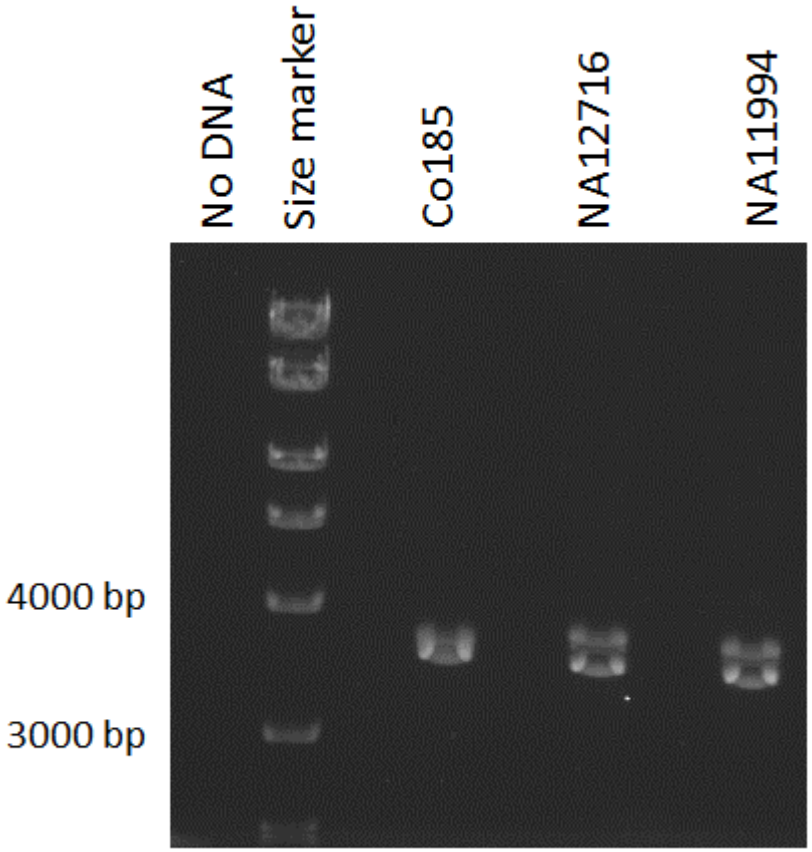



**Figure 3.54: RFLP assay gel electrophoresis result.** The *HinDIII* restriction enzyme was successfully cut for all of the samples. Co185 sample shows two very close bands (3747 bp and 3647 bp). Both NA12716 and NA11994 samples (positive controls) show two less close bands (3725 bp and 3132 bp). The third expected band (502 bp) was not shown, however, it should be below these bands. The size marker is the Bioline HyperLadderTM 1kb ladder.

After sequencing *GYPE-B-E* gene specific PCR product for NA12716 and NA11994 (positive control samples) and the multiple sequence alignments, the breakpoints of the gene conversion were identified by the switches from *GYPE* to *GYPB* and *GYPB* to *GYPE* paralogue specific variants. The multiple alignment result shows that *GYPE-B-E* gene conversion has two breakpoints; the breakpoint A has a proximal breakpoint in *GYPE* unit with a range of 6 bp (chr4:144,807,057-144,807,062), while the distal breakpoint in *GYPB* unit with a range of 6 bp (chr4:144,925,341-144,925,346) (Table 3.2). The proximal and

distal breakpoint sequences do not map to any repeat element. The breakpoint B has a proximal breakpoint in *GYPB* unit with a range of 147 bp (chr4:144,927,782-144,927,928), while the distal breakpoint is in *GYPE* unit with a range of 147 bp (chr4:144,809,511-144,809,657) (Table 3.2). The proximal and distal breakpoint sequences do not map to any repeat element. The *GYPE-B-E* gene conversion contains a duplication of *GYPB* and a deletion of *GYPE*. The duplicated *GYPB* fragment is 2,441 bp in size and the deleted *GYPE* fragment is 2,455 bp in size (see Appendices 2 and 4 for sequencing stage primers and full alignments).

## 3.7 Discussion

A total of 12 variants were detected and confirmed using high throughput sequencing data and fibre-FISH at the Wellcome Sanger Institute. The variants were analysed using the long-PCR technique and Sanger sequencing, and the breakpoints were identified using the MAFFT alignment program. The resolution of the breakpoint mapping was very high and more specific, as it depended on the spacing of single nucleotide variants distinguishing the *GYP* units by aligning all of the Sanger sequencing results of the PCR product of the variant against *GYPE*, *GYPB* and *GYPA*. However, breakpoints were only identified for eight of the 12 variants. All of the variants appeared to be gains or losses of whole numbers of glycophorin repeat units (Table 3.1), except DUP5 and the gene conversion variant (Table 3.2).

There was none of the variants shared the same breakpoints with any of with any other variant. However, DEL2 and DEL1 overlap together with 32.518 kb (chr4:144,913,001 - 144,945,518) covering the entire *GYPB* gene. DEL1 and DEL7 overlap with 66.176 kb (chr4:144,835,160 - 144,901,335) covering the non-coding region between the *GYPE* and *GYPB* genes. In addition, DUP29 and DUP14 overlap with 27.969 kb (chr4: 144,825,644 - 144,853,612) covering a part of the *GYPE* (chr4:144,825,644 - 144,826,716), which includes the last exon of the *GYPE* gene. DEL6 overlaps with all of the deletion variants as it is the largest deletion among the glycophorin variants.

At the beginning of the analysis, it was assumed that DEL7 deletion is the reciprocal copy of the DUP14 deletion copy, because they have almost identical estimated breakpoints. However, the alignment from the Sanger sequencing result suggests that this is not the case, as the difference between each breakpoint was 57.402 kb towards the 5' end and 47.723 kb towards the 3' end. However, DUP14 is larger than DEL7 at only 9.679 kb in size.

**Table 3.1: Summary of all breakpoints identified for each deletion and duplication variant.** N/A means that the breakpoint sequence does not map to any repeat element.

| Variant | Proximal breakpoint | Distal breakpoint | Repeat element | Variant size | Genes Involved |
|---|---|---|---|---|---|
| **DEL1** | Chr4:144835143-144835279 | Chr4:144945375-144945517 | L1PBa, L1ME3 & L1PBa (LINE) | 110.359 kb | *GYPB* |
| **DEL2** | Chr4:144912872-144913001 | Chr4:145016127-145016256 | THE1C (LTR) | 103.257 kb | *GYPB* |
| **DEL6** | Chr4:144780045-144780137 | Chr4:145004120-145004212 | L1PA5 (LINE) | 224.168 kb | *GYPE* and *GYPB* |
| **DEL7** | Chr4:144780450-144780498 | Chr4:144901287-144901335 | N/A | 120.839 kb | *GYPE* |
| **DUP14** | Chr4:144853613-144853688 | Chr4:144723019-144723094 | L1PA5 (LINE) | 130.518 kb | *GYPE* |
| **DUP29** | Chr4:144939393-144939453 | Chr4:144825583-144825643 | MIR & MIRb (SINE) | 113.809 kb | *GYPE* and *GYPB* |

**Table 3.2: Summary of the DUP5 complex variant and the gene conversion identified breakpoints.** N/A means that the breakpoint sequence does not map to any repeat element. NRC means to non-coding region.

| Variant | Breakpoints | Proximal breakpoints | Distal breakpoints | Repeat element | Variant size | Genes Involved |
|---|---|---|---|---|---|---|
| **DUP5** | A | Chr4:145113700 | Chr4:144936865 | MLT1J2 (LTR) | 299 bp | NDR (duplication) |
| | B | Chr4:144933933-144933945 | Chr4:144820933-144820945 | L1PA2 & L1MB5 (LINE) | 2.933 kb | *GYPB* (inverted) |
| | C | Chr4:144819856-144819887 | Chr4:144932860-144932891 | | 1.090 kb | *GYPE* (inverted) |
| | | | | N/A | 1.086 kb | *GYPB* (inverted) |
| ***GYPE-B-E*** | A | Chr4:144807057-144807062 | Chr4:144925341-144925346 | N/A | 2.455 kb | *GYPE* (deletion) |
| | B | Chr4:144927782-144927928 | Chr4:144809511-144809657 | N/A | 2.441 kb | *GYPB* (duplication) |

Although the DUP3 duplication breakpoints were not identified, it is quite possible that DUP3 is the reciprocal copy of the DEL6 deletion. This expectation is based on the comparison between their 5 kb window plots and fibre-fish results (Figure 3.55). The overlaps between these variants suggest that the *GYP* gene area is a highly variable region with many recombination events.

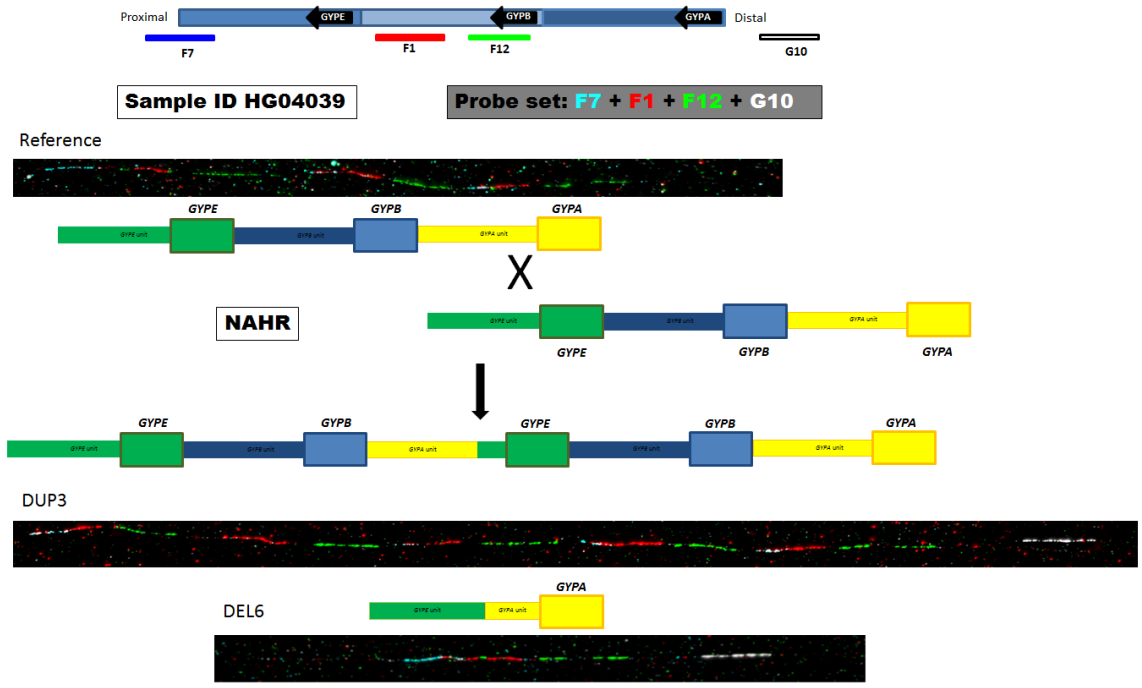



**Figure 3.55: A comparison of the DUP3 duplication and DEL6 deletion fibre-FISH results.** The digram shows the expected NAHR possible results and how they are matched with the DUP3 and DEL6 fibre-FISH results, which suggests that they are reciprocal copies.

The breakpoints for each variant can be used to design a specific PCR assay that could be applied to a large cohort in order to investigate the frequency of any of these variants and apply any further association analysis. The four remaining variants need to be sequenced and their breakpoints identified in order to cover all of the variants. DUP2, DUP3 and DUP7 primers were optimised and their products extracted and purified. They must be sequenced in order to determine their exact breakpoints. However, as the designed primer pairs for DUP24 cannot be optimised, a new primer pair needs to be designed specifically for this variant.

Leffler *et al*. (2017) study has Sanger sequenced positive samples for DUP4 and DEL1 and they have identified one of the DUP4 breakpoints and the DEL1 breakpoints, the sequencing result of DEL1 in this project confirms the reported breakpoints of this variant from the Leffler *et al*. study. Comparison between bother Sanger sequencing results are shown in (Appendix 12).

DUP29 shows a (*GYPB-E*) hybrid gene with duplication of a small exon 1 of the *GYPE* gene and all of the *GYPB* gene exons except exon 1, which encodes the leading peptide. Therefore, DUP29 will have a normal GPB protein and the duplicated GPB protein missing the leading peptide. However, exon 1 of the *GYPE* is duplicated, which could cover the missing expressed leading peptide. This hybrid gene has not been reported before to be one of the MNS antigens, but it might be a new (not reported) Miltenberger phenotype.

The *GYPE-B-E* gene conversion event was very interesting because it has been occurred within the *GYPE* and *GYPB* genes themselves. However, the gene conversion breakpoints indicate that no deletion of any exon of *GYPE* gene only a (2,450-2,601) bp deletion of the intron 1 and no duplication of any exon of the *GYPB* only a (2,437-2,588) bp duplication of the intron 2. In addition, the gene conversion breakpoints have not indicated for any a frameshift in the sequencing of any possible result of a stop codon, which means no exons were affected. Thus, there is not an expectation of any change in the gene expression and the generated protein.

Apart from the gene conversion, DUP29 and DUP4 (GP.Dantu) (see chapter 5), no fusion glycophorins are predicted to be generated. Compared to the blood group antigens many of which are fusion products, DNA analysis of the detected variants suggest that there is no variant related to the known blood group antigens. However, deletion of *GYPB* (exon 2 to 6) is responsible for the null phenotype RBCs (S-s-U-) (Blumenfeld and Huang, 1997; Reid, 2009). Therefore, it is expected that individuals homozygous for DEL1, 2 and 6 or carrying combination of these alleles have the S-s-U- phenotype as these variants have full deletion of *GYPB* gene. However, the differences in their breakpoints could lead to a different blood group antigen. For example, GP.Hil and GP.JL MNS antigens are the results of the *GYPA-B* hybrid gene but with different breakpoints that lead to different GPA-B proteins (Velliquette *et al*., 2008).

Moreover, $M^k/M^k$ phenotype is a result of absent of *GYPA* and *GYPB* which leads to a resistant to malaria infection (Blumenfeld and Huang, 1997). However, from the analysed variants in this project, no one of the deletion variants has covered either *GYPB* and *GYPA* genes together or *GYPA* alone. Therefore, for future work analysis of more variants with a deletion of the *GYPA* would be useful to expand the study of the glycophorins relationship with malaria infection and other blood antigens that could be expressed from the absence of the *GYPA* gene, such as En(a-) RBCs antigen.

To conclude this chapter, it would be interesting to sequence and analyse the deletions and duplications that were not analysed and studied in this project, such as DEL4 and DUP6 in order to study and analyse them to determine the exact breakpoints for use in any further analysis and try to cover most of the glycophorin variants. For future work, sequencing of the Tanzanian and Benin malaria cohort is recommended for investigate other unknown glycophorin CNVs.

# Chapter 4: Developing paralogue ratio tests for measuring CNV at the glycophorin gene cluster

## 4.1 Evidence for copy number variable glycophorin genes

The glycophorin proteins act as receptors for the *P. falciparum* malaria parasite and glycophorin genes are within a copy number variable region. This chapter describes development of a PRT approach to measure glycophorin copy number in order to genotype large cohorts. The PRT PCR was chosen to type the copy numbers of the glycophorin region, because is able to determine the multiallelic copy number variations (mCNVs) and is an accurate and cost effective technique uses small amount of DNA, which is useful when cohorts are limited (Aldhous *et al*., 2010). However, the reference regions should be not copy number variable in order to call the integer glycophorin copy numbers accurately. PRT can be very difficult in the initial stages of designing primers, although once they have been designed perfectly, PRT becomes a robust test for typing CNV.

Genome wide analysis reports that glycophorins (*GYPA*, *GYPB* and *GYPE*) are highly variable genes in terms of copy number (Conrad *et al*., 2010). There is evidence, in the copy number information given in the Database of Genomic Variants (DGV) (MacDonald *et al*., 2014), of deletions, duplication, and polymorphism, and the Agilent array-CGH data reveals three glycophorin genes are copy number variable (Figure 4.1). The Database of Genomic Variants is the most comprehensive repository of CNVs, and it details insertions, deletions, inversions and inversion breakpoints in the human genome. The data included within it was obtained from healthy individuals. DGV considers segments of DNA larger than 1000 bp, and insertions or deletions greater than 50 bp are also included (http://dgv.tcag.ca/gb2/gbrowse/dgv2_hg19/).

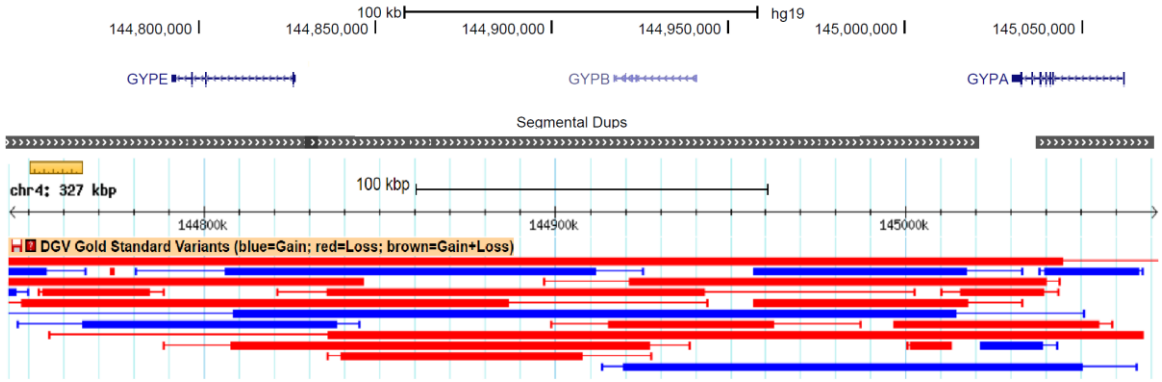



**Figure 4.1: Copy number variable regions in glycophorin genes.** Database of Genomic Variants shows a comprehensive summary of structural variation (blue indicates a gain, red indicates a loss and brown indicates both a gain and loss in size) in glycophorins. The copy number variable regions cover all the *GYPA*, *GYPB* and *GYPE*. As the figure shows, the *GYPE* is a less copy number variable gene of the glycophorin gene family. DGV tracks were used from http://dgv.tcag.ca/gb2/gbrowse/dgv2_hg19/.

In addition, normalised high throughput sequencing data examining 1000 Genomes Project samples has detected different deletions and duplications in the glycophorin genes, as shown in chapter 3, indicating that *GYPA*, *GYPB* and *GYPE* are copy number variable genes. As mentioned previously the Leffler *et al*. (2017) study identified various deletions and duplication within the glycophorin region and each variant was confirmed in at least 2 unrelated individuals, confirming some CNVs were previously detected by normalised high throughput data sequencing of the 1000 genome project samples (Chapter 3).

## 4.2 Strategy for PRT measurement of glycophorin copy number

There are two types of PRT assay: cis PRT assays, and trans PRT assays. The difference between cis PRT and trans PRT is the location of the reference amplicon; such that the reference amplicon of a cis PRT will be on the same chromosome as that the test amplicon, while the reference amplicon of a trans PRT will be on different chromosome to the test amplicon. For this project both PRT assay types were designed, three cis PRT assays and two trans PRT assays. They were chosen using the PRTPrimer software (Veal *et al*. 2013) after establishing certain criteria; i.e. the PCR products had to be between 90-450 bases in size, the primer sizes had to be between 18-27 bases in length and the size difference between the test and the reference was between 3-29 bases (Chapter 2).

For the cis PRT assays (cis_PRT1, cis_PRT2 and cis PRT4) the *GYPE* was used as a reference. That because from aCGH data and blood groups, it was initially thought that *GYPE* to be the least copy number variable of the glycophorin gene family, while *GYPB* and *GYPA*

are the most copy number variable and so were considered as test regions (Figure 4.1). However, in case of the trans PRT assays (trans_PRT2 and trans_PRT4) the reference amplicon of trans_PRT2 is on chromosome 3, whereas the reference amplicons of trans_PRT3 are on chromosome 3 and chromosome 9. However, the DGV has confirmed that the three reference amplicons for both the trans_PRTs are not commonly copy number variable (Figure 4.2), and the reason for designing the trans_PRTs was to achieve full copy number measurement for all the glycophorin genes as a set.

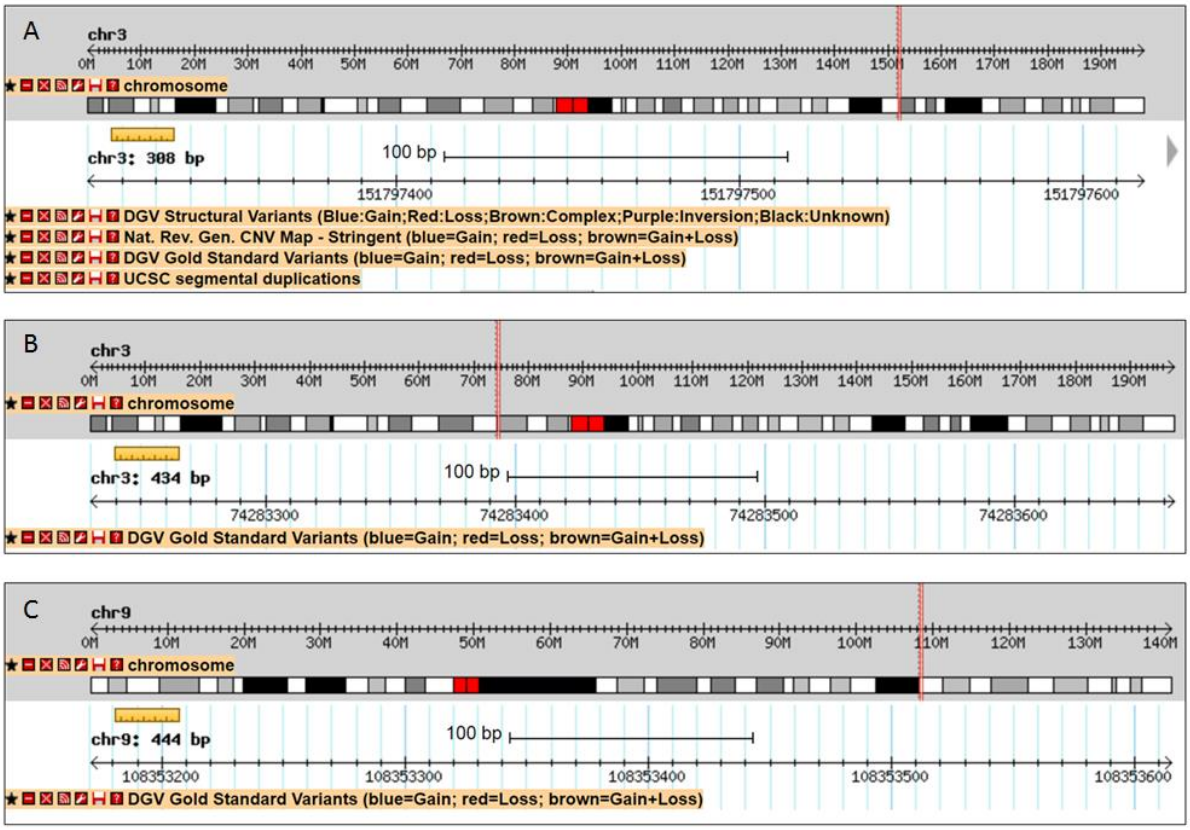



**Figure 4.2: Screen shots of the Database of Genomic Variants for the trans_PRTs reference regions.** DGV summarises structural variation in these regions. The top figure (A) represents the reference amplicon of trans_PRT2 on chr3. The middle figure (B) represents the first reference amplicon of trans_PRT3 on chr3. The bottom figure (C) represents the second reference amplicon of trans_PRT3 on chr9. DGV tracks were used from http://dgv.tcag.ca/gb2/gbrowse/dgv2_hg19/.

Each product from each cis_PRT amplifies a certain location in the glycophorin regions with different amplicon sizes (Figure 4.3) allowing test and reference amplicons to be distinguished by size. Each forward primer of these assays was fluorescently labelled from the 3' end of the primer sequence.

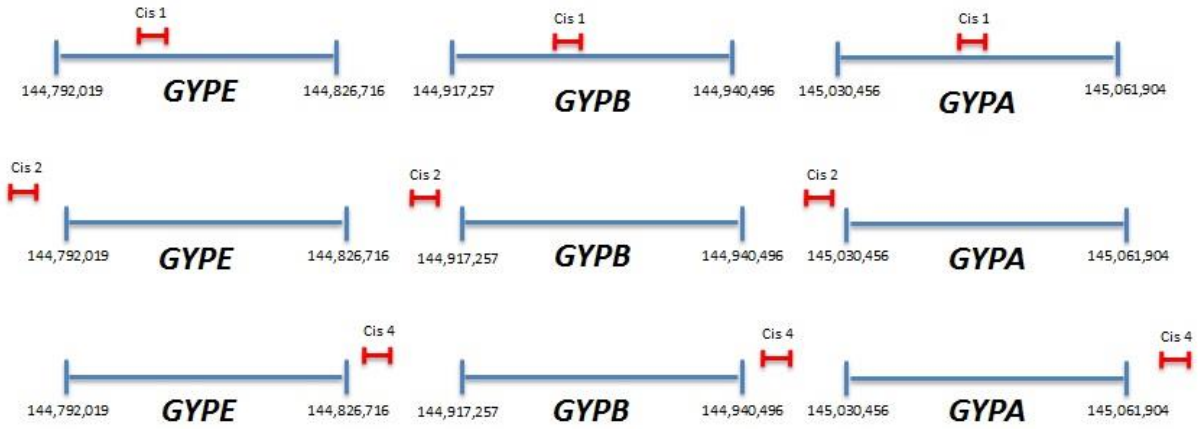



**Figure 4.3: Location of cis_PRT amplicons relative to glycophorin genes.** All cis_PRT primers used *GYPA* and *GYPB* for testing and *GYPE* as a reference. The reference sequence for the glycophorin genes was collected from the UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly.

The cis_PRT1 forward primer was labelled with a FAM dye and the assay yielded three PCR products, 399 bp, 396 bp and 418 bp from the *GYPA* (test), *GYPB* (test) and *GYPE* (reference) respectively. The cis_PRT2 forward primer was labelled with a FAM dye, and the assay yielded three PCR products 366 bp, 363 bp and 382 bp from the *GYPA* (test), *GYPB* (test) and *GYPE* (reference) respectively. The cis_PRT4 forward primer was labelled with a HEX dye and the assay yielded three PCR products 105 bp, 102 bp and 105 bp from the *GYPA* (test), *GYPB* (test) and *GYPE* (reference) respectively (Table 2.1, Chapter 2). All the test and reference PCR products were aligned to show the sequence basis of the size differences between amplicons (Figure 4.4).

Each product of each trans_PRT is amplified in a certain location within the glycophorin regions, but the test PCR products (from *GYPA, GYPB* and *GYPE*) gave same sizes when estimating the CNV on the whole glycophorin genes region using references on another chromosome with a different size (Figure 4.5) and each forward primer of the two assays was fluorescently labelled from the 3' end of the primer sequence.

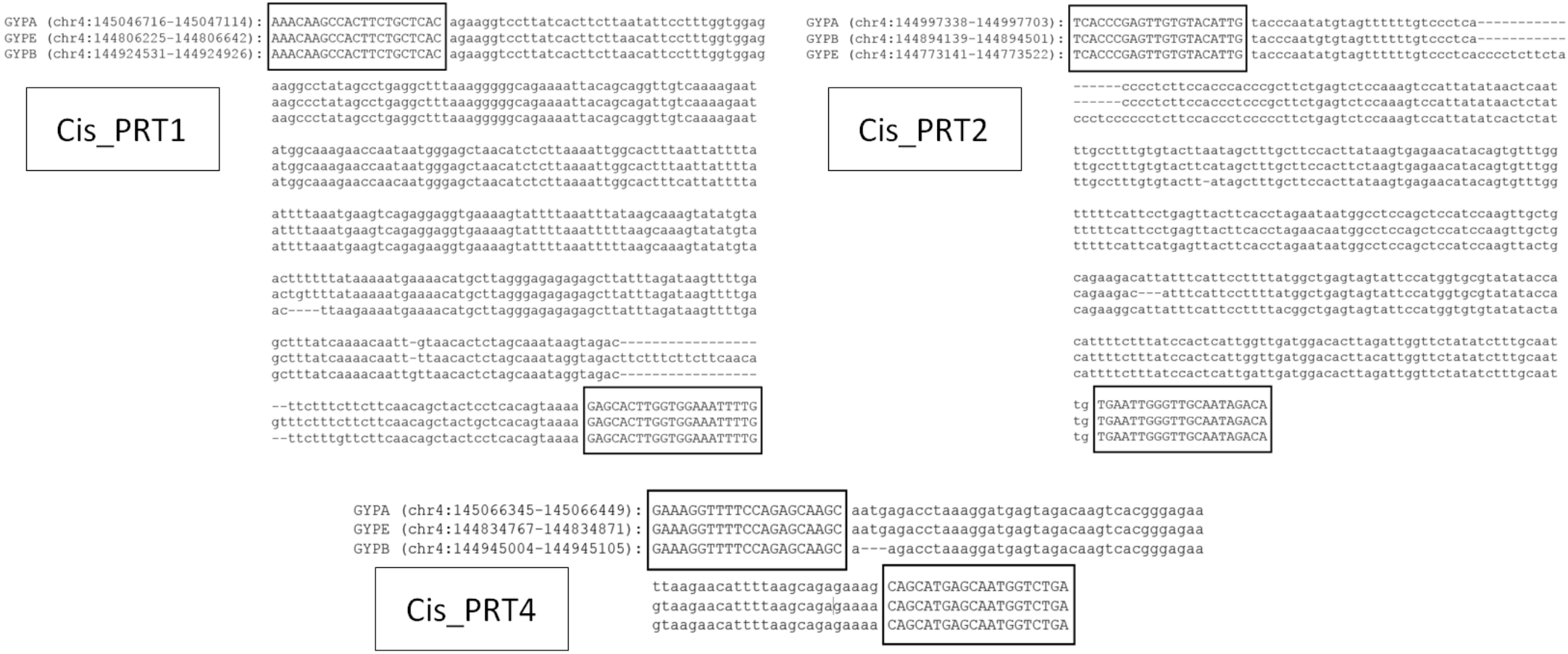



**Figure 4.4: Alignment of cis_PRT sequences from test and reference PCR products.** In each assay the top box represents the forward primer, while the bottom box represents the reverse primer. Insertions / deletions are indicated with dashes. Sequence coordinates are from the assembly (GRCh37/hg19).

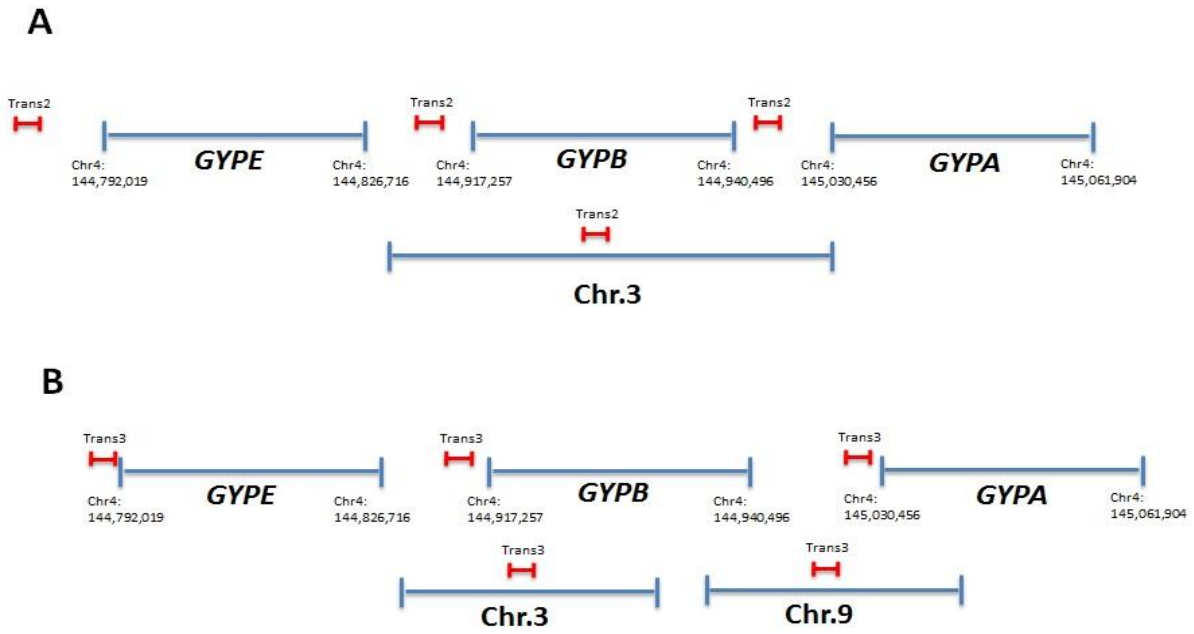



**Figure 4.5: Location of trans_PRT amplicons relative to glycophorin genes.** (A) trans_PRT2 primer pair uses *GYPA*, *GYPB*, and *GYPE* for tests and chr.3 as a reference, whereas (B) trans_PRT3 primer pair uses *GYPA*, *GYPB*, and *GYPE* for tests and chr.3 and chr.9 as references. The reference sequence for the glycophorin genes were collected from the UCSC Genome Browser on the Human Feb. 2009 (GRCh37/hg19) Assembly.

In contrast, the trans_PRT2 had two designed forward primers, one was labelled with a FAM dye and the other with a HEX dye, to attain two separate measurements of trans_PRT2 to produce more accurate CN calling for this assay. The trans_PRT2 assay yields three 314 bp PCR products from the *GYPA*, *GYPB* and *GYPE* genes (tests) and a 308 bp PCR product from chromosome 3 (reference). As with trans_PRT2, the trans_PRT3 includes two designed forward primers, one was labelled with a HEX dye and the other with a NED dye, to produce two separate measurements of trans_PRT3 to attain a more accurate CN calling for this assay. The trans_PRT3 assay produced three 415 bp PCR products from the *GYPA*, *GYPB* and *GYPE* genes (tests), a 434 bp PCR product from chromosome 3 (reference), and a 444 bp PCR product from chromosome 9 (reference) (Table 2.1, Chapter 2). The sequence of the trans PRT amplicons are shown in (Figure 4.6).

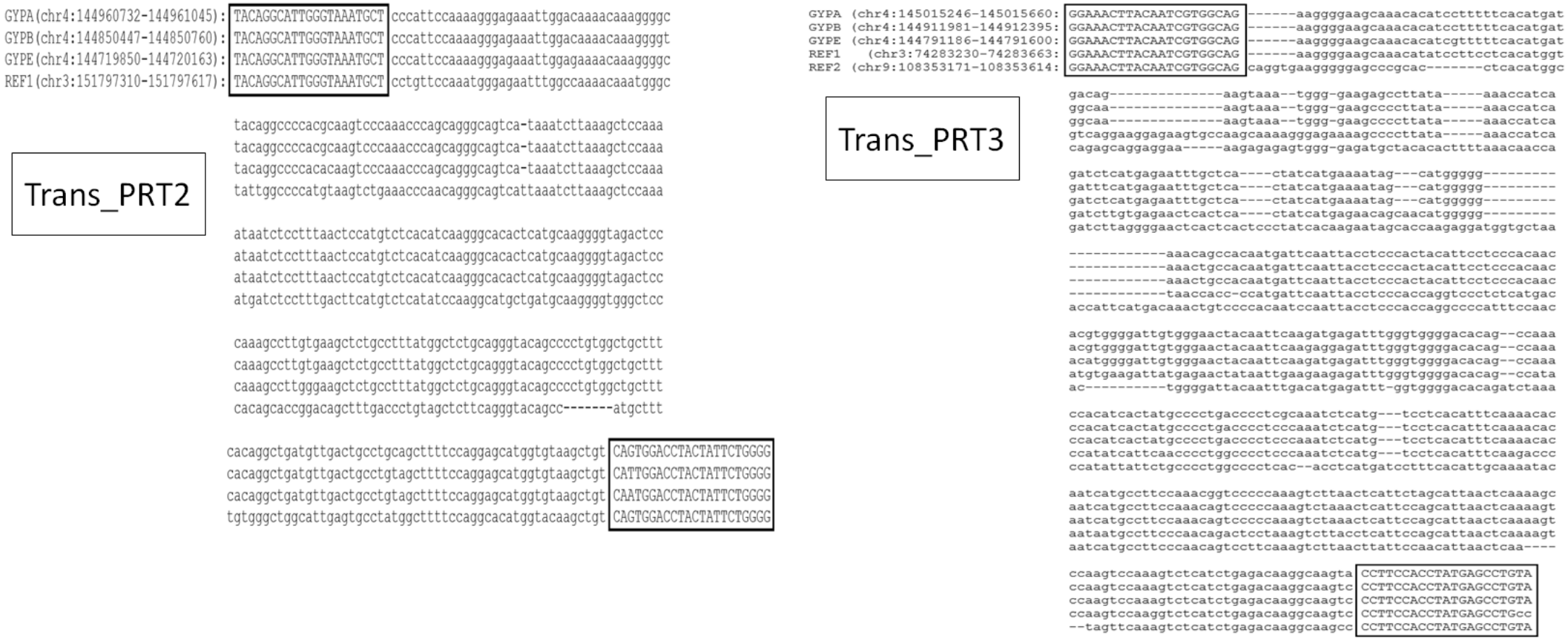



**Figure 4.6: Alignment of trans_PRT sequences from test and reference PCR products.** In each assay the top box represents the forward primer, while the bottom box represents the reverse primer. Insertions / deletions are indicated with dashes. Sequence coordinates are from the assembly (GRCh37/hg19).

The gradient PCR was applied with each cis_PRT primer pair to select the best annealing temperature using a control DNA sample (Co185) from the human random control samples plate (HRC). 65°C was identified as the best annealing temperature for all the cis_PRT assays (Figures 4.7).

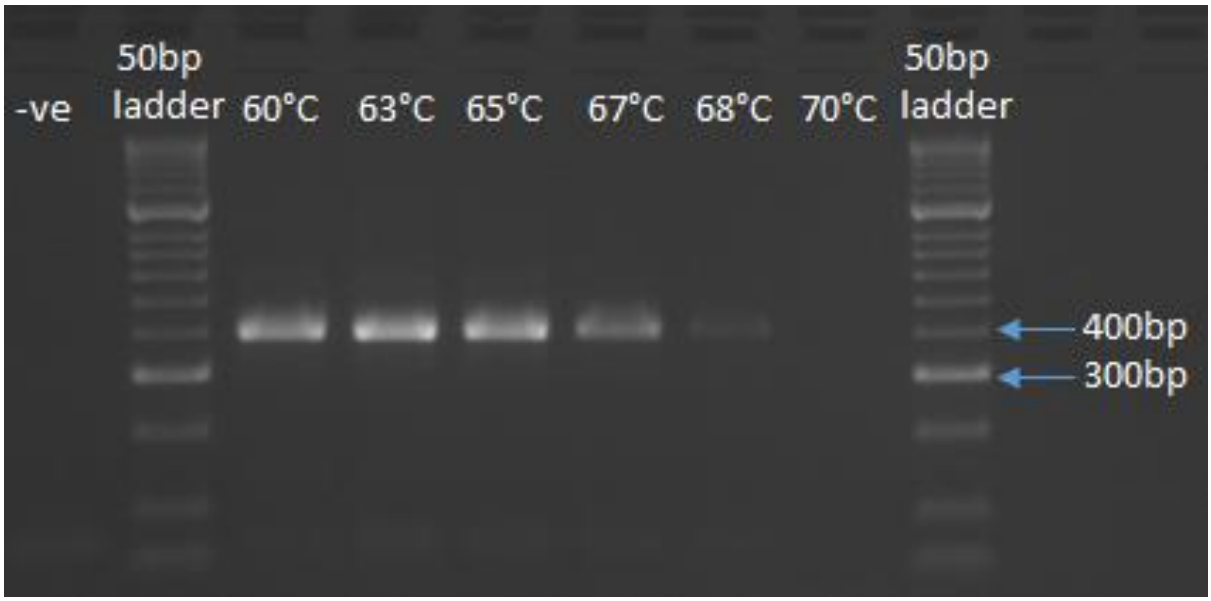



**Figure 4.7: An example of optimisation of one of the cis_PRT assays.** The gel result of the cis_PRT1 gradient PCR shows the predicted size bands at most annealing temperatures.

However, the gradient PCR was also applied for each trans_PRT primer pair, to select the best annealing temperature using a controlled DNA sample (Co185) from the human random control samples plate (HRC). 68°C was identified as the best annealing temperature for all the trans_PRT assays (Figures 4.8).

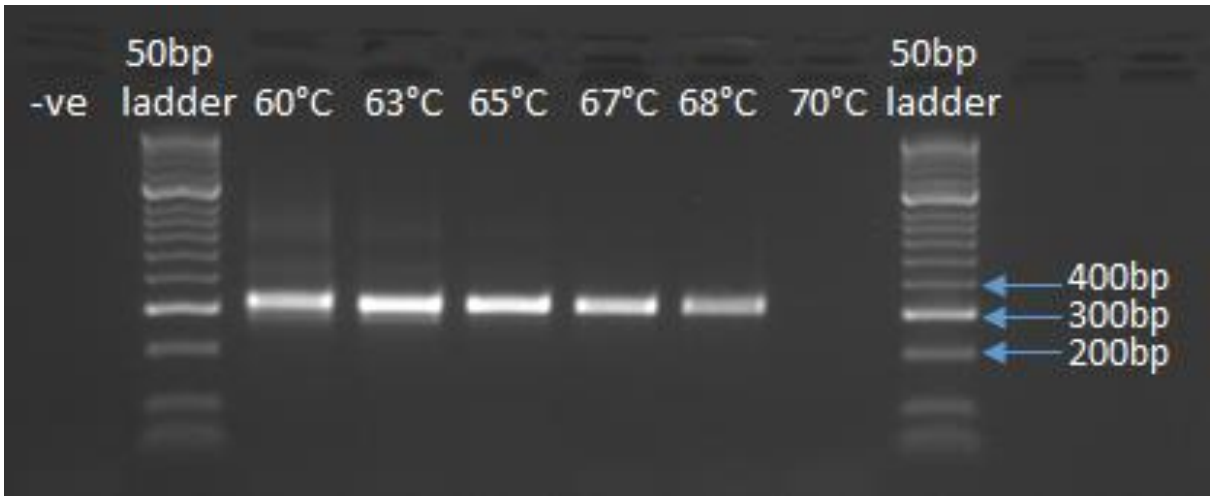



**Figure 4.8: An example of optimisation of one of the trans_PRT assays.** The gel result of the trans_PRT2 gradient PCR shows the predicted size bands at most of the annealing temperatures.

After designing the assays, seven separate measurements of CNVs within glycophorin genes were made; it was expected that each PRT assay would yield a similar copy number. After optimisation of each of the PRT assays, a triplex PCR combining cis_PRT1, cis_PRT2 and cis_PRT4 primers was designed (Chapter 2) because they share the same annealing temperature. A duplex PCR that combines trans_PRT2 (FAM) and trans_PRT3 (HEX) primers was also designed (Chapter 2) because it has the same annealing temperature and another duplex PCR that combines trans_PRT2 (HEX) and trans_PRT3 (NED) primers was also designed (Chapter 2). This strategy helps facilitate the use of the PCR products for all of the PRT assays in capillary electrophoresis, and saves time when typing copy numbers for a large cohort (Figure 4.9).

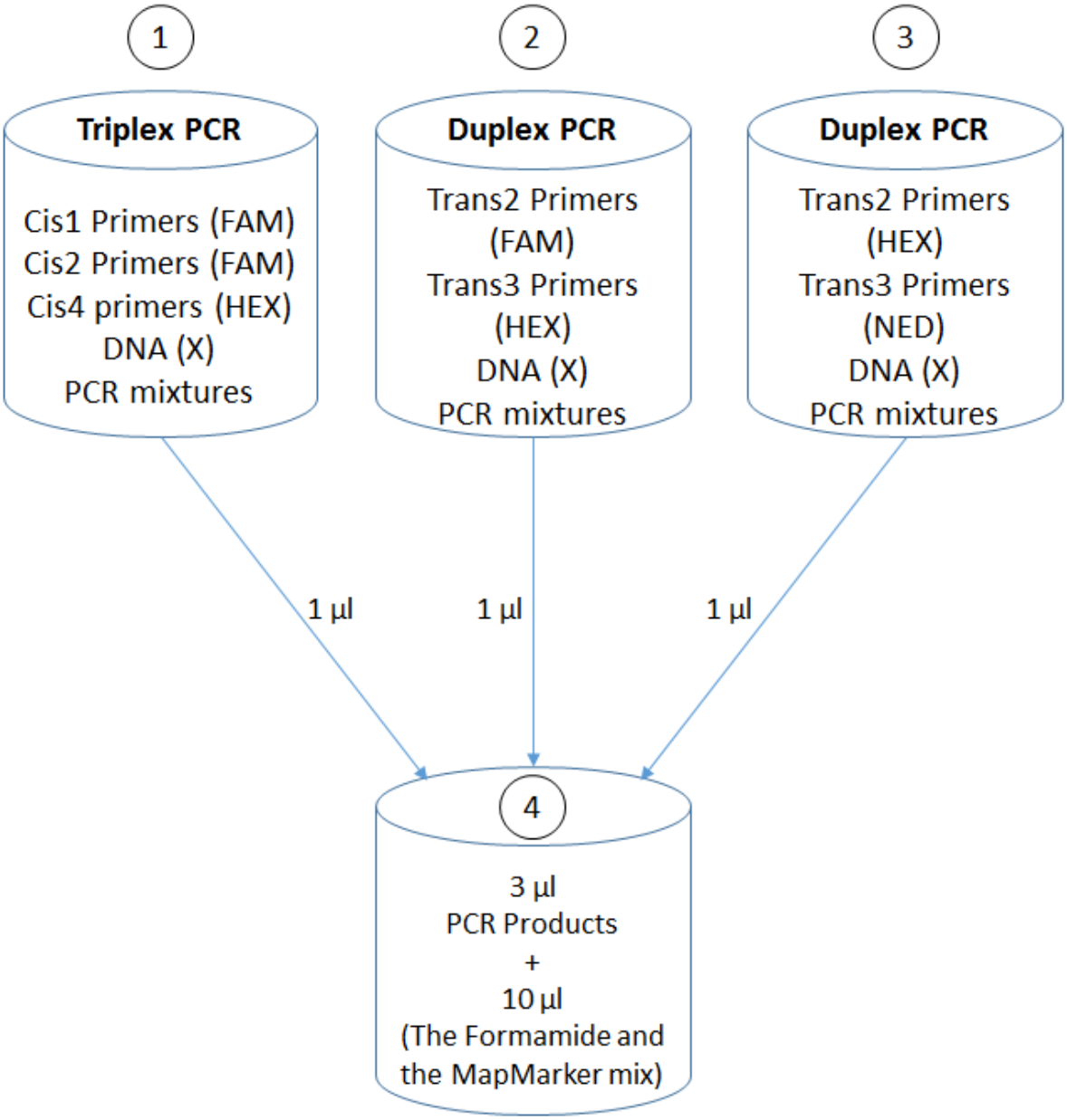



**Figure 4.9: Strategy for PRT measurement of glycophorin copy number.** Tube 1 represents the triplex PCR that combines all cis_PRT assays (three separate CN measurements) using a certain DNA. Tube 2 represents the duplex PCR that combines trans_PRT2 (FAM) and trans_PRT3 (HEX) assays (two separate CN measurements) using the same DNA in tube 1. Tube 3 represents the duplex PCR that combines trans_PRT2 (HEX) and trans_PRT3 (NED) assays (two separate CN measurements) using the same DNA in tubes 1 and 2. Tube 4 has 1 µl from tube 1 PCR product, tube 2 PCR product and tube 3 PCR product (3 µl in total) and 10 µl from the HIDI Formamide and the MapMarker® 1000XL Rhodamin (Rox1000XL) mix, this tube (4) will be used for the ABI (fragmental analysis). All details are given in Chapter 2.

PRT assays were used on samples Co185 to confirm correct detection of peaks from from PCR products. Using the GeneMapper® software (chapter 2), the electropherogram shows that cis_PRT1 (FAM), cis_PRT2 (FAM) and cis PRT4 (HEX) assays produced the expected peak sizes. Under each peak all the important information about the peak is given, including the peak area, which is necessary in any PRT assay to calculate and estimate the copy number of each sample. The blue peaks represent products with the FAM dye while the green peaks represent the products with the HEX dye and black NED dye (Figure 4.10 – 4.12). However, split peaks are shown from run to run as a technical issue, the GeneMapper® software cans calculate the peak areas and combine as a one peak.

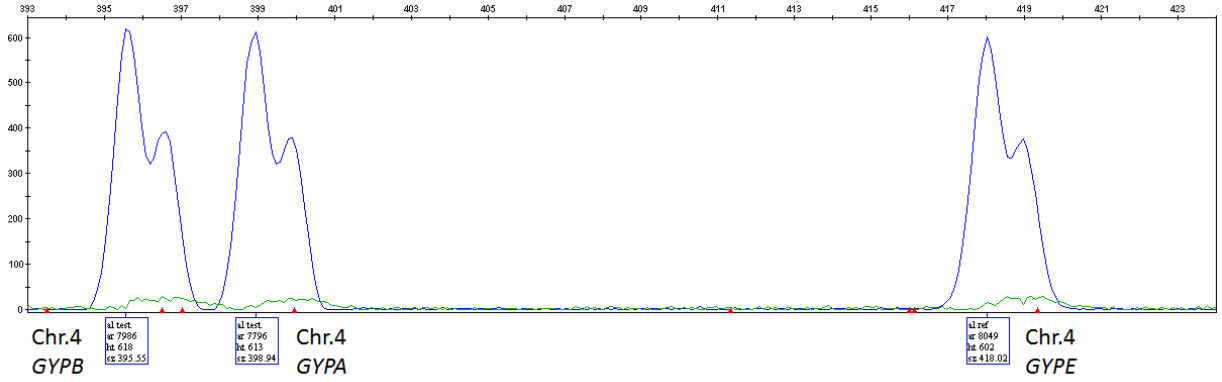



**Figure 4.10: Electropherogram of test loci and reference locus of cis_PRT1 assay.** The Y axis represents the peak height and the X axis represents the peak size by (bp). The box under each peak gives the peak name (test/reference), peak area, peak height and the peak size. The peaks are in blue because of the FAM dye.

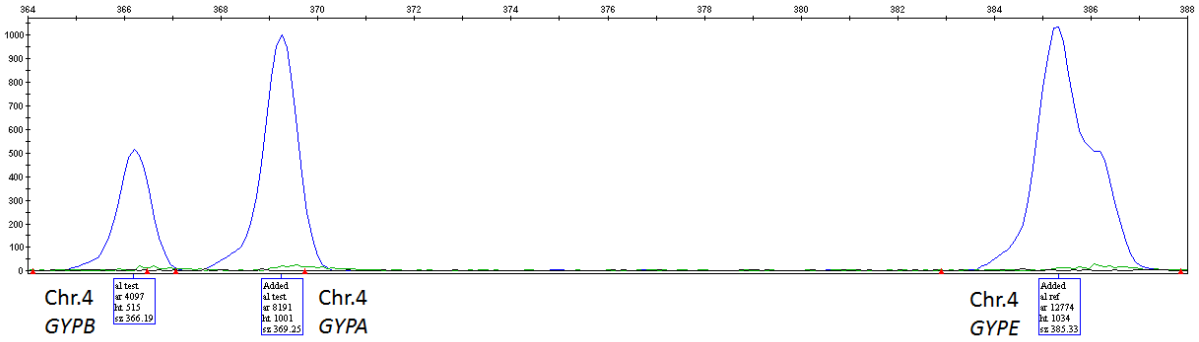



**Figure 4.11: Electropherogram of test loci and reference locus of cis_PRT2 assay.** The Y axis represents the peak height and the X axis represents the peak size by (bp). The box under each peak gives the peak name (test/reference), peak area, peak height and peak size. The peaks are in blue because of the FAM dye.

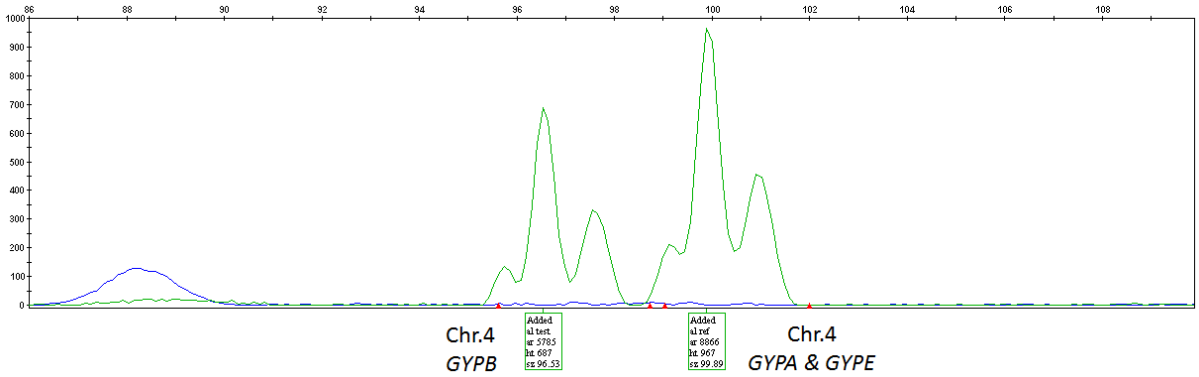



**Figure 4.12: Electropherogram of test loci and reference locus of cis_PRT4 assay.** The Y axis represents the peak height and the X axis represents the peak size by (bp). The box under each peak gives the peak name (test/reference), peak area, peak height and peak size. The peaks are in green colour because of the HEX dye. In this assay, the reference shows the same product as that in one of the tests that binds to *GYPA*, which explains why there is a single peak instead of two peaks.

In addition, the electropherogram shows the trans_PRT2 (FAM), trans_PRT2 (HEX), trans_PRT3 (HEX) and trans_PRT3 (NED) assays produce the same expected peak sizes. Under each peak all of the important information about the peak is given, including the peak area, which is essential in any PRT assay, as it makes it possible to calculate and estimate the copy number for each sample (Figure 4.13 – 4.14).

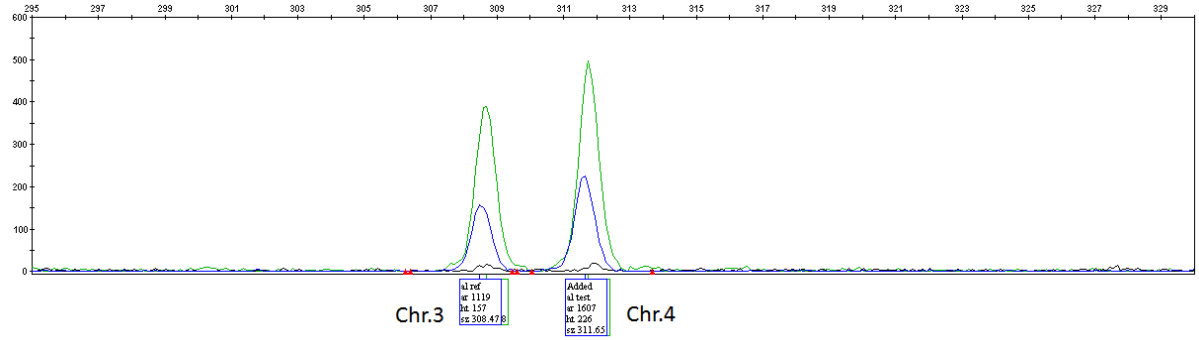



**Figure 4.13: Electropherogram of test loci and reference locus of trans_PRT2 (FAM) and trans_PRT2 (HEX) assays.** The Y axis represents the peak height and the X axis represents the peak size by (bp). The box under each peak gives the peak name (test/reference), peak area, peak height and peak size.

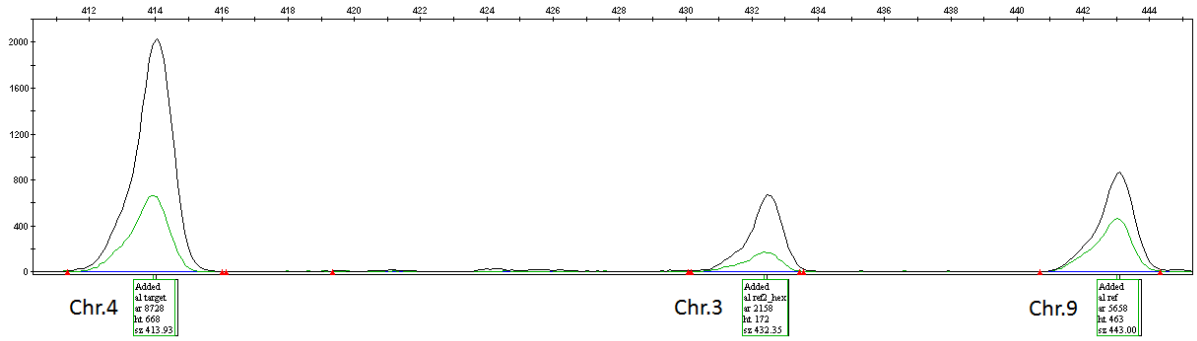



**Figure 4.14: Electropherogram of test loci and reference locus of trans_PRT3 (HEX) and trans_PRT2 (NED) assays.** The Y axis represents the peak height and the X axis represents the peak size by (bp). The box under each peak gives the peak name (test/reference), peak area, peak height and peak size.

Table 4.1 summarises the expected peak sizes of each PRT products.

**Table 4.1: Summary of all PRTs designed.**

| Product Size | PRT | Place | Dye |
|---|---|---|---|
| 102 bp | Cis_PRT 4 | Chr.4 *GYPB* (test) | HEX |
| 105 bp | Cis_PRT 4 | Chr.4 *GYPA* (test) | HEX |
| 105 bp | Cis_PRT 4 | Chr.4 *GYPE* (reference) | HEX |
| 308 bp | Trans_PRT 2 | Chr.3 (reference) | HEX+FAM |
| 314 bp | Trans_PRT 2 | Chr.4 *GYPA* (test) | HEX+FAM |
| 314 bp | Trans_PRT 2 | Chr.4 *GYPB* (test) | HEX+FAM |
| 314 bp | Trans_PRT 2 | Chr.4 *GYPE* (test) | HEX+FAM |
| 363 bp | Cis_PRT 2 | Chr.4 *GYPB* (reference) | FAM |
| 366 bp | Cis_PRT 2 | Chr.4 *GYPA* (test) | FAM |
| 382 bp | Cis_PRT 2 | Chr.4 *GYPE* (reference) | FAM |
| 396 bp | Cis_PRT 1 | Chr.4 *GYPB* (test) | FAM |
| 399 bp | Cis_PRT 1 | Chr.4 *GYPA* (reference) | FAM |
| 415 bp | Trans_PRT 3 | Chr.4 *GYPA* (test) | HEX+NED |
| 415 bp | Trans_PRT 3 | Chr.4 *GYPB* (test) | HEX+NED |
| 415 bp | Trans_PRT 3 | Chr.4 *GYPE* (test) | HEX+NED |
| 418 bp | Cis_PRT 1 | Chr.4 *GYPE* (reference) | FAM |
| 434 bp | Trans_PRT 3 | Chr.3 (reference) | HEX+NED |
| 444 bp | Trans_PRT 3 | Chr.9 (reference) | HEX+NED |

## 4.3 Normalisation of PCR results using positive controls

The PRT ratios from each PRT assay were normalised by using a linear regression of the 6 control DNA samples of known glycophorin copy number were chosen (Chapter 2.2 and Table 2.2) ratios against their known integer copy numbers (Figure 4.15). The regression equation was used to adjust all other PRT data for each of the seven PRT assays.

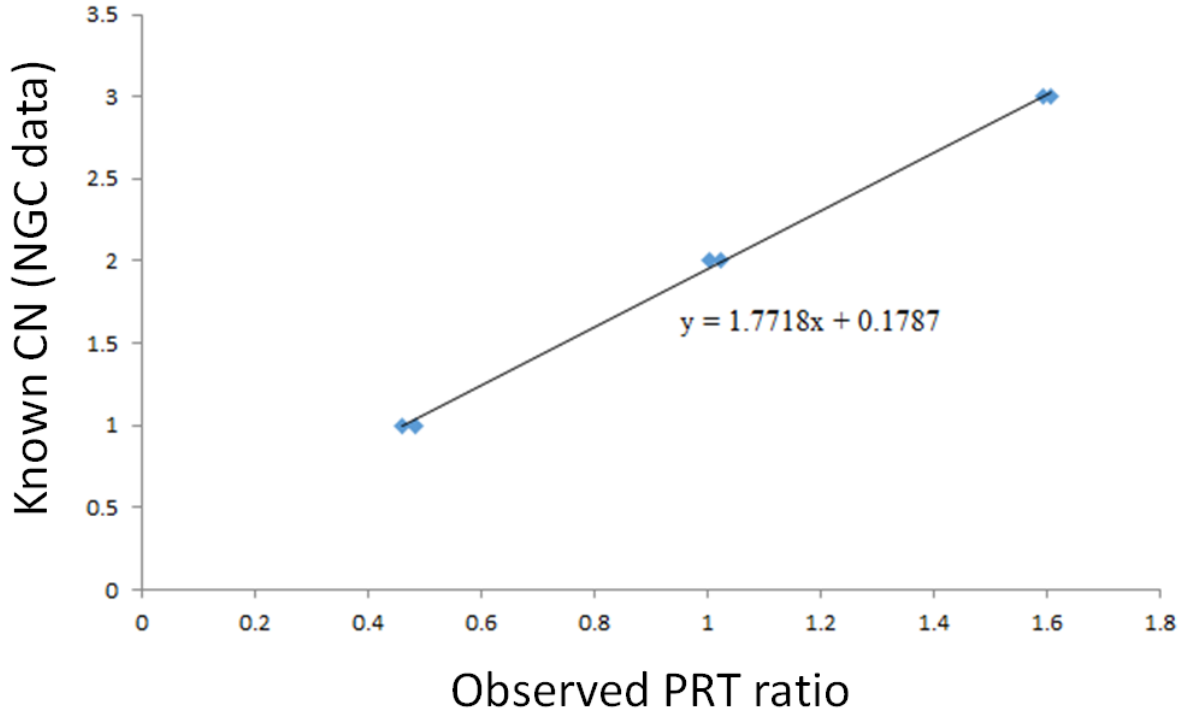



**Figure 4.15: A linear regression for the Cis_PRT1 assay with the 6 positive control DNA samples of known glycophorin copy number for a single PCR reaction.** The known CN derived from the NGS data appear on the Y axis and the observed PRT ratio (test/reference) appear on the X axis of controls; the normalized estimated CN of the positive DNA samples were plotted and the correlation between the known and unknown normalized CN ratios were calculated.

## 4.4 Validation of the designed PRT assays using HapMap samples

In order to evaluate the quality of the PRT assays and to explain how each assay works with large cohort, the PRT assays were compared with published high-throughput sequence read depth data to identify distinct clusters giving the copy number of the glycophorin genes. Further, it was important to show that the PRT assays yielded the same copy number results, so as to confirm that in future CNV studies the average copy number of glycophorin genes can be used.

177 DNA samples drawn from the 1000 Genomes project were typed by cis PRT assays, the copy number data from the three PRT assays was compared with estimates of copy number from the NGS read depth data taken from the 1000 Genomes Project. For each of the 177 samples, the total number of reads counted for the test region were (chr4: 144,745,739-145,069,133); and copy number variable (Figure 4.1) and a diploid reference region (chr4: 145,518,270-145,842,585), which is not copy number variable region (Figure 4.16) were calculated using Samtools. The same samples were typed for glycophorin copy number using the PRT strategy described in section 4.2.

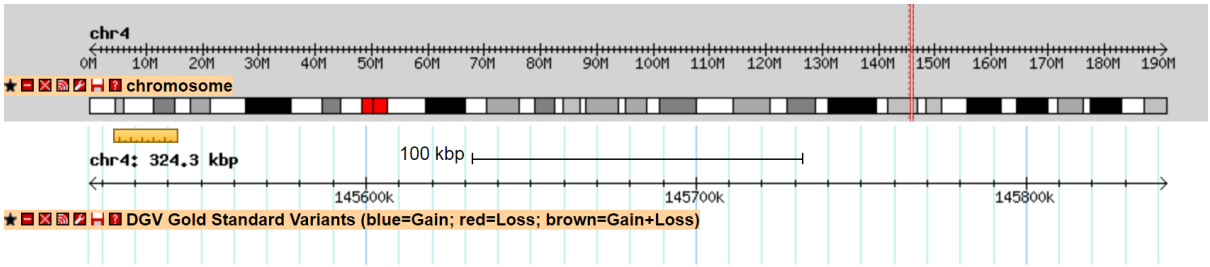



**Figure 4.16: A Screen shot of the Database of Genomic Variants for the diploid reference region (chr4:145,518,270-145,842,585).** DGV summarises the structural variation in this region. DGV tracks were used from http://dgv.tcag.ca/gb2/gbrowse/dgv2_hg19/.

### 4.4.1 Comparison of the PRT assay results with the published NGS data

#### 4.4.1.1 Cis_PRT1 validation

The resulting copy number estimates were the ratio of the total number of reads for the test region vs. the reference region. Therefore, a value of 1.0 was considered a normal diploid copy number (two copies), a value of 0.5 is a copy number of one (heterozygous deletion) and a value of 1.5 a copy number of three (heterozygous duplication).

The results of the comparison of the cis_PRT1 assay with the normalised high-throughput sequence read depth data shows the cis_PRT1 assay has a 100% concordance with the NGS data. Therefore, the cis_PRT1 assay works well and is an excellent choice for determining the glycophorin genes' copy number, because of the good clustering and because the deletions and duplications match the NGS data (Figure 4.17).

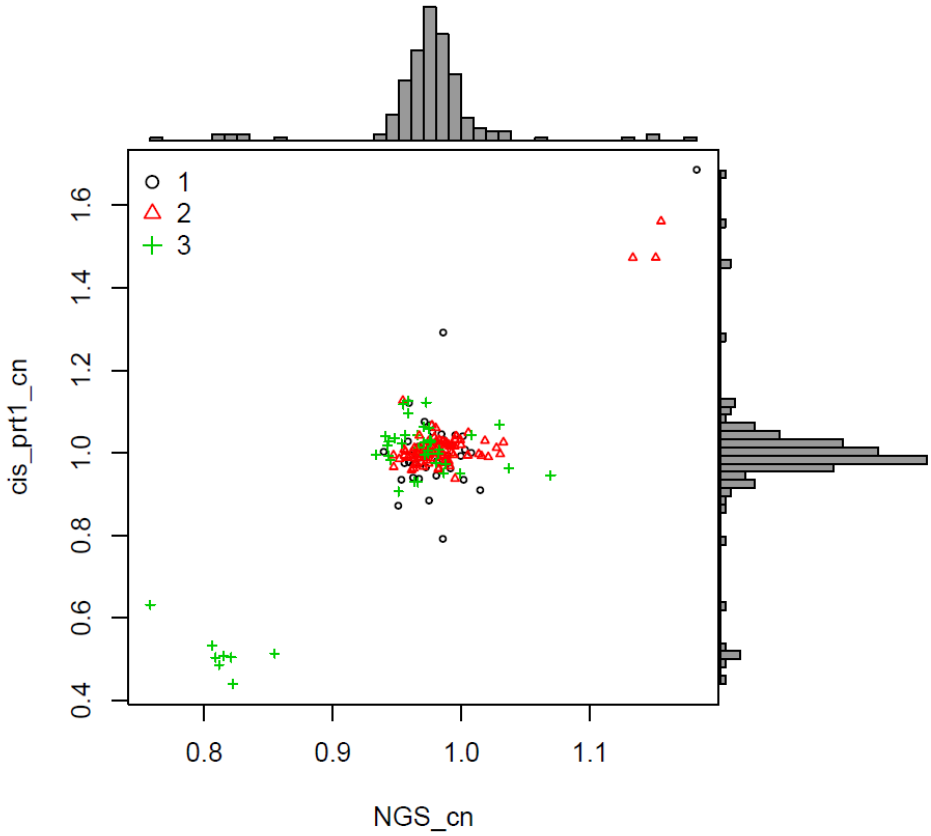



**Figure 4.17: Scatterplot for the cis_PRT1 assay of 177 of the 1000 Genome Project samples.** The X axis represents the (CN/Ref) ratios (known CNs) published in the NGS data, and the Y axis represents the predicted relative CNs from cis_PRT1 assay. The numbers 1, 2 and 3 indicate CEU, JPT/CHB and YRI populations respectively.

#### 4.4.1.2 Cis_PRT2 validation

In addition, although the overall result for the cis_PRT2 assay has a good clustering (Figure 4.18), there are some clusters which do not agree and might reflect different structural variants. Notwithstanding, it ascertained the copy number of most samples (170 out of 177) based on the NGS data. However, five of the tested samples in the assay (cis_PRT2 vs NGS) have high cis_PRT2 values, since they should be presented as deletion samples according to the NGS data, and the results from the other PRT assays (cis_PRT1 and cis_PRT4) correctly judged these samples (Table 4.2).

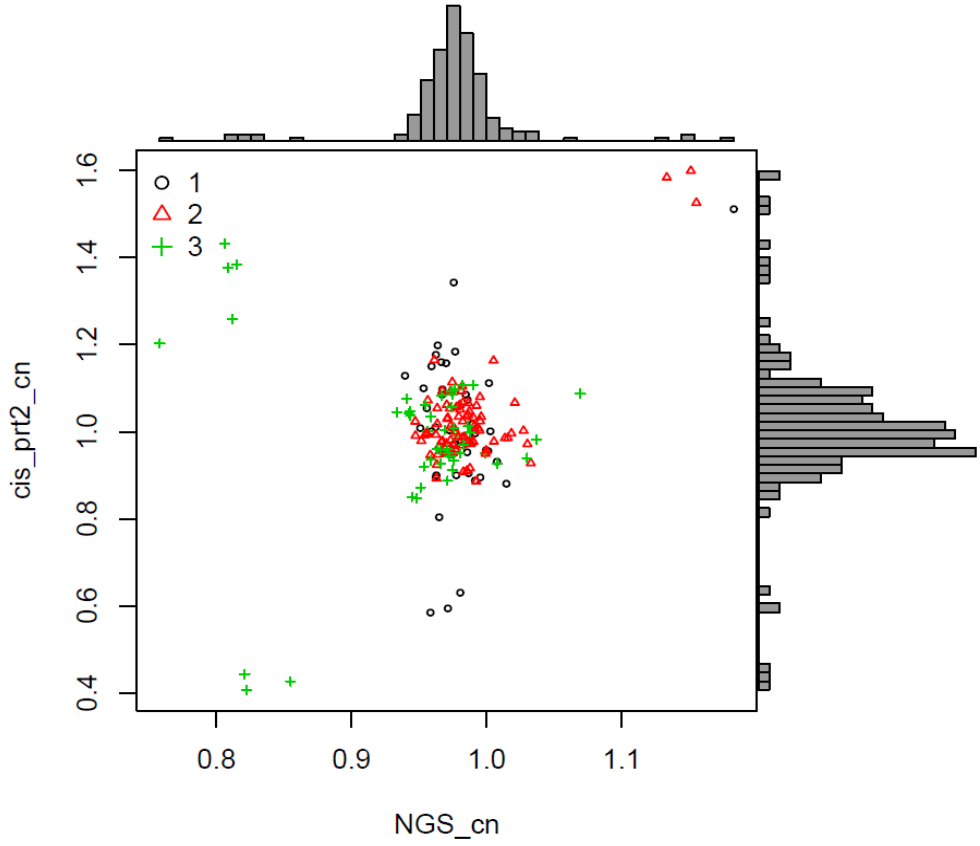



**Figure 4.18: Scatterplot for the cis_PRT2 assay of 177 of the 1000 Genome Project samples.** The X axis represents the (CN/Ref) ratios (known CNs) published in the NGS data, and the Y axis represents the predicted relative CNs from the cis_PRT2 assay. The numbers 1, 2, and 3 indicate CEU, JPT/CHB and YRI populations respectively.

**Table 4.2: The normalized copy number values of cis_PRT1, 2, 4 and NGS estimates for the five 1000 Genome Project samples (expected to be DEL1).**

| Sample | NGS_cn | Cis_PRT1_cn | Cis_PRT2_cn | Cis_PRT4_cn |
|---|---|---|---|---|
| **NA18523** | 0.81 | 0.51 | 1.38 | 0.68 |
| **NA19093** | 0.81 | 0.49 | 1.26 | 0.623 |
| **NA19102** | 0.80 | 0.50 | 1.38 | 0.65 |
| **NA19207** | 0.80 | 0.53 | 1.43 | 0.67 |
| **NA19223** | 0.75 | 0.63 | 1.20 | 0.73 |

In the same analysis (cis_PRT2 vs NGS) (Figure 4.18), three samples on the bottom middle of the scatterplot show low cis_PRT2 values. However, these should also be shown as normal samples, based on the NGS data and the results from the cis_PRT1 and cis_PRT4 assays (Table 4.3). According to the cis_PRT2 test and reference loci, it was anticipated that a gene conversion event would occur in *GYPE* with these three samples. This expected gene conversion event was tested and it was confirmed in these samples (see Chapter 2.5 and Chapter 3.6).

**Table 4.3: A table compares the normalized copy number values of cis_PRT1, 2, 4 and NGS estimates for the three gene conversion (*GYPE-B-E*) samples of the 1000 Genomes Project.**

| Sample | NGS_cn | Cis_PRT1_cn | Cis_PRT2_cn | Cis_PRT4_cn |
|---|---|---|---|---|
| **NA11994** | 0.98 | 0.94 | 0.63 | 1.00 |
| **NA12006** | 0.96 | 1.02 | 0.58 | 1.00 |
| **NA12716** | 0.97 | 1.07 | 0.54 | 0.97 |

#### 4.4.1.3 Cis_PRT4 validation

The data for Cis_PRT4 are shown in figure 4.19 are generally correct. Although the majority of the samples (171 out of 177) show high concordance with the NGS data, seven samples showed high values, since they were classed as normal samples based on the NGS data and the results from the other PRT assays (Table 4.4).

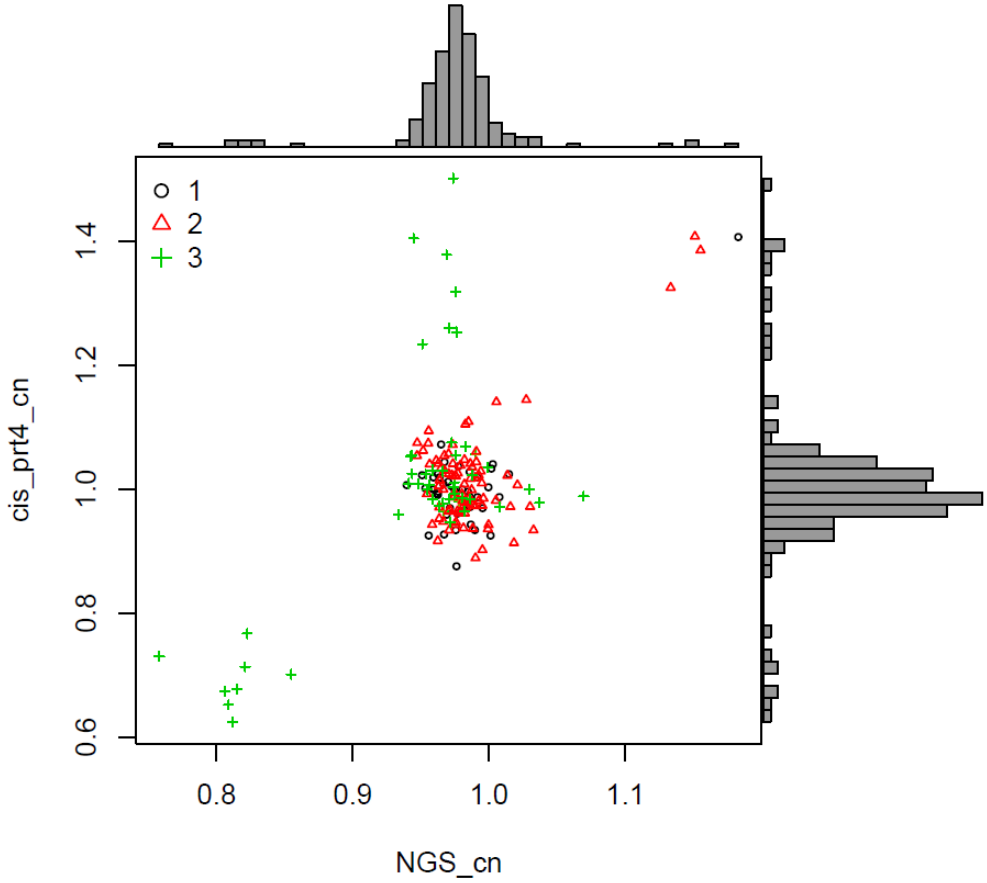



**Figure 4.19: Scatterplot for the cis_PRT4 assay of 177 of the 1000 Genome Project samples.** The X axis represents the (CN/Ref) ratios (known CNs) published in the NGS data, and the Y axis represents the predicted relative CNs from the cis_PRT4 assay. The numbers 1, 2 and 3 indicate CEU, JPT/CHB and YRI populations respectively.

**Table 4.4: A table compares the normalized copy number values of cis_PRT1, 2, 4 and NGS estimates for the seven 1000 Genome Project samples (expected *GYPA-B-A* gene conversion).**

| Sample | NGS_cn | Cis_PRT1_cn | Cis_PRT2_cn | Cis_PRT4_cn |
|---|---|---|---|---|
| **NA18502** | 0.97 | 1.06 | 0.89 | 1.26 |
| **NA18504** | 0.95 | 0.90 | 0.87 | 1.23 |
| **NA18516** | 0.96 | 1.00 | 1.00 | 1.38 |
| **NA18870** | 0.94 | 0.98 | 0.85 | 1.40 |
| **NA19119** | 0.97 | 1.03 | 0.93 | 1.31 |
| **NA19137** | 0.98 | 1.03 | 1.09 | 1.25 |
| **NA19138** | 0.97 | 1.00 | 1.05 | 1.50 |

The explanation of the high values of the seven samples is a small duplication of the glycophorin B region was occurred with normal value of the reference amplicon (glycophorin E region). Moreover, that could be due to a gene conversion event in *GYPB* according to the test and reference loci of the PRT assay, which could denote a *GYPA-B-A* gene conversion (Figure 4.20).

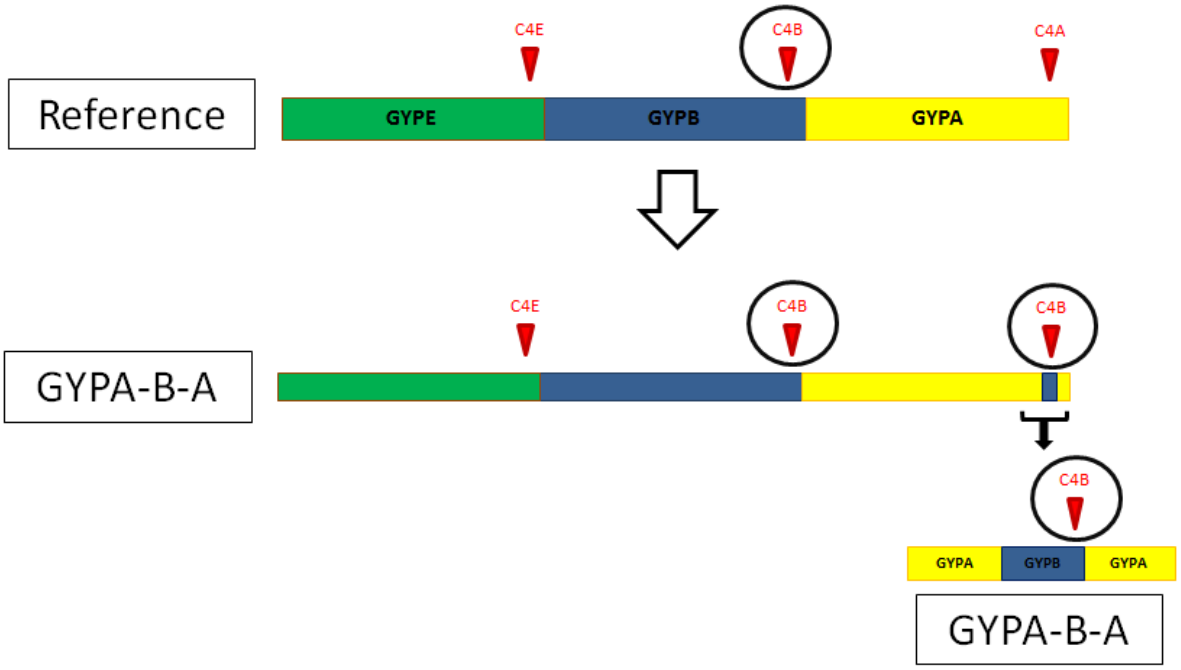



**Figure 4.20: An illustration of the expected gene conversion event possibility for the seven samples that given high values with the cis_PRT4 assay.** The diagram shows a reference haplotype with the cis_PRT4 primers amplicons. According to the position of the test (C4B) and the reference (C4E), these seven samples could be a carrier for *GYPA-B-A* gene conversion allele. The diagram shows that the possible allele has two amplicons of the test, while the reference amplicon hasn't deleted or duplicated.

#### 4.4.1.4 Trans_PRT2 validation

In contrast, although the trans PRT assays identified the majority of the duplications, they have not detected the deletions. In addition, these assays have yielded inconsistent clusters and low accuracy in glycophorin CN calling compared with all the cis PRT assays (Figures 4.21 and 4.22). Therefore, the trans PRT assays were removed from all further copy number analysis.

Trans_PRT2 detected all of the known duplicated samples and showed a fine clustering of the normal samples, only eleven samples were shifted up to have high values instead of normal values. However, this assay cannot detect any of the deletions, which are very important variant to investigate because most of the blood group antigens are resulting from full or partial deletion of glycophorin genes, also because deletion of these genes could affect the invasion of the RBCs by the *P. falciparum* malaria parasite (Figure 4.21).

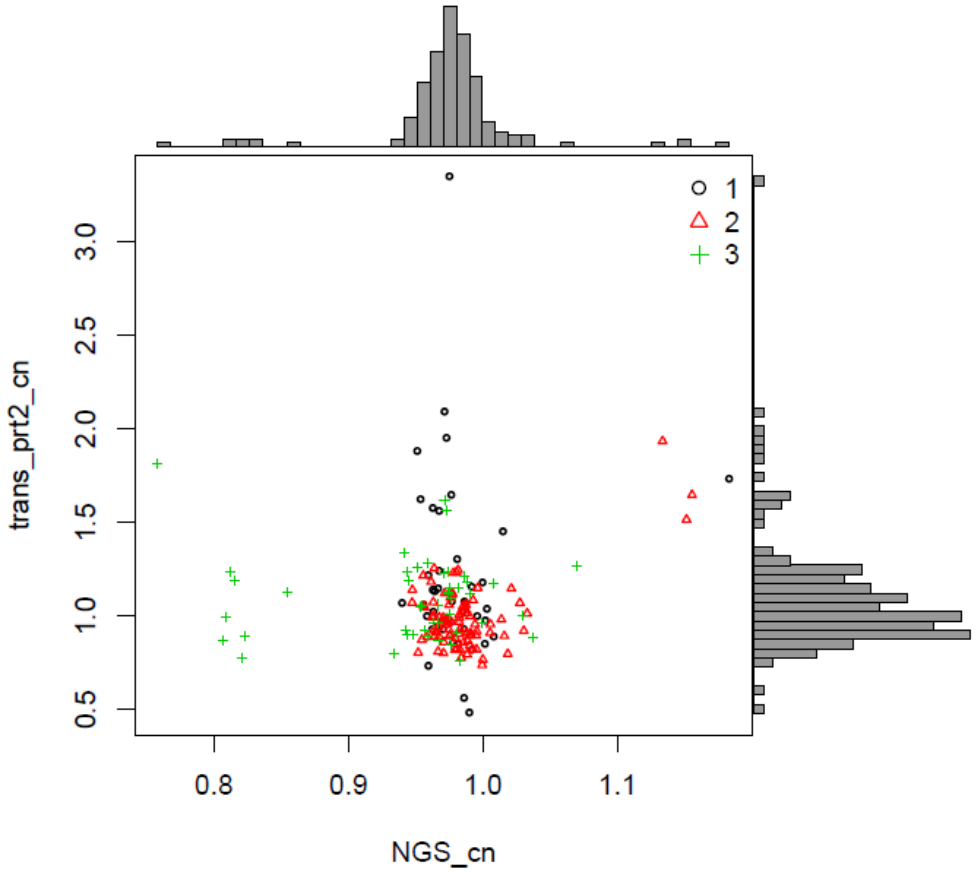



**Figure 4.21: Scatterplot for the trans_PRT2 assay of 177 of the 1000 Genome Project samples.** The X axis represents the (CN/Ref) ratios (known CNs) published in the NGS data, and the Y axis represents the predicted relative CNs produced from the trans_PRT2 assay. The numbers 1, 2, and 3 indicate CEU, JPT/CHB and YRI populations respectively.

#### 4.4.1.5 Trans_PRT3 validation

Trans_PRT3 has detected only three of the known duplicated samples and showed a very random and inconsistent clustering of the normal samples, with many samples showing high and low values. Although, this assay has showed an ability to detect some of the known deleted samples (5 out of 8) the normalized CN values of these samples are not low enough to confirm any unknown deletion (Figure 4.22). Therefore, the trans PRT assays have been discarded from all further copy number analysis.

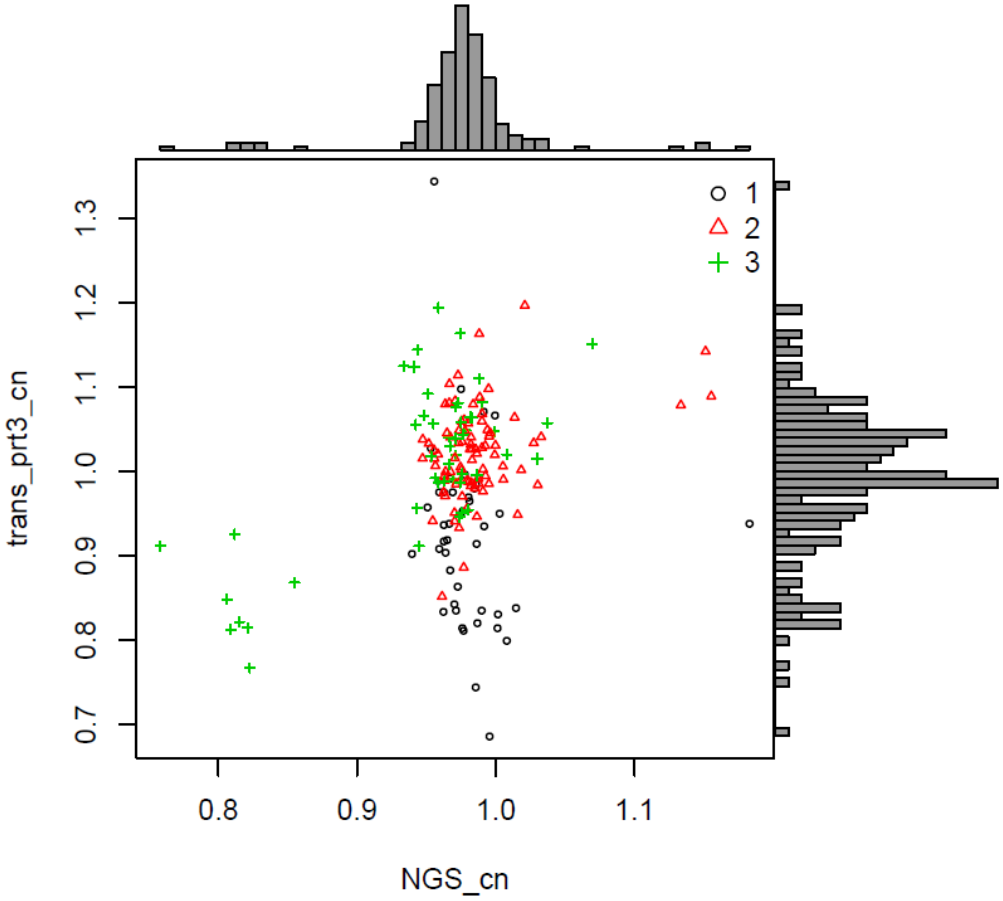



**Figure 4.22: Scatterplot for the trans_PRT3 assay of 177 of the 1000 Genome Project samples.** The X axis represents the (CN/Ref) ratios (known CNs) published in the NGS data, and the Y axis represents the predicted relative CNs from the trans_PRT3 assay. The numbers 1, 2, and 3 indicate CEU, JPT/CHB and YRI populations respectively.

### 4.4.2 Analysis of PRT results in the light of glycophorin variant breakpoints

The cis PRT PCR product positions (Table 4.5) and the information for each glycophorin variant breakpoints (Chapter 3, Tables 3.1 and 3.2) have been used to evaluate each of the PRT assays for each glycophorin variant. Because cis assays use amplicons within the glycophorin cluster as a reference, and because *GYPB* is used as the test, some variants will be missed or will generate unexpected results. Therefore, the PRT locations were compared with the known locations of variants (Figure 4.23).

**Table 4.5: Summary of cis PRT product positions.**

| Cis PRT assay | Place | Position |
|---|---|---|
| **Cis_PRT 4** | Chr.4 *GYPB* (test) | Chr4:144945004 - 144945105 |
| **Cis_PRT 4** | Chr.4 *GYPA* (test) | Chr4:145066345 - 145066449 |
| **Cis_PRT 4** | Chr.4 *GYPE* (reference) | Chr4:144834767 - 144834871 |
| **Cis_PRT 2** | Chr.4 *GYPB* (reference) | Chr4:144894139 - 144894501 |
| **Cis_PRT 2** | Chr.4 *GYPA* (test) | Chr4:144997338 - 144997703 |
| **Cis_PRT 2** | Chr.4 *GYPE* (reference) | Chr4:144773141 - 144773522 |
| **Cis_PRT 1** | Chr.4 *GYPB* (test) | Chr4:144924531 - 144924926 |
| **Cis_PRT 1** | Chr.4 *GYPA* (reference) | Chr4:145046716 - 145047114 |
| **Cis_PRT 1** | Chr.4 *GYPE* (reference) | Chr4:144806225 - 144806642 |

After calculating the expected cis_PRT1 result values for the test (C1B) and the references (C1A) for each variant using the 5 kb windows plot for each variant from their figures, exact breakpoints are given in chapter 3, cis_PRT1 assay (C1E reference) cans detect DEL1, DEL2, DEL6, DUP2, DUP3, DUP29 and DUP24. Nevertheless, normal values (~ 0.1) will be given for DEL7, DUP7, and DUP14, so these three variants cannot be detected using cis_PRT1 (Figure 4.23 and Figure 4.24).

For cis_PRT2, the result values for the test (C2A) and the references (C2B and C2E) for each variant were as predicted using 5 kb windows plots and exact breakpoints in chapter 3; therefore, this assay can detect DEL2 and DUP24 with a value ~ 0.5 and ~1.5 (Figure 4.23). However, it will give a normal value (~ 0.1) with DUP7 and DEL6. Cis_PRT2 gives high values with DEL1 and DEL7. For an example, the cis_PRT2 scatterplot (Figure 4.18) shows five of the tested samples in the assay have high values, since these samples should be presented as deletions according to the NGS data and the results from cis_PRT1 and cis_PRT4 results, which correctly identify these samples. The 5 kb window plot of deletion type 1 (chapter 3) explains the situation of the five samples (high values) in terms of cis_PRT2 assay (Table 4.2), which shows one of the cis_PRT2 references (C2B) is deleted (Figure 4.23).

Thus, the end values will be high, because in this assay, the test amplicon is C2A, and the equation is (C2A/ (C2B+C2E)), which means these samples are highly likely to have the type 1 deletion allele as the cis_PRT1, cis_PRT4, and the NGS data confirms that. Although the same value (high value) will be given with DEL7, the value will be normal with cis_PRT1 and high with cis_PRT4 (Figure 4.23). In addition, the cis_PRT2 gives low values with DUP3, DUP14 and DUP29, this is because one of the cis_PRT2 references (C2B or C2E) falls within the duplication region; therefore, the end values will be low according to the equation for this assay (C2A/ (C2B+C2E)), which means that samples with the DUP3, DUP14 or DUP29 allele are highly likely to have low values in the cis_PRT2 assay (Figure 4.24).

The third cis PRT (cis_PRT4) shows the result value from the test (C4B), and the reference (C4E) for each variant was anticipated using the 5 kb windows plots and their exact breakpoints in chapter 3, therefore, the cis_PRT4 assay can detect DEL1 and DEL2. Nevertheless, it will give a normal value with DEL6, DUP3, DUP29 and DUP24, so this variant cannot be detected with this assay (Figure 4.23 and Figure 4.24). Cis_PRT4 affords a high value with DEL7 because the reference (C4E) is in the deletion region and the test (C4B) is outside the deletion region. The result is that the end values will be high, according to the equation for this assay (C4B/C2E), which means that samples with DEL7 allele are likely to have high values with the cis_PRT4 assay (Figure 4.23). In addition, cis_PRT4 gives low values with DUP7 and DUP14, because the reference (C4E) is in the duplication region and the test (C4B) applies outside the duplication region; therefore, the end values will be low according to the equation for this assay (C4B/C2E), which means that samples with the DUP7 or DUP14 allele are highly likely to have low values with the cis_PRT4 assay (Figure 4.24). Table 4.6 illustrates the possible cis PRT values, normal, low and high values.

**Table 4.6: Copy number calling values of Cis PRT assays for each *GYP* variant.**

| Variant | Cis_PRT1 value | Cis_PRT2 value | Cis_PRT4 value |
|---|---|---|---|
| **DEL1** | Low (0.0 -0.5) | High (1.5 - 2.0) | Low (0.0 -0.5) |
| **DEL2** | Low (0.0 -0.5) | High (1.5 - 2.0) | Low (0.0 -0.5) |
| **DEL6** | Low (0.0 -0.5) | Normal (1.0) | Normal (1.0) |
| **DEL7** | Normal (1.0) | High (1.5 - 2.0) | High (1.5 - 2.0) |
| **DUP2** | High (1.5 - 2.0) | High (1.5 - 2.0) | High (1.5 - 2.0) |
| **DUP3** | High (1.5 - 2.0) | Normal (1.0) | Normal (1.0) |
| **DUP7** | Normal (1.0) | Low (0.0 -0.5) | Low (0.0 -0.5) |
| **DUP14** | Normal (1.0) | Low (0.0 -0.5) | Low (0.0 -0.5) |
| **DUP29** | High (1.5 - 2.0) | Low (0.0 -0.5) | Low (0.0 -0.5) |
| **DUP24** | High (1.5 - 2.0) | Normal (1.0) | Normal (1.0) |

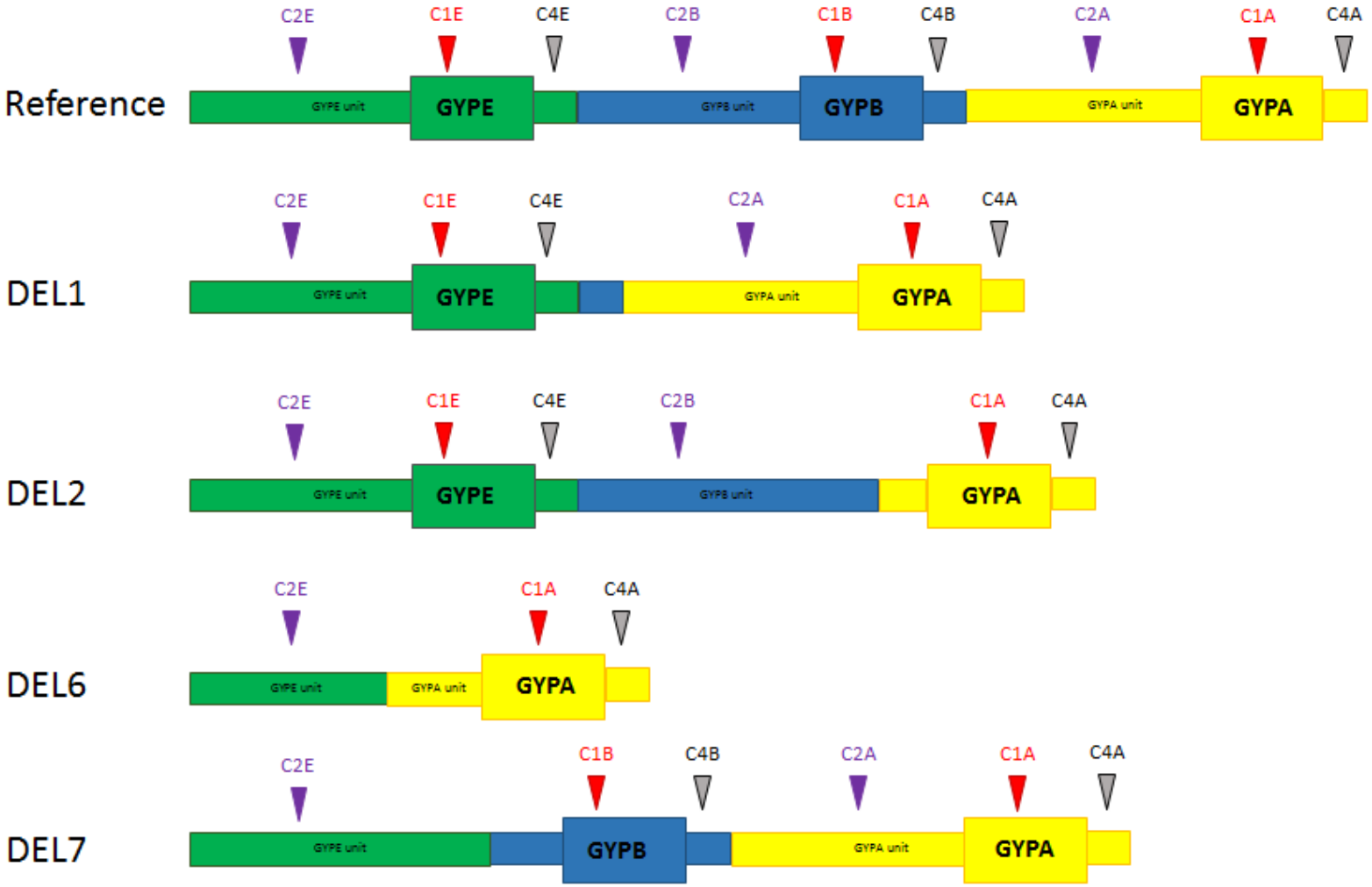



**Figure 4.23: A diagram shows reference and test amplicon positions of the PRT assays on the different deletion alleles.** Red, purple and black-grey triangles represent the cis_PRT1, 2 and 4 amplicons respectively.


**Figure 4.24: A diagram shows reference and test amplicon positions of the PRT assays on the different duplication alleles.** Red, purple and black-grey triangles represent the cis_PRT1, 2 and 4 amplicons respectively.

## 4.5 Discussion

In conclusion, the robust PRT assays were designed and applied to measure copy number variations across the entire human glycophorin gene cluster. The effectiveness of these assays was demonstrated by the common CNV, as shown in the NGS data set detected by the PRT assays. Based on these results, the most common glycophorin variants in the HapMap samples were DEL1 and DEL2, which have been found in the Yoruba population from Ibadan in Nigeria (YRI) (African ancestry) and DUP2, which have been found in the Han Chinese population from Beijing (CHB), and northern and western European populations from Utah (CEU) (Table 4.7).

**Table 4.7: Deletion and duplication frequencies in 1000 Genome Project samples.**

| Variant | Total samples | CEU samples | JPT samples | CHB samples | YRI samples |
|---|---|---|---|---|---|
| **Normal** | 165 (0.93) | 42 (0.98) | 45 (1.00) | 37(0.925) | 41 (0.84) |
| **DEL1** | 5 (0.029) | 0 (0.00) | 0 (0.00) | 0 (0.00) | 5 (0.1) |
| **DEL2** | 3 (0.017) | 0 (0.00) | 0 (0.00) | 0 (0.00) | 3 (0.06) |
| **DUP2** | 4 (0.02) | 1 (0.03) | 0 (0.00) | 3 (0.075) | 0 (0.00) |
| **Total** | 177 | 43 | 45 | 40 | 49 |

Cis_PRT1 and cis_PRT4 can detect the most common glycophorin variants, such as DEL1, DEL2 and DUP2. The Cis_PRT2 assay could be used for purposes such as confirming the existence of deletion type 2, differentiating between deletion types 1 and 2 and detecting the gene conversion event that occurred in the *GYPE* (*E-B-E* gene conversion), which was detected and its breakpoints identified (Chapter 3).

In addition, the cis_PRT4 assay could be used to detect gene conversion event (*GYPA-B-A*). However, this gene conversion is likely to be not within the actual *GYPA* and *GYPB* genes because both of the cis_PRT4 (C4B and C4A) are located in the *GYPB* and *GYPA* repeat units but not in the actual genes, therefore no such blood group is possible to be related to this gene conversion.

Some variants are missed with the PRT assays, such as DEL6, DEL7, DUP3, DUP7 and DUP14. It is possible to detect most of them by using *GYPA* as references instead of *GYPE*, and that because most of the detected variants were covered *GYPE* and *GYPB* rather than *GYPB* and *GYPA* as was predicted before. Thus, PRT assay analysis can be modified to detect additional known variants.

Using all the cis PRT results to analyse the glycophorin copy number is recommended to differentiate between glycophorin variants, between deletion types, and between duplication types. In addition, more accurate CN calling will be obtained if future work allows the design of more PRT assays with suitable test and reference loci on the glycophorin genes. It would be better in the future to focus on the glycophorin genes during designing the PRT assays, and according to chapter 3, figure 4.23 and 4.24 and the latest study by Leffler *et al*., (2017), test need to be within *GYPB* and *GYPE* genes, whereas the reference would be better to be on *GYPA* gene.

# Chapter 5: Association of CNV within glycophorins with malaria

## 5.1 Introduction

Many studies have discussed the relationship between variation in the glycophorin genes, which encode receptors for *P.falciparum* (Chapter 1) and the susceptibility to malaria (Band *et al*., 2015; Leffler *et al*., 2017; Ndila *et al*., 2018). This chapter studies the relationship between the glycophorin CNVs and *P.falciparum* malaria infection in two different cohorts from Benin in West Africa (infant malaria cohort) and Tanzania in East Africa (family-based, total population malaria cohort) using the PRT strategy described in (Chapter 4).

According to the glycophorin variant breakpoint results (chapter 3), 3 out of the 4 glycophorin deletion variants have entire absence of the *GYPB* gene and according to the Leffler *et al*. (2017) study, most of the common glycophorin deletions are DEL1 and DEL2. However, 5 out of the 6 glycophorin duplication variants have a duplicated *GYPE* gene. Therefore, this chapter uses the PRT assays for typing the *GYPB* copy number of the Tori-Bossito Benin malaria cohort and Tanzanian malaria cohort. *GYPB* gene encodes one of the most important proteins that responsible for the *P.falciparum* malaria parasite invasion, thus any mutation in the *GYPB* gene could affect the proteins biological function, which could lead to a full/partial prevention of the malaria parasite invasion (Ndila *et al*., 2018).

Following further analysis of the glycophorins region, a new SNP (rs186873296) was identified in one of the genome wide association studies (Band *et al*., 2015) to be associated with resistance to severe malaria in a region of ancient balancing selection, that has been argued to be due to its linkage disequilibrium with DUP4 (Leffler *et al*., 2017). This SNP is close to the glycophorin gene family and is located downstream of *GYPE* and upstream of the *FRAS1* related extracellular matrix 3 (*FREM3*) on chromosome 4.

The recent study about the duplication copy number variant (DUP4) has illustrated the variant molecular structure and identified breakpoints (Leffler *et al*., 2017). They have found a nearby association of DUP4 with resistance to severe malaria. This complex variant has six copy number changes that cannot have arisen by a single unequal crossover, DUP4 consists a duplication of *GYPE*, deletion of the 3’ end of *GYPB*, duplication of the 5’ end of *GYPB* and triplication of the 3’ end of *GYPA*. The DUP4 encoded protein would join the extracellular domain of *GYPB* to the transmembrane and intracellular domains of *GYPA*, creating an amino acid sequence at their junction, which is specific for the Dantu antigen in the MNS

blood group system (Blumenfeld *et al*., 1987). Leffler *et al*. (2017) has confirmed that DUP4 is like the common Dantu type (NE), both have two such hybrid genes (*GYPB-A*) and lack of a full *GYPB* gene.

This chapter aims to test association between the SNP (rs186873296) and *GYPB* copy number in the Tori-Bossito Benin malaria cohort in order to investigate if there is any association of *GYPB* copy number with rs186873296 SNP. That is because all deletions in this cohort have the absence of the *GYPB* gene. In addition, this chapter aims to design and optimise a specific DUP4 assay to be applied on the Tanzanian malaria cohort in order to genotype the DUP4 variant and calculate the allele frequency in this cohort because DUP4 previously found by Leffler *et al*. (2017). The genotyping data and the Tanzanian cohort clinical information have been utilized in this chapter to investigate if there is any association between the DUP4 genotype and any of the malaria phenotypes in the cohort, such as clinical episodes, parasitemia levels and haemoglobin levels.

## 5.2 Description of the malaria cohorts used

The two cohorts used in this study were obtained from two longitudinal multidisciplinary projects. The Tori-Bossito cohort is financed by France's Agence Nationale de la Recherche (ANR) (http://www.agence-nationale-recherche.fr/en/). This project is managed in the Atlantique Department of southern Benin by David Courtin and André Garcia. The malaria data analysed in this project were from the PALNOUGENENV cohort in Tori-Bossito (following 600 children from birth to 18 months). The Tori-Bossito Cohort is an infant malaria cohort with all clinical and non-clinical information for each infant. All these characteristics are shown and summarised in the table below (Table 5.1).

The Nyamisati Tanzanian cohort was drawn from a family-based study of 922 Tanzanian Bantu people in Nyamisati. The epidemiological and clinical phenotypes (parasite load, mean number of clinical episodes of malaria and haemoglobin levels) were generated from non-selected 'total' population surveys carried out annually during March and April between 1993 and 1999. A complete annual record of episodes of clinical malaria in Nyamisati between 1993 and 1999 was documented by Dr Ingegerd Rooth, who lived in the village since 1985 (Carpenter *et al*., 2012). See Chapter 2.1 for more details about the cohorts and the locations of the malaria samples. The cohort contains 1048 males and 1014 females with full clinical characteristics for each individual. A summary table of the clinical characteristics of the Tanzanian malaria cohort is shown below (Table 5.2).

**Table 5.1: Characteristics of Tori-Bossito cohort.**

| Characteristic | Code | Cases number |
|---|---|---|
| **Mother age** | Median (Range) by years | 28 (16-49) |
| **Sex** | Female = 0 | 272 |
| | Male = 1 | 283 |
| **Ethnic group** | Tori = 0 | 397 |
| | Fon = 1 | 56 |
| | Other = 2 | 92 |
| **Sickle cell** | Yes = 1 | 9 |
| | No = 2 | 532 |
| | **N/A** | 14 |
| **Chloroquine intake during pregnancy** | No = 0 | 101 |
| | Yes =1 | 441 |
| | Unknown = 2 | 13 |
| **Mosquito net** | No = 0 | 182 |
| | Yes = 1 | 364 |
| | **N/A** | 9 |
| **Malaria suspicion during pregnancy** | No = 0 | 419 |
| | Yes = 1 | 124 |
| | **N/A** | 0 |

**Table 5.2: Characteristics of The Nyamisati Tanzanian cohort.**

| Characteristic | Median (Range) | Cohort average |
|---|---|---|
| **Age** | 16 (0-89) | 20.7 |
| **EPI** (Clinical episodes) | -1.0728 | -0.99 |
| **PARA** (Parasitemia levels) | -1.1179 | -0.98 |
| **HB** (Uncorrected Haemoglobin levels) | 1.2071 | 0.17 |
| **HBYBY** (Hb levels corrected for age and sex) | 1.3319 | 0.20 |
| **HBPYBY** (Hb levels corrected for age, sex and parasitemia levels) | 1.1241 | 0.30 |

## 5.3 Results of studies on the Tori-Bossito Cohort

### 5.3.1 Glycophorin copy number typing on the Tori-Bossito Cohort

The Tori-Bossito cohort was successfully typed for *GYPB* copy numbers using cis PRT 1, 2 and 4. Three samples from this cohort were excluded from the study because of some uncertainties about the basic information (such as the sample name) recorded for them, and no DNA was present in another six sample tubes. Therefore, the total sample number typed for CNV with PRT assays was 574 samples.

The histogram analysis of raw PRT ratios shows clustering of values and indicated a total of three clusters for cis_PRT1 and cis_PRT4 (Figures 5.1-5.3). The first cluster indicated a

diploid copy number of zero and the remaining clusters were assigned diploid copy numbers of 1 and 2 copies respectively. However, the histogram analysis of raw PRT ratios reveals a clustering of values and indicated a total of six clusters for cis_PRT2. The first cluster indicated a diploid copy number of zero, and the remaining clusters were assigned diploid copy numbers of 1, 2 and 3 or more copies (4 or 5 copy numbers are rare in all populations) interpreted as homozygous deletions, heterozygous deletions, normal and heterozygous duplications respectively Figures 5.1-5.3).

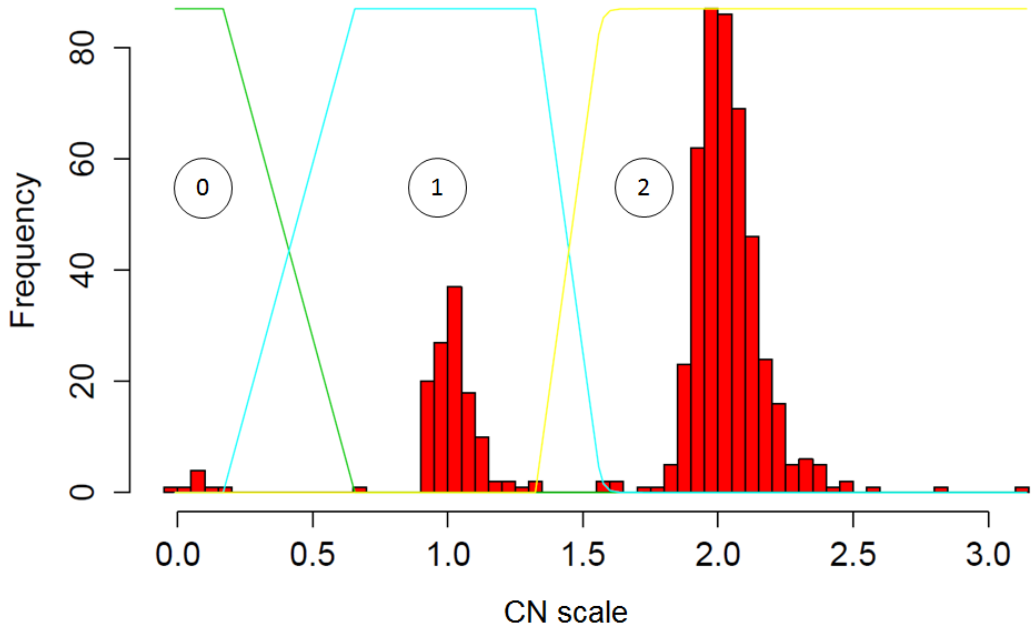



**Figure 5.1: Population distribution of copy number in Tori Bossito cohort for cis_PRT1.** The coloured lines show the Gaussian distributions for each copy number class (copy number = 0, 1 or 2). The x-axis indicates diploid copy number data after division by the standard deviation of the entire cis_PRT1 dataset. The y-axis represents the frequency scale.

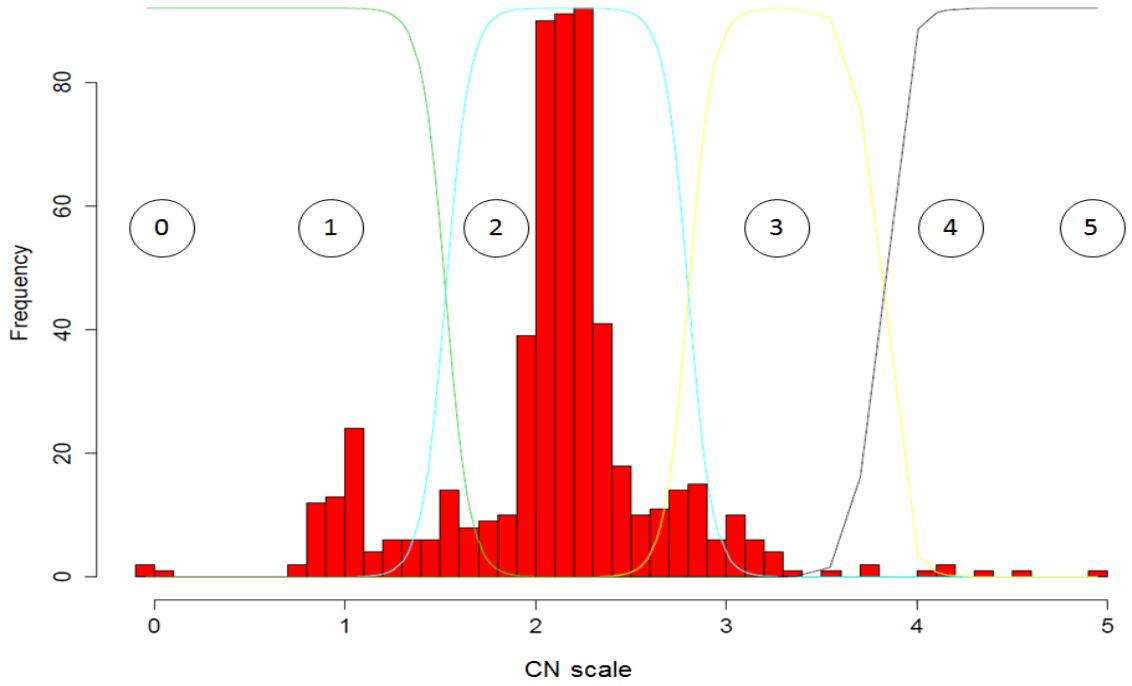



**Figure 5.2: Population distribution of copy number in Tori Bossito cohort for cis_PRT2.** The coloured lines show the Gaussian distributions for each copy number class (copy number = 0, 1, 2 3, 4 or 5). The x-axis indicates diploid copy number data after division by the standard deviation of the entire cis_PRT2 dataset. The y-axis represents the frequency scale.

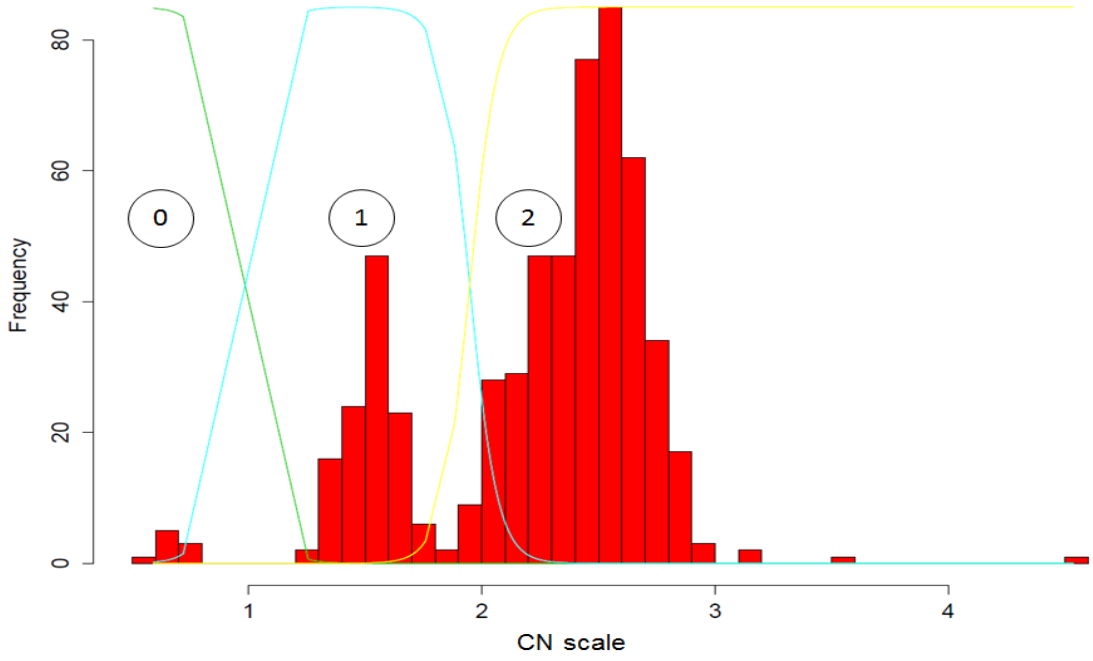



**Figure 5.3: Population distribution of copy number in Tori Bossito cohort for cis_PRT4.** The coloured lines show the Gaussian distributions for each copy number class (copy number = 0, 1 or 2). The x-axis indicates diploid copy number data after division by the standard deviation of the entire cis_PRT4 dataset. The y-axis represents the frequency scale.

These differences between the cis PRT assays are due to the differences between the tests and the reference locations and the glycophorin CNVs (Chapter 4, Figures 4.23 and 4.24). As a result, all cis PRTs have to be run together and compared using scatterplots to clearly determine whether the glycophorin variant is deletion or duplication.

The copy numbers of the 574 Benin samples were typed as for the 1000 Genome Project samples (Chapter 4.4). A value of a sample or a cluster of < 1 could indicate a deletion and a value of a sample or a cluster of > 1 could indicate a duplication. After the creation of the population distribution of the glycophorin copy number in the Tori-Bossito cohort for each cis PRT assay, the data for each cis PRT assay were compared by creating a scatterplot.

The scatterplot of Cis_PRT2 against Cis_PRT1 of the 574 Benin malaria cohort samples shows seven clusters in general (Figure 5.4), each cluster represents a specific genotype (Figure 5.5). The scatterplot shows that 446 samples cluster together under the value of 1 (the black cluster), which indicates that these samples have *GYPB* copy number of two. However, the two samples at the top of the black cluster show values of 2.0 and 1.0 with cis_PRT2 and cis_PRT1 respectively (DEL7), two samples at the right of the black cluster show values of 1.0 and 1.5 with cis_PRT2 and cis_PRT1 respectively (DUP3 or DUP24), and one sample at the bottom of the black cluster shows values of <0.5 and 1.0 with cis_PRT2 and cis_PRT1 respectively (DUP7 or DUP14). The prediction of each variant was relying on the chapter 3 and 4 results and the glycophorin PRT assays guideline table that has been generated in order to show the possible values of the cis PRTs for each glycophorin variant and it would facilitate the first prediction of the possible variant (Chapter4, Table 4.6).

Additionally, the scatterplot shows that 70 samples cluster together under the value of 0.5 for cis_PRT1 and under the value of (1.5 – 2) for cis_PRT2 (the top green cluster), which means that these samples have a heterozygous deletion (one copy). However, the high value from cis_PRT2 is the same scenario during validation of the cis PRT assays with the known copy numbers from the published data (Chapter 4). Therefore, these 70 samples are highly likely to be heterozygous for the DEL2 allele. In addition, the scatterplot shows that 50 samples cluster together under the value of 0.5 for both cis PRT assays (the bottom green cluster), which means that these samples also have a heterozygous deletion (one copy). However, these 50 samples are highly likely to be heterozygous for DEL2 allele (Figure 5.5).

Nevertheless, the last eight samples shown in the cis_PR1 vs cis_PRT2 scatterplot are divided into three red clusters. All clusters are under the value of zero for cis PRT1, whereas

different values were given for each cluster by cis_PRT2. Two of the eight samples cluster together under the value of 2 for cis_PRT2 (the top red cluster), which means that these samples have a homozygous deletion of DEL1 (zero copy). In addition, three of the eight samples cluster together under the value of 0.5 for cis_PRT2 (the middle red cluster), which means that these samples have both heterozygous deletions of DEL1 and DEL2 (zero copy). However, the last three of the eight samples cluster together under the value of 0.0 for both assays (the bottom red cluster), which means that these samples have a homozygous deletion of DEL2 (zero copy).

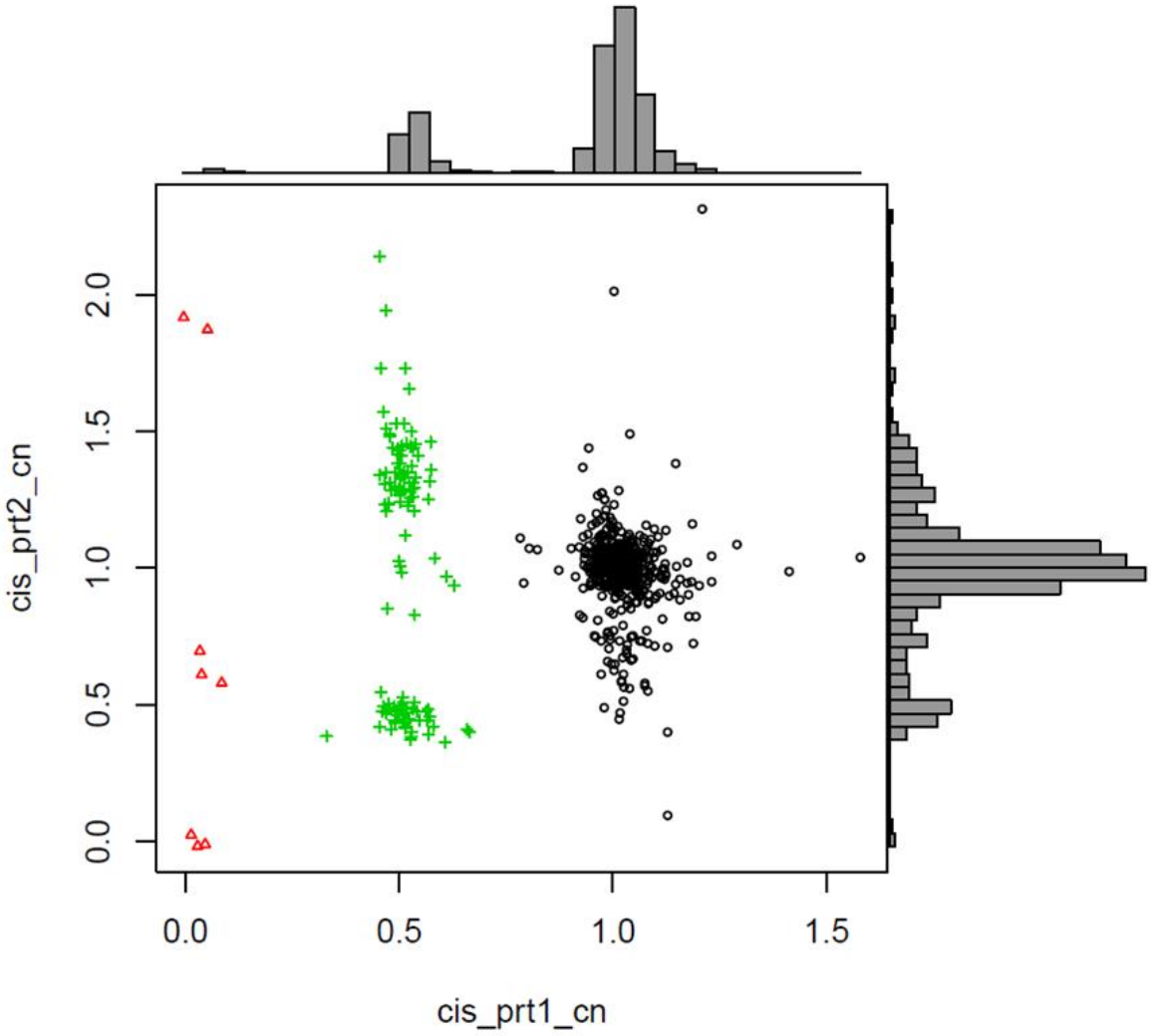



**Figure 5.4: Raw data comparison of cis_PRT1 and cis_PRT2 for the Benin malaria cohort.** The scatterplot shows the comparison of the raw data normalized to determine the copy number of the cis_PRT1 data (X-axis) and raw data normalized to determine the copy number of the cis_PRT2 data (Y-axis) in the Tori Bossito cohort (Benin malaria cohort). The colours black, green and red indicate two copies (normal), one copy (heterozygous deletion), and zero copy (homozygous deletion) respectively.

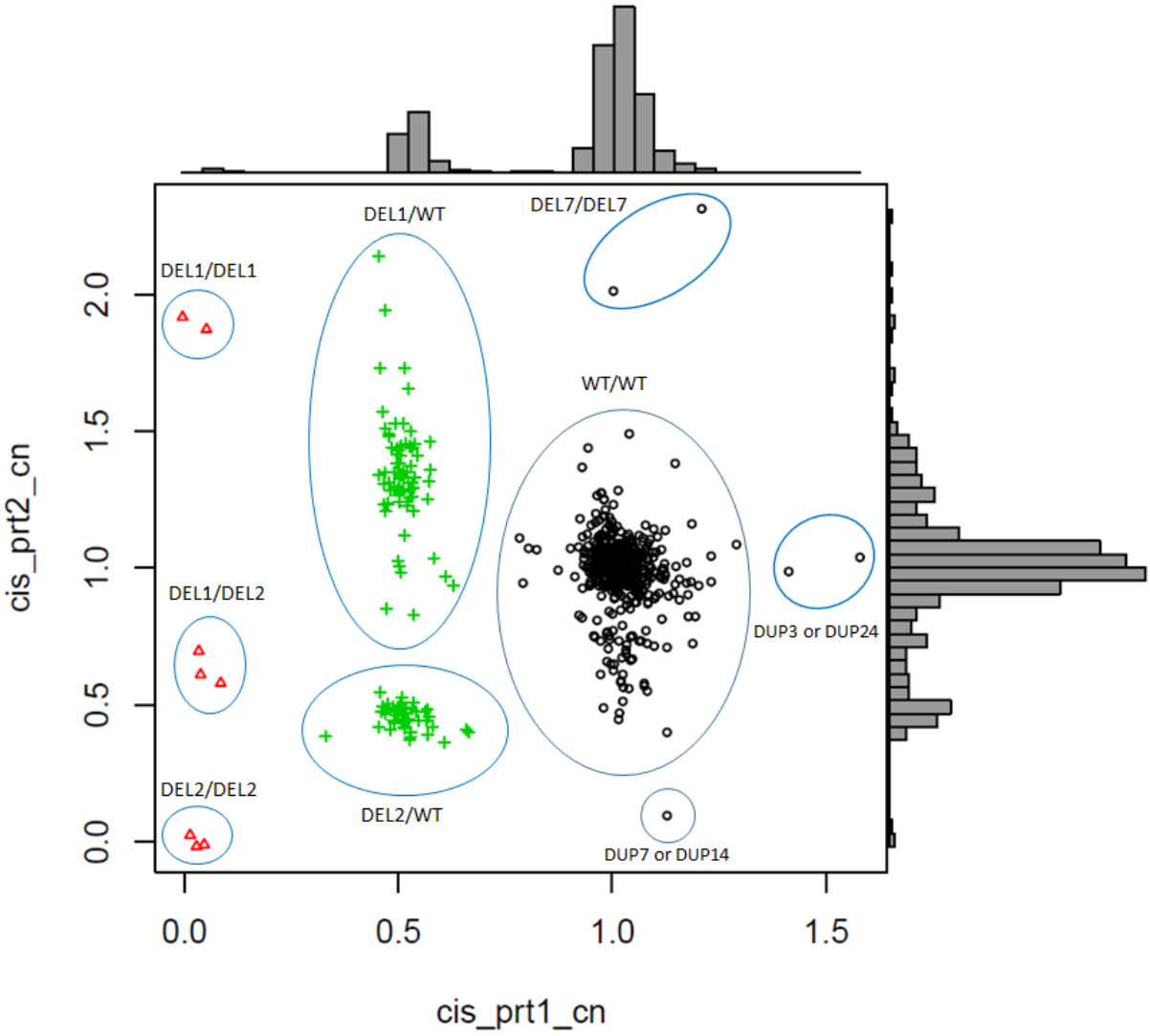



**Figure 5.5: The cis_PRT1 vs cis_PRT2 clusters for the Benin malaria cohort.** The scatterplot shows the figure 5.6 with highlighting the genotypes. The colours black, green and red indicate two copies (normal), one copy (heterozygous deletion), and zero copy (homozygous deletion) respectively.

On the other hand, the scatterplot of Cis_PRT4 against Cis_PRT1 of the 574 Benin malaria cohort samples shows three general clusters, each representing a specific genotype (Figure 5.6). The scatterplot shows that 446 samples cluster together under the value of 1 for both assays (the black cluster), indicating that these samples have normal copy numbers (two copies). However, two samples at the top of the black cluster show values of >1.5 and 1.0 with cis_PRT4 and cis_PRT1 respectively (DEL7), two samples on the right of the black cluster show values of 1.0 and 1.5 with cis_PRT4 and cis_PRT1 respectively (DUP3 or DUP24), and one sample at the bottom of the black cluster show values of <0.5 and 1.0 with cis_PRT4 and cis_PRT1 respectively (DUP7 or DUP14). The prediction of each variant was

relying on the chapter 3 and 4 results and the glycophorin PRT assays guideline table (Chapter4, Table 4.6).

The scatterplot shows that 120 samples cluster together under the value of 0.5 for both assays (the green cluster), which means that these samples are heterozygous deletions (one copy). However, from this scatterplot alone it is hard to predict which samples carry the DEL1 allele and which carry the DEL2 allele. Moreover, the scatterplot shows that eight samples cluster together under the value of zero for both assays (the red cluster), which means that these samples are homozygous deletions (zero copy). Although cis_PRT4 can detect *GYPB* deletions, it is hard to say from this scatterplot alone which samples from the red cluster (homozygous for *GYPB* deletion) and the green cluster (heterozygous for *GYPB* deletion) have the DEL1 allele and which have the DEL2 allele. That is why it is important to compare both scatterplots (cis_PRT2 vs cis_PRT1) and (cis_PRT4 vs cis_PRT1) as the cis_PRT2 scatterplot separates the red cluster to differentiate between the homozygous DEL1 and the homozygous DEL2 and it separates the red cluster differentiate between the heterozygous DEL1 and the heterozygous DEL2 (Figure 5.4).

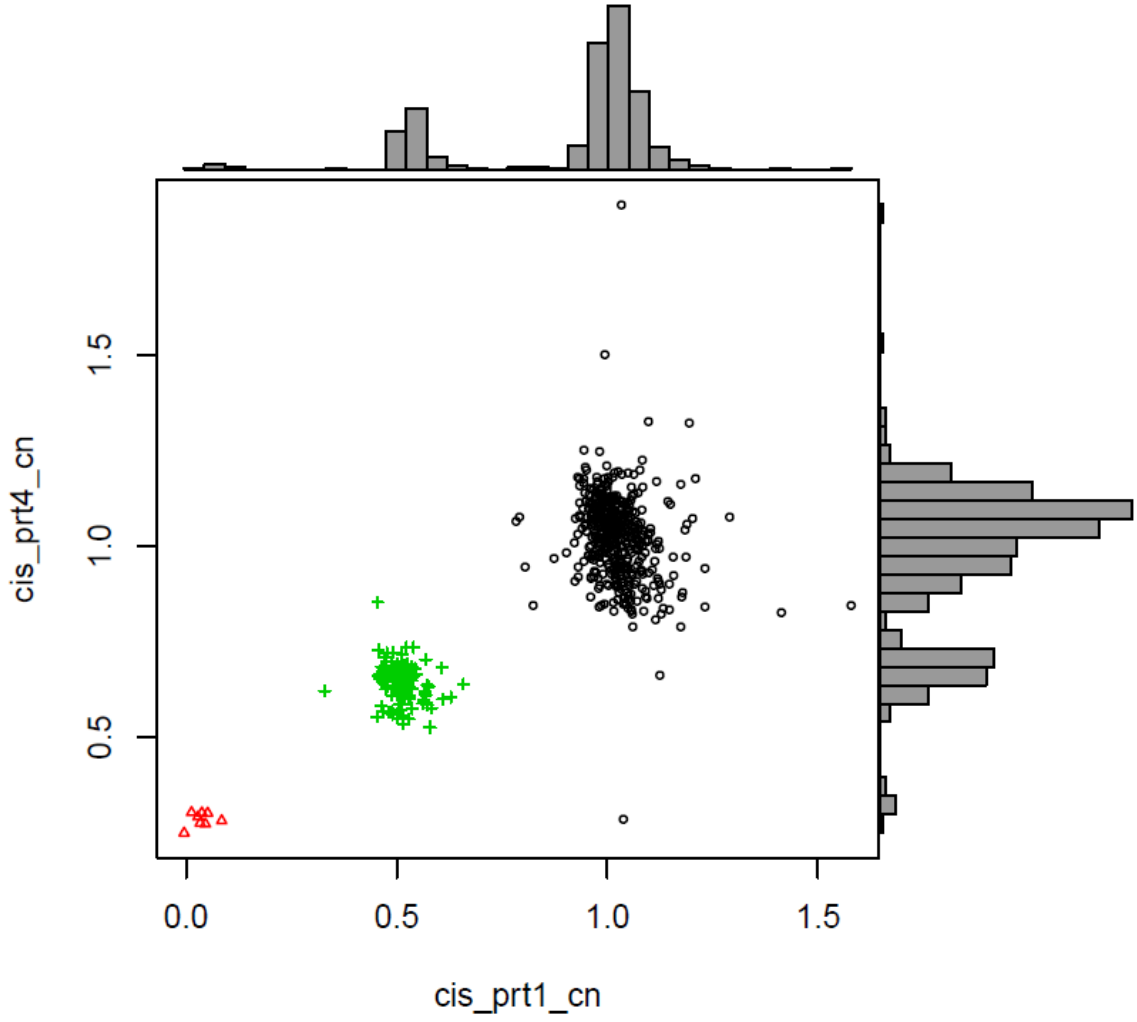



**Figure 5.6: Raw data comparison of cis_PRT1 and cis_PRT4 for the Benin malaria cohort.** The scatterplot shows the comparison between raw data normalized to determine the copy number of cis_PRT1 data (X-axis) and raw data normalized to determine the copy number of cis_PRT4 data (Y-axis) in the Tori Bossito cohort (Benin malaria cohort). The colours black, green and red indicate two copies (normal), one copy (heterozygous deletion), and zero copy (homozygous deletion) respectively.

The two scatterplots show that from the 574 Tori-Bossito cohort (Benin malaria cohort) samples, 446 samples are normal with two copies, 120 samples are heterozygous deletions (one copy) and eight samples are homozygous deletions (zero copy).

As has been observed in the cis_PRT2 vs cis_PRT1 scatterplot, seven clusters are shown, one cluster for the normal copies, two green clusters for the heterozygous deletions and three red clusters for the homozygous deletions. Cluster analysis was used (Appendix 9) to classify heterozygous samples for DEL1/WT and samples for DEL2/WT (Figure 5.7). 70 samples are heterozygous for DEL1/WT, which is shown as a black cluster, and 50 samples are heterozygous for DEL2/WT, which is shown as a red cluster.

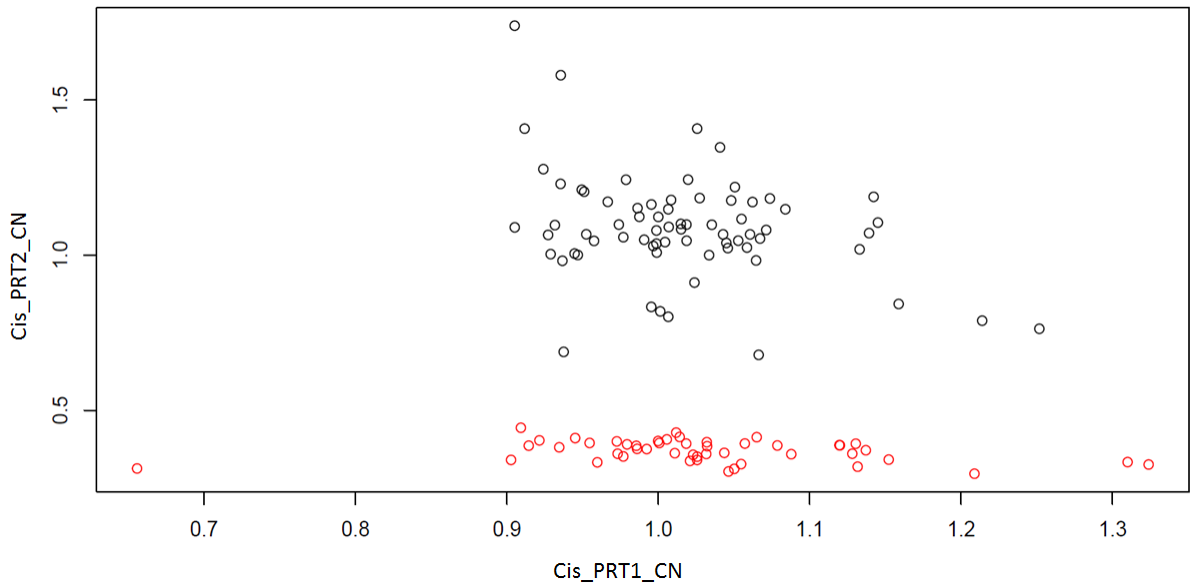



**Figure 5.7: A scatterplot of the cluster analysis of the (cis_PRT2 vs cis_PRT1) heterozygous deletion clusters.** The X axis represents the normalized raw data of cis_PRT1 and the Y axis represents the normalized raw data of cis_PRT2. The black circles are grouped together as one cluster (DEL1/WT) and the red circles are grouped together as one cluster (DEL2/WT).

Cluster analysis was applied (Appendix 9) to the homozygous deletion clusters in the cis_PRT2 vs cis_PRT1 scatterplot to confirm the separation between these three clusters in order to confidently differentiate between DEL1 and DEL2. Two samples are homozygous for DEL1, which is shown as a green cluster, three samples are homozygous for DEL2, which is shown as a black cluster, and three samples are carrying both DEL1 and DEL2 alleles, which is shown as a red cluster (Figure 5.8).

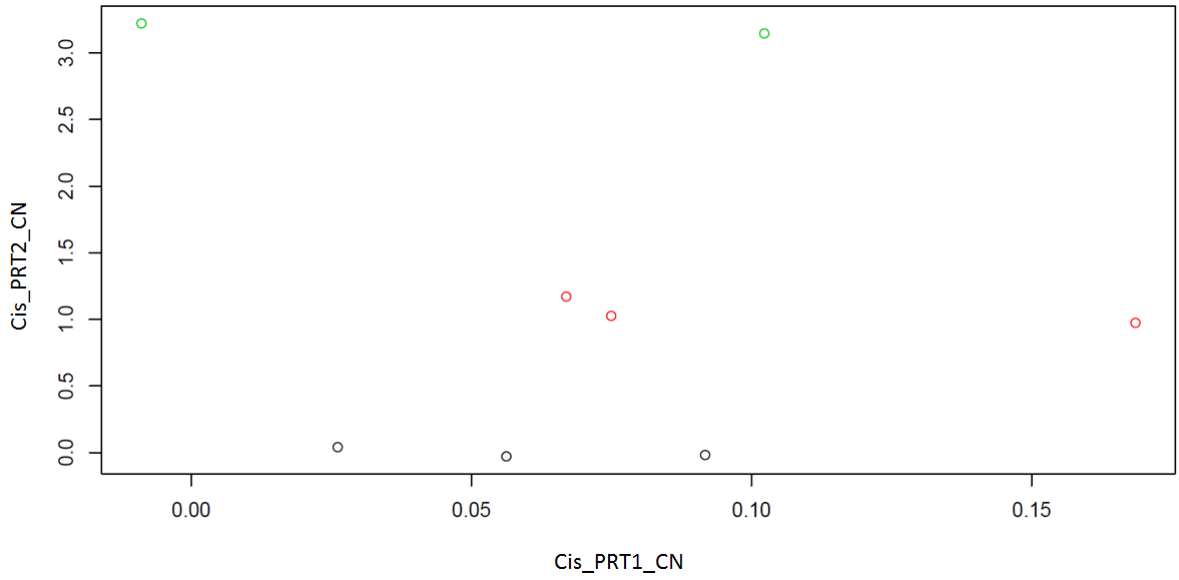



**Figure 5.8: A scatterplot of the cluster analysis of the (cis_PRT2 vs cis_PRT1) homozygous deletion clusters.** The X axis represents the normalized raw data of cis_PRT1 and the Y axis represents the normalized raw data of cis_PRT2. Each coloured circle is grouped together as an

independent cluster, green, red and black for DEL1/DEL1, DEL1/DEL2 and DEL2/DEL2 respectively.

The genotype count for *GYPB* CNV for DEL1, DEL2 and *GYPB*_DEL (both DEL1 and DEL2) are shown in the tables (5.3 - 5.5). The allele frequencies have been calculated for DEL1, DEL2 and *GYPB*_DEL variant and show no significant departure from Hardy-Weinberg equilibrium ($p>0.05$) (Tables 5.3-5.5).

**Table 5.3: Table showing the number of individuals observed for each DEL1 genotype.**

| Genotype | Observed number of individuals |
|---|---|
| **No DEL1** | 499 |
| **DEL1/WT** | 73 |
| **DEL1/DEL1** | 2 |
| **Total** | 574 |

| | |
|---|---|
| **DEL1 allele frequency** | DEL1 – ((2x2) + 73)/ 574x2 = 0.07 |

**Table 5.4: Table showing the number of individuals observed for each DEL2 genotype.**

| Genotype | Observed number of individuals |
|---|---|
| **No DEL2** | 518 |
| **DEL2/WT** | 53 |
| **DEL2/DEL2** | 3 |
| **Total** | 574 |

| | |
|---|---|
| **DEL2 allele frequency** | DEL2 – ((3x2) + 53)/ 574x2 = 0.05 |

**Table 5.5: Table showing the number of individuals observed for each (DEL1+DEL2) genotype.**

| Genotype | Observed number of individuals |
|---|---|
| **No DEL1 or DEL2** | 446 |
| ***GYPB*_DEL/WT** | 120 |
| ***GYPB*_DEL/*GYPB*_DEL** | 8 |
| **Total** | 574 |

| | |
|---|---|
| ***GYPB*_DEL allele frequency** | *GYPB*_DEL – ((8x2) + 120)/ 574x2 = 0.12 |

### 5.3.2 The effect of glycophorin variants on the time to first malarial infection

Considering the environmental parameters linked with the onset of early malaria infections is fundamental to estimate the risk of malaria exposition for every newborn. Several environmental parameters were evaluated, with the collection of entomologic, geographic, nutritional and climatic data (Le Port *et al*., 2012). The first malarial infection was defined with reference to first parasite density after the malaria attack. The Tori-Bossito cohort comprised 580 samples, but a few samples were not included in the analysis as some of the data for them were missing. Therefore, the total number of the samples used for the analysis was 555. In this cohort, a test was conducted to examine the association between the glycophorin copy numbers (DEL1, DEL2 and DEL_*GYPB*) and the time to first malarial infection using sex, maternal age, use of mosquito nets, sickle cell anaemia, ethnicity and malarial treatments, as covariates. In this analysis an additive model for the effect of variant on outcome phenotype was associated. Both DEL1 and DEL2 are full deletions of *GYPB* but with different breakpoints, so DEL1 and DEL2 are combined in a third analysis (DEL_*GYPB*) and test the association of the *GYPB* copy number with time to first malaria infection in this cohort.

The Cox regression analysis shows a protective effect of the use of mosquito nets, which a p value of 0.013 and the presence of sickle cell anaemia which a p value of 0.027. In terms of the use of mosquito nets, the absence of a mosquito net leads to an HR (hazard ratio) of 0.744, because having a mosquito net (Yes,1) is the reference. The presence of sickle cell anaemia leads to an HR of 0.447, because not having the sickle cell trait (No,1) is the reference (Table 5.6). Therefore, individuals with sickle cell anaemia or individuals using mosquito nets are protected against malaria infection.

There was no association between the DEL1 glycophorin copy number (p=0.856) and malaria in the Tori-Bossito cohort. The DEL1 result shows a HR of 1.029 with lower 0.753 and upper 1.406 values of the 95.0% CI (confidence interval). The results are summarized below (Table 5.6). Individuals with DEL1 genotypes (homozygous or heterozygous) had no effect on the time to onset of malaria disease, which means there is no delay or acceleration of the disease.

The same analysis for DEL1 was repeated for DEL2 and the result shows that there was no association between the DEL2 glycophorin copy number (p=0.856) and malaria in the Tori-Bossito cohort. The DEL2 result shows a HR of 1.265 with lower 0.902 and upper 1.773 values of the 95.0% CI (confidence interval). The results are summarized below (Table 5.7).

Individuals with DEL2 genotypes (homozygous or heterozygous) had no effect on the time to the onset of malaria disease, which means there is no delay or acceleration of the disease.

In addition, the same analysis was repeated for a third time counting all *GYPB* deletions and the result shows that there was no association between DEL_*GYPB* (p=0.856) and malaria in the Tori-Bossito cohort. The DEL_*GYPB* result shows a HR of 1.146 with lower 0.895 and upper 1.467 values of the 95.0% CI (confidence interval). The results are summarized below (Table 5.8). Individuals with DEL1, DEL2 or both of them (a homozygous or heterozygous deletion of the *GYPB* gene) had no effect on the time to the onset of malaria disease, which means there is no delay or acceleration of the disease.

**Table 5.6: Cox regression analysis for DEL1 for the Tori-Bossito cohort, with days until disease.** The dependent variable was the time to first malarial infection (days). **Bold Results**=significant, **df**= degree of freedom, **sig**.=significance, **CI**=the confidence interval, **Exp(B)**=the exponentiation of the B coefficient, **SE**=standard errors associated with the coefficients, **Wald**= Wald chi-square value and **B**=the coefficient for the constant.

| Parameter | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| **Maternal Age** | -0.010 | 0.010 | 1.000 | 1 | 0.317 | 0.990 | 0.970 | 1.010 |
| **Malaria suspicion during pregnancy** (No=0; Yes=1, reference) | -0.047 | 0.134 | 0.124 | 1 | 0.725 | 0.954 | 0.733 | 1.241 |
| **Sex** (M=1, reference; F=0) | -0.021 | 0.114 | 0.033 | 1 | 0.856 | 0.980 | 0.784 | 1.224 |
| **Birth term** | 0.006 | 0.028 | 0.038 | 1 | 0.846 | 1.006 | 0.951 | 1.063 |
| **Mosquito net** (No=0; Yes=1, reference) | -0.295 | 0.119 | 6.116 | 1 | **0.013** | 0.744 | 0.589 | 0.941 |
| **Ethnic group** (Tori=0; Fon=1; other=2) | | | 3.898 | 2 | 0.142 | | | |
| **Ethnic group** (Tori vs other/Fon (reference)) | 0.102 | 0.183 | 0.309 | 1 | 0.579 | 1.107 | 0.773 | 1.587 |
| **Ethnic group** (Fon vs Tori/other(reference)) | -0.296 | 0.164 | 3.244 | 1 | 0.072 | 0.744 | 0.539 | 1.026 |
| **Sickle cell** (Yes=0; No=1, reference) | -0.805 | 0.363 | 4.916 | 1 | **0.027** | 0.447 | 0.219 | 0.911 |
| **Chloroquine intake during pregnancy** (Yes=0; No=1; Unknown=2) | | | 1.613 | 2 | 0.446 | | | |
| **Chloroquine intake during pregnancy** (Yes vs No/unknown(reference)) | 0.029 | 0.146 | 0.040 | 1 | 0.841 | 1.030 | 0.773 | 1.371 |
| **Chloroquine intake during pregnancy** (No vs Yes/unknown(reference)) | -1.244 | 1.016 | 1.499 | 1 | 0.221 | 0.288 | 0.039 | 2.112 |
| **DEL1** (0,1 and 2 copies) | 0.029 | 0.159 | 0.033 | 1 | 0.856 | 1.029 | 0.753 | 1.406 |

**Table 5.7: Cox regression analysis for DEL2 for the Tori-Bossito cohort, with days until disease.** The dependent variable was the time to first malarial infection (days). **Bold Results**=significant, **df**= degree of freedom, **sig**.=significance, **CI**=the confidence interval, **Exp(B)**=the exponentiation of the B coefficient, **SE**=standard errors associated with the coefficients, **Wald**= Wald chi-square value and **B**=the coefficient for the constant.

| Parameter | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| **Maternal Age** | -0.009 | 0.010 | 0.784 | 1 | 0.376 | 0.991 | 0.971 | 1.011 |
| **Malaria suspicion during pregnancy** (No=0; Yes=1, reference) | -0.046 | 0.134 | 0.117 | 1 | 0.732 | 0.955 | 0.734 | 1.242 |
| **Sex** (M=1, reference; F=0) | -0.016 | 0.113 | 0.019 | 1 | 0.890 | 0.984 | 0.789 | 1.229 |
| **Birth term** | 0.008 | 0.029 | 0.077 | 1 | 0.781 | 1.008 | 0.953 | 1.066 |
| **Mosquito net** (No=0; Yes=1, reference) | -0.288 | 0.119 | 5.812 | 1 | **0.016** | 0.750 | 0.594 | 0.948 |
| **Ethnic group** (Tori=0; Fon=1; other=2) | | | 3.717 | 2 | 0.156 | | | |
| **Ethnic group** (Tori vs other/Fon (reference)) | 0.118 | 0.184 | 0.415 | 1 | 0.520 | 1.126 | 0.785 | 1.614 |
| **Ethnic group** (Fon vs Tori/other(reference)) | -0.281 | 0.164 | 2.915 | 1 | 0.088 | 0.755 | 0.547 | 1.042 |
| **Sickle cell** (Yes=0; No=1, reference) | -0.806 | 0.363 | 4.936 | 1 | **0.026** | 0.447 | 0.220 | 0.909 |
| **Chloroquine intake during pregnancy** (Yes=0; No=1; Unknown=2) | | | 1.566 | 2 | 0.457 | | | |
| **Chloroquine intake during pregnancy** (Yes vs No/unknown(reference)) | 0.024 | 0.146 | 0.027 | 1 | 0.869 | 1.024 | 0.770 | 1.363 |
| **Chloroquine intake during pregnancy** (No vs Yes/unknown(reference)) | -1.234 | 1.016 | 1.474 | 1 | 0.225 | 0.291 | 0.040 | 2.134 |
| **DEL2** (0,1 and 2 copies) | 0.235 | 0.173 | 1.851 | 1 | 0.174 | 1.265 | 0.902 | 1.773 |

**Table 5.8: Cox regression analysis for DEL1 and DEL2 (DEL_GYPB) for the Tori-Bossito cohort, with days until disease.** The dependent variable was time to first malarial infection (days). **Bold Results**=significant, **df**= degree of freedom, **sig**.=significance, **CI**=the confidence interval, **Exp(B)**=the exponentiation of the B coefficient, **SE**=standard errors associated with the coefficients, **Wald**= Wald chi-square value and **B**=the coefficient for the constant.

| Parameter | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| **Maternal Age** | -0.009 | 0.010 | 0.874 | 1 | 0.350 | 0.991 | 0.971 | 1.010 |
| **Malaria suspicion during pregnancy** (No=0; Yes=1, reference) | -0.053 | 0.134 | 0.158 | 1 | 0.691 | 0.948 | 0.729 | 1.233 |
| **Sex** (M=1, reference; F=0) | -0.029 | 0.114 | 0.066 | 1 | 0.797 | 0.971 | 0.777 | 1.213 |
| **Birth term** | 0.008 | 0.029 | 0.074 | 1 | 0.785 | 1.008 | 0.953 | 1.066 |
| **Mosquito net** (No=0; Yes=1, reference) | -0.294 | 0.119 | 6.086 | 1 | **0.014** | 0.745 | 0.590 | 0.941 |
| **Ethnic group** (Tori=0; Fon=1; other=2) | | | 3.524 | 2 | 0.172 | | | |
| **Ethnic group** (Tori vs other/Fon (reference)) | 0.112 | 0.184 | 0.374 | 1 | 0.541 | 1.119 | 0.780 | 1.604 |
| **Ethnic group** (Fon vs Tori/other(reference)) | -0.276 | 0.165 | 2.789 | 1 | 0.095 | 0.759 | 0.549 | 1.049 |
| **Sickle cell** (Yes=0; No=1, reference) | -0.788 | 0.363 | 4.713 | 1 | **0.030** | 0.455 | 0.223 | 0.926 |
| **Chloroquine intake during pregnancy** (Yes=0; No=1; Unknown=2) | | | 1.514 | 2 | 0.469 | | | |
| **Chloroquine intake during pregnancy** (Yes vs No/unknown(reference)) | 0.016 | 0.146 | 0.012 | 1 | 0.912 | 1.016 | 0.763 | 1.354 |
| **Chloroquine intake during pregnancy** (No vs Yes/unknown(reference)) | -1.226 | 1.017 | 1.455 | 1 | 0.228 | 0.293 | 0.040 | 2.152 |
| **DEL_*GYPB*** (0,1 and 2 copies) | 0.136 | 0.126 | 1.166 | 1 | 0.280 | 1.146 | 0.895 | 1.467 |

### 5.3.3 Analysis of glycophorin variants and number of malarial infections

An alternative outcome variable was used in the next set of association studies: the number of malaria infections. Because this outcome is represented by count data, a Poisson regression was used to assess the effect of the following factors on the risk of contracting malaria: maternal age, malaria suspicion during pregnancy, sex, birth term, mosquito net, ethnic group Tori, chloroquine intake during pregnancy, sickle cell anaemia, and genetic polymorphism. Three glycophorin variants (DEL1, DEL2 and DEL_*GYPB*) were analysed in turn. None of the three variants showed a statistically significant result.

The results for DEL1 association show that the absence of a mosquito net and sickle cell anaemia increase the number of infections. Interestingly, maternal age (p=0.044) and ethnic group (p=0.028) have a significant effect on the number of breakthrough infections in this cohort (Table 5.9). In the case of DEL2, the results show a confirmation of the effect of the use of a mosquito net and the presence of sickle cell anaemia on the number of infections. In addition, in these results, the ethnic group (p=0.026) has a significant effect on the number of breakthrough infections in this cohort. However, maternal age has an almost significant effect on the number of breakthrough infections in this cohort (p= 0.057), as shown below (Table 5.10). However, the association results for the combination of both DEL1 and DEL2 data (DEL_*GYPB*) show that the presence of a mosquito net and the sickle cell anaemia decrease the number of infections. In addition, the results confirm the significant effect of ethnic group on the number of breakthrough infections in this cohort (p=0.033). Nevertheless, maternal age has an almost significant effect on the number of breakthrough infections in this cohort (p= 0.053), as shown below (Table 5.11). Therefore, maternal age could be a crucial cofactor that affects the number of malaria infections of infants.

**Table 5.9: For the Poisson generalized linear model, the same covariates (with DEL1) were used with the number of independent malarial infections used as the dependent variable. Bold Results**=significant, **df**= degree of freedom, **sig.**=significance, **CI**=the confidence interval, **Std. Error**=standard error and **B**=the coefficient for the constant.

| Parameters Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Parameter** | **B** | **Std. Error** | **95% Wald Confidence Interval** | | **Hypothesis Test** | | |
| | | | **Lower** | **Upper** | **Wald Chi-Square** | **df** | **Sig.** |
| **Maternal Age** | -0.015 | 0.0072 | -0.029 | 0.000 | 4.045 | 1 | **0.044** |
| **Malaria suspicion during pregnancy** (0) (No=0; Yes=1) | 0.150 | 0.0959 | -0.038 | 0.338 | 2.438 | 1 | 0.118 |
| **Malaria suspicion during pregnancy** (1, reference) (No=0; Yes=1) | 0[a] | | | | | | |
| **Sex** (0) (M=1, reference ; F=0) | 0.095 | 0.0782 | -0.058 | 0.248 | 1.469 | 1 | 0.226 |
| **Sex** (1, reference) | 0[a] | . | . | . | . | . | . |
| **Birth term** | 0.016 | 0.0197 | -0.023 | 0.054 | 0.644 | 1 | 0.422 |
| **Mosquito net** (0) (No=0; Yes=1, reference) | 0.233 | 0.0819 | 0.073 | 0.394 | 8.124 | 1 | **0.004** |
| **Mosquito net** (1, reference) | 0[a] | | | | | | |
| **Ethnic group** (Tori vs other/Fon (reference)) | 0.258 | 0.1173 | 0.028 | 0.488 | 4.851 | 1 | **0.028** |
| **Ethnic group** (Fon vs Tori/other(reference)) | 0.250 | 0.1591 | -0.062 | 0.562 | 2.460 | 1 | 0.117 |
| **Ethnic group** (Tori=0; Fon=1; other=2, reference) | 0[a] | | | | | | |
| **Sickle cell** (1) (Yes=0; No=1, reference) | 0.687 | 0.2107 | 0.274 | 1.100 | 10.642 | 1 | **0.001** |
| **Sickle cell** (1, reference) | 0[a] | | | | | | |
| **Chloroquine intake during pregnancy** (Yes vs No/unknown (reference)) | 0.354 | 0.4702 | -0.567 | 1.276 | 0.567 | 1 | 0.451 |
| **Chloroquine intake during pregnancy** (No vs Yes/unknown (reference)) | 0.284 | 0.4670 | -0.632 | 1.199 | 0.369 | 1 | 0.544 |
| **Chloroquine intake during pregnancy** (Yes=0; No=1; Unknown=2, reference) | 0[a] | | | | | | |
| **DEL1** | 0.080 | 0.1074 | -0.131 | 0.290 | 0.548 | 1 | 0.459 |

**Table 5.10: For the Poisson generalized linear model, the same covariates (with DEL2) were used, with the number of independent malarial infections used as the dependent variable. Bold Results**=significant, **df**= degree of freedom, **sig.**=significance, **CI**=the confidence interval, **Std. Error**=standard error and **B**=the coefficient for the constant.

| Parameters Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Parameter** | **B** | **Std. Error** | **95% Wald Confidence Interval** | | **Hypothesis Test** | | |
| | | | **Lower** | **Upper** | **Wald Chi-Square** | **df** | **Sig.** |
| **Maternal Age** | -0.014 | 0.0073 | -0.028 | 0.000 | 3.619 | 1 | **0.057** |
| **Malaria suspicion during pregnancy** (0) (No=0; Yes=1) | 0.153 | 0.0959 | -0.035 | 0.341 | 2.531 | 1 | 0.112 |
| **Malaria suspicion during pregnancy** (1, reference) (No=0; Yes=1) | 0[a] | | | | | | |
| **Sex** (0) (M=1, reference ; F=0) | 0.082 | 0.0777 | -0.070 | 0.235 | 1.127 | 1 | 0.289 |
| **Sex** (1, reference) | 0[a] | | | | | | |
| **Birth term** | 0.016 | 0.0197 | -0.022 | 0.055 | 0.672 | 1 | 0.412 |
| **Mosquito net** (0) (No=0; Yes=1, reference) | 0.228 | 0.0818 | 0.068 | 0.388 | 7.766 | 1 | **0.005** |
| **Mosquito net** (1, reference) | 0[a] | | | | | | |
| **Ethnic group** (Tori vs other/Fon (reference)) | 0.260 | 0.1170 | 0.030 | 0.489 | 4.930 | 1 | **0.026** |
| **Ethnic group** (Fon vs Tori/other(reference)) | 0.260 | 0.1590 | -0.052 | 0.572 | 2.677 | 1 | 0.102 |
| **Ethnic group** (Tori=0; Fon=1; other=2, reference) | 0[a] | . | . | . | . | . | . |
| **Sickle cell** (1) (Yes=0; No=1, reference) | 0.679 | 0.2106 | 0.266 | 1.092 | 10.397 | 1 | **0.001** |
| **Sickle cell** (1, reference) | 0[a] | | | | | | |
| **Chloroquine intake during pregnancy** (Yes vs No/unknown (reference)) | 0.335 | 0.4702 | -0.586 | 1.257 | 0.509 | 1 | 0.476 |
| **Chloroquine intake during pregnancy** (No vs Yes/unknown (reference)) | 0.273 | 0.4666 | -0.642 | 1.187 | 0.342 | 1 | 0.559 |
| **Chloroquine intake during pregnancy** (Yes=0; No=1; Unknown=2, reference) | 0[a] | | | | | | |
| **DEL2** | 0.152 | 0.1081 | -0.060 | 0.364 | 1.983 | 1 | 0.159 |

**Table 5.11: For the Poisson generalized linear model, the same covariates (with DEL_GYPB) were used with dependent variable as number of independent malarial infections. Bold Results**=significant, **df**= degree of freedom, **sig.**=significance, **CI**=the confidence interval, **Std. Error**=standard error and **B**=the coefficient for the constant.

| Parameters Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Parameter** | **B** | **Std. Error** | **95% Wald Confidence Interval** | | **Hypothesis Test** | | |
| | | | **Lower** | **Upper** | **Wald Chi-Square** | **df** | **Sig.** |
| **Maternal Age** | -0.014 | 0.0073 | -0.028 | 0.000 | 3.732 | 1 | **0.053** |
| **Malaria suspicion during pregnancy** (0) (No=0; Yes=1) | 0.158 | 0.0961 | -0.031 | 0.346 | 2.692 | 1 | 0.101 |
| **Malaria suspicion during pregnancy** (1, reference) (No=0; Yes=1) | 0[a] | | | | | | |
| **Sex** (0) (M=1, reference ; F=0) | 0.094 | 0.0777 | -0.058 | 0.247 | 1.474 | 1 | 0.225 |
| **Sex** (1, reference) | 0[a] | | | | | | |
| **Birth term** | 0.016 | 0.0197 | -0.022 | 0.055 | 0.683 | 1 | 0.408 |
| **Mosquito net** (0) (No=0; Yes=1, reference) | 0.233 | 0.0819 | 0.073 | 0.394 | 8.116 | 1 | **0.004** |
| **Mosquito net** (1, reference) | 0[a] | | | | | | |
| **Ethnic group** (Tori vs other/Fon (reference)) | 0.250 | 0.1174 | 0.020 | 0.480 | 4.530 | 1 | **0.033** |
| **Ethnic group** (Fon vs Tori/other(reference)) | 0.251 | 0.1590 | -0.061 | 0.562 | 2.487 | 1 | 0.115 |
| **Ethnic group** (Tori=0; Fon=1; other=2, reference) | 0[a] | | | | | | |
| **Sickle cell** (1) (Yes=0; No=1, reference) | 0.668 | 0.2110 | 0.254 | 1.082 | 10.021 | 1 | **0.002** |
| **Sickle cell** (1, reference) | 0[a] | | | | | | |
| **Chloroquine intake during pregnancy** (Yes vs No/unknown (reference)) | 0.322 | 0.4705 | -0.600 | 1.244 | 0.468 | 1 | 0.494 |
| **Chloroquine intake during pregnancy** (No vs Yes/unknown (reference)) | 0.254 | 0.4672 | -0.662 | 1.169 | 0.295 | 1 | 0.587 |
| **Chloroquine intake during pregnancy** (Yes=0; No=1; Unknown=2, reference) | 0[a] | | | | | | |
| **DEL_*GYPB*** | 0.128 | 0.0809 | -0.030 | 0.287 | 2.520 | 1 | 0.112 |

### 5.3.4 Analysis of linkage disequilibrium between glycophorin CNVs and rs186873296 SNP in the Tori-Bossito Cohort

Genotyping the rs186873296 SNP directly in this cohort allow to test whether this SNP is in linkage disequilibrium with DEL1 or DEL2. The high-throughput sequencing read depth published data shows that the NA19190 DNA sample, which is one of the 1000 Genomes Project samples and used as one of the PRT DNA positive controls, is heterozygous for the G allele. This is the rare allele for SNP (rs186873296) according to the UCSC Genome Browser (GRCh37/hg19) assembly. Therefore, the NA19190 DNA sample was used as a heterozygous positive control for the SNP (rs186873296) genotyping. However, a homozygous positive control for the G allele of the SNP (rs186873296) could not be found as it is a very rare SNP. As a result, the ARMS assay was designed with a positive primer pair in addition to the main ARMS primer pair to ensure that the PCR had worked successfully; for each sample at least one amplification should occur (Chapter 2).

The ARMS assay designed in Chapter 2 was applied on the six PRT positive controls in order to evaluate them and to assess the NA19190 DNA sample before it was used as a heterozygous positive control for the rare allele (G) of the SNP. As was expected, the ARMS gel result shows that all six PRT positive controls had the expected band of the β-defensin positive primer pair (301 bp), while the NA19190 DNA sample showed an extra band (200 bp). This is a result of an amplification of the ARMS primer pair due to the presence of at least one G allele of the SNP (rs186873296) (Figure 5.9). However, the gel result for a homozygous sample of the rare allele is expected to be the same as for the NA19190 DNA sample, which is a heterozygous positive control for the rare allele.

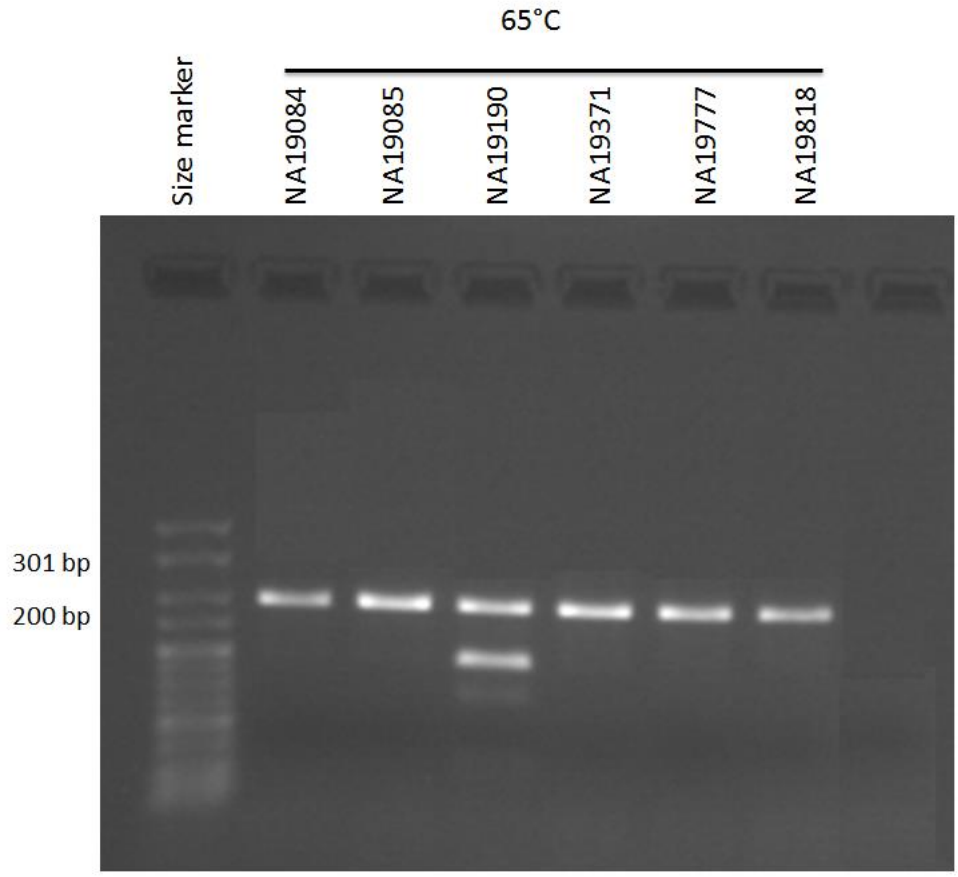



**Figure 5.9: Successful confirmation of the ARMS assay specificity.** The PCR gel result confirms the specificity of the ARMS assay. The ARMS DNA positive control (NA19190) (a G allele carrier) shows two bands, one from the positive control primer pair (301 bp) and one from the ARMS primer pair (200 bp). The other DNA negative controls (AA allele carrier) show only the band from the positive control primer pair (301 bp).

After confirming the success of the ARMS assay, 555 DNA samples from the Tori-Bossito cohort (Benin malaria cohort) were subjected to the assay to investigate the relationship between the SNP (rs186873296) and resistance to severe malaria. Specifically, the assay was applied on these samples in order to determine the allele frequency among this cohort and whether this rare (G) allele of SNP (rs186873296) exists in linkage disequilibrium with the copy number variations of glycophorin genes in the case of resistance to severe malaria.

The ARMS gel results for the 583 DNA samples from the Benin malaria cohort show that only seven samples showed amplification of the allele specific primers. The figure below shows an example of one set of the ARMS results (Figure 5.10). These seven samples (see Appendix 13 for the sample names) from the entire Benin cohort could be homozygous or heterozygous for the G allele of the SNP. The agarose gel cannot indicate whether the samples are homozygous or heterozygous.

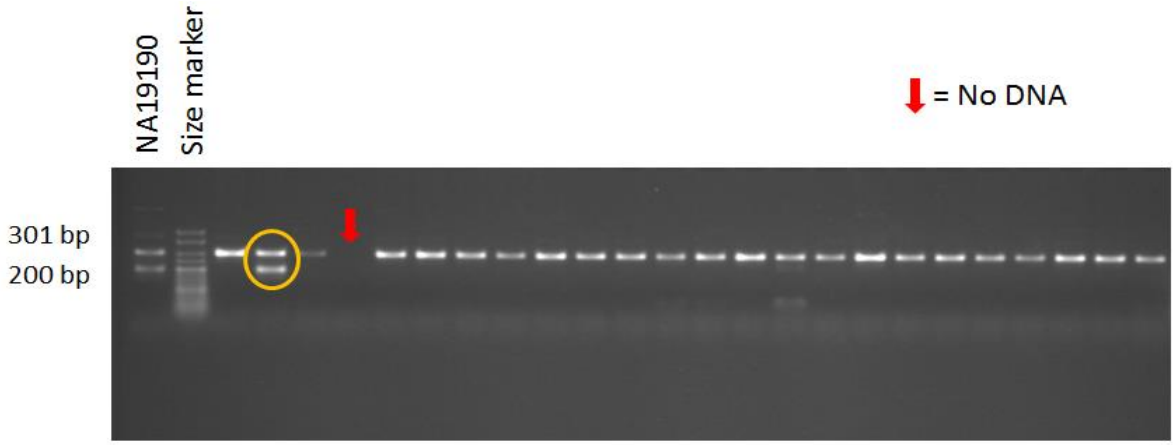



**Figure 5.10: Part of the ARMS gel results for 23 samples from the Benin cohort.** Samples with one band are homozygous for the A allele, whereas the sample surrounded by a yellow circle shows two bands, which means this sample is homozygous or heterozygous for the G allele. However, the positive control (NA19190) is known to be heterozygous for the G allele according to the NGS published data.

A standard PCR primer pair was specifically designed (Chapter 2) for the samples with the rare allele in order to sequence their product, which was designed to include the novel SNP (rs186873296) with a 397 bp size. The PCR assay was applied to the seven samples and sequenced by Sanger sequencing. The sequencing results of the PCR products of the seven G allele positive samples indicate that all seven Benin samples are heterozygous for the rare allele of the SNP (rs186873296) (Figure 5.11).

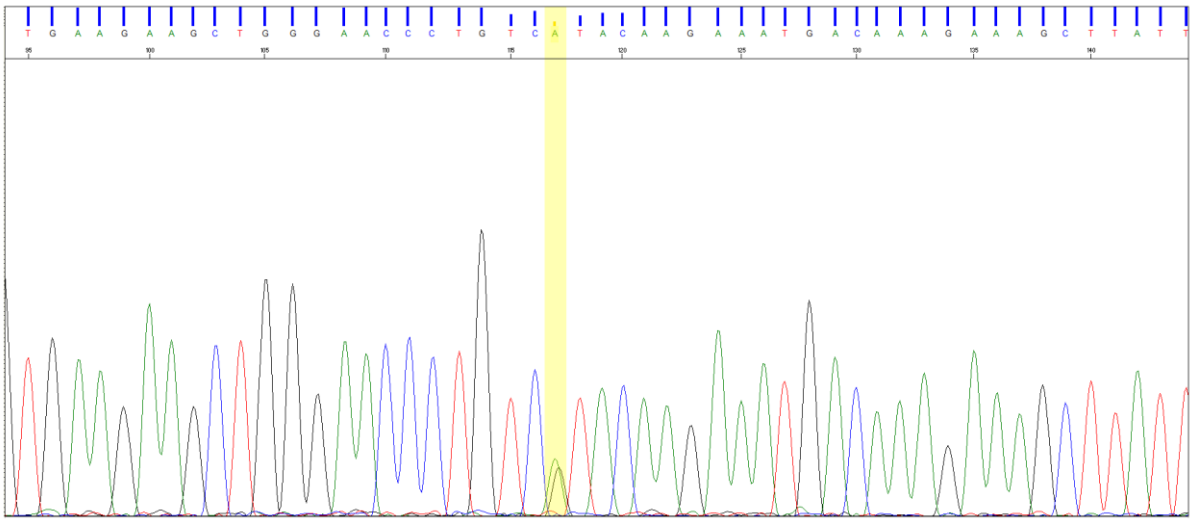


**Figure 5.11: A sequencing trace result of the ARMS sequencing primer set.** This is an example of one of the positive samples (A023) for the G allele. The yellow highlight indicates the SNP base, which shows two peaks (heterozygous): one green (A allele) and one black (G allele).

Linkage disequilibrium (LD) analysis was conducted for rs186873296 and the CNVs detected in the Benin malaria cohort across the glycophorin genes to determine whether the (rs186873296) SNP is in linkage disequilibrium with DEL1 and DEL2. Therefore, linkage disequilibrium analysis for the Tori-Bossito cohort was assessed once for the SNP and DEL1 and once for the SNP and DEL2. According to the results, rs186873296 was in low linkage disequilibrium with both DEL1 (Table 5.12) and DEL2 (Table 5.13) glycophorin variants. However, if rs186873296 is a protective allele against severe malaria, as has been reported in Band *et al*. (2015), it may be in linkage disequilibrium with another glycophorin variant that provides protection, but not the DEL1 and DEL2 glycophorin variants.

**Table 5.12: 3x3 table of observed and expected diplotype numbers.** Black numbers on white are the original data entered for DEL1. The table below represents the LD statistics table using the http://www.oege.org/software/cubex/ online software.

| SNP | CNV | | |
|---|---|---|---|
| | DEL1/DEL1 | DEL1 / + | + / + |
| **AA** | 2 | 76 | 470 |
| **AG** | 0 | 0 | 7 |
| **GG** | 0 | 0 | 0 |

| LD statistics | | |
|---|---|---|
| D' | $r^2$ | $x^2$ |
| 1.0 | 0.0005 | 0.28 |

**Table 5.13: 3x3 table of observed and expected diplotype numbers.** Black numbers on white are the original data entered for DEL2. The table below represents the LD statistics table using the http://www.oege.org/software/cubex/ online software.

| SNP | CNV | | |
|---|---|---|---|
| | DEL2/DEL2 | DEL2 / + | + / + |
| **AA** | 3 | 50 | 495 |
| **AG** | 0 | 0 | 7 |
| **GG** | 0 | 0 | 0 |

| LD statistics | | |
|---|---|---|
| D' | $r^2$ | $x^2$ |
| 1.0 | 0.0003 | 0.17 |

## 5.4 Analysis of DUP4 glycophorin variant in Tanzanians

### 5.4.1 Genotyping of DUP4 in the Nyamisati Tanzanian malaria cohort

As described in the previous chapters, DUP4 is an important and complex glycophorin variant. Because it has been reported that DUP4 is associated with severe malaria, a specific DUP4 PCR assay was designed and optimised by Paulina Bajer (MSc student) in order to assess the role of DUP4 in malaria infections, and to determine whether it is linked with any of the phenotypes of malaria in the Nyamisati Tanzanian malaria cohort. The DUP4 specific assay was applied on different glycophorin variant positive controls, which were NA19120 heterozygous for DEL1, HG04039 heterozygous for DEL6, HG03837 heterozygous for DUP24 and HG03686 heterozygous for DUP29, in order to evaluate the PCR assay before applying it to a large cohort (the Tanzanian cohort).

The expected gel result for any sample positive for the DUP4 allele is two bands. One is a 397 bp PCR product from a positive control primer pair, expected to amplify all samples, and the other band is a 500 bp PCR product from the specific DUP4 primer pair. The gel result confirms the specificity of the DUP4 PCR assay, and shows that only the known DUP4 positive control has two bands, whereas the other glycophorin variant positive controls show only one band. This is the result of the amplification of positive primers (Figure 5.12).

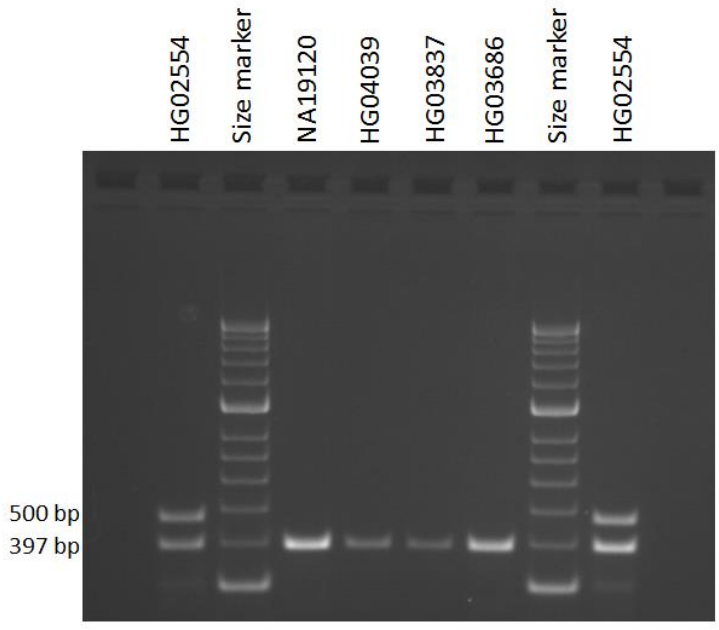



**Figure 5.12: Successful confirmation of the specific DUP4 assay.** The PCR gel result confirms the specificity of the specific DUP4 assay. The DUP4 DNA positive control (HG02554) (a DUP4 allele carrier) shows two bands, one from the positive control primer pair (397 bp) and one from the specific DUP4 primer pair (500 bp). The other DNA negative controls (with the DUP4 allele absent) show only the band from the positive control primer pair (397 bp). The marker size was HyperLadder™ 50bp – Bioline.

In order to calculate the frequency of the DUP4 allele in the Tanzanian cohort and to determine whether this genotype is associated with any of the phenotypes of malaria, the DUP4 specific assay was applied to 348 samples from the 922 Nyamisati Tanzanian cohorts. The DUP4 genotyping results indicate that 90 individuals from the Tanzanian malaria cohort carry a DUP4 allele that is either heterozygous or homozygous (Figure 5.13). The assay cannot distinguish between the heterozygous and homozygous DUP4 alleles because the primers only amplify when the DUP4 is present but it does not give any amplification with normal sequence, so at any situation (heterozygous or homozygous) the DUP4 specific primers will give only the 500 bp PCR product.

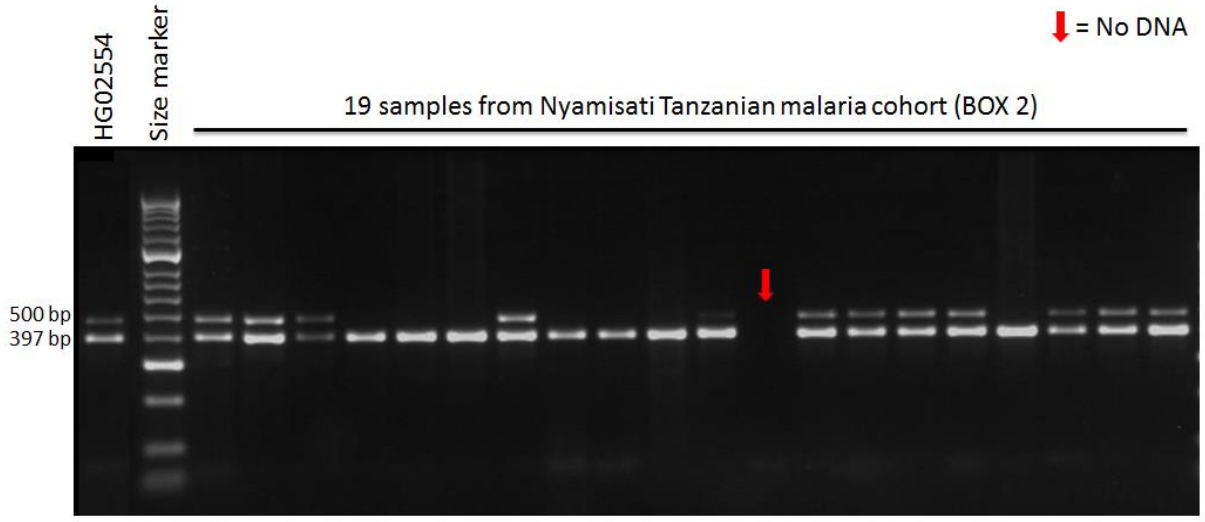



**Figure 5.13: Part of the Tanzanian cohort samples DUP4 assay gel results.** Samples with one band are the DUP4 negative allele, whereas the samples two bands are homozygous or heterozygous for the DUP4 allele. The HG02554 DNA sample is the DUP4 positive control.

Homozygous and heterozygous were distinguished by quantification of DUP4 and control band intensity on agarose gels using ImageJ version 1.51 software (Schneider *et al.*, 2012; https://imagej.net/Welcome), and calculating the ratio of DUP4: control band intensity for each individual. At low allele frequencies, homozygotes are expected to be rare. After log2 transformation, a cluster of four outliers of high ratio (>2SD, log2ratio>1.43) were clearly separated from the 256 other DUP4 positive samples, which were classified as heterozygotes (Figure 5.14).

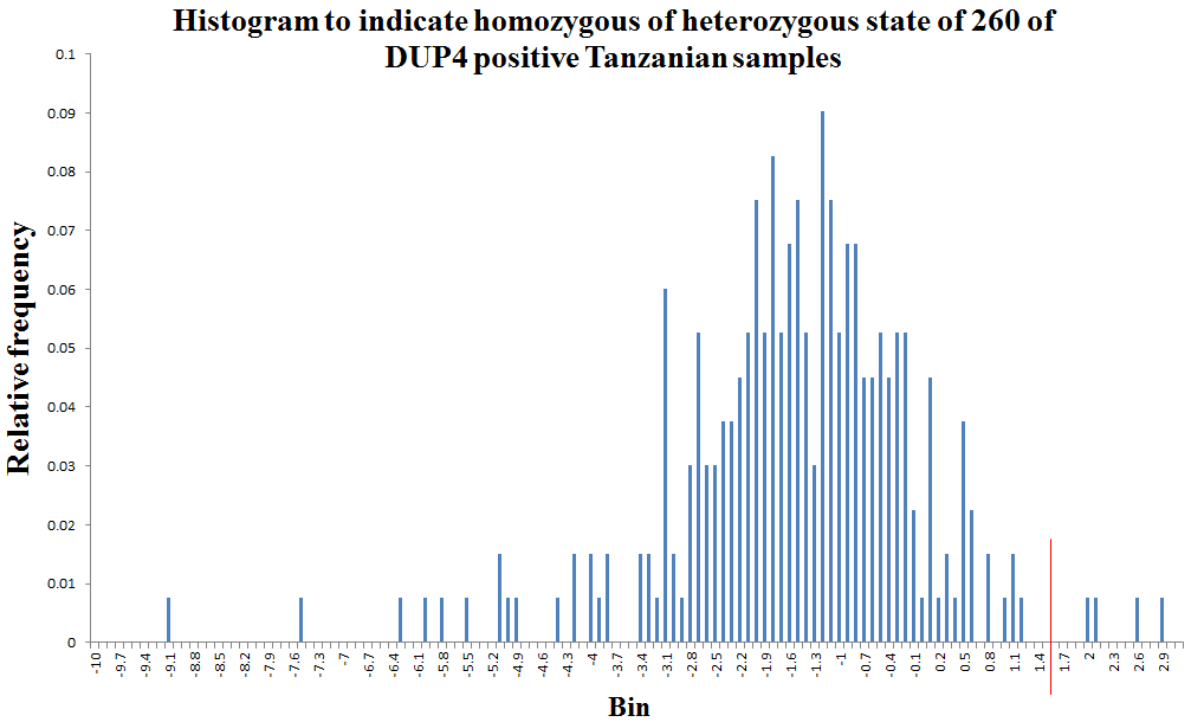



**Figure 5.14: Histogram used to identify homozygous or heterozygous state of the DUP4 positive samples from box 2 (260 samples) of the Tanzanian cohort.** The X axis represents the bin, which refers to the sample log2ratios. The Y axis represents the relative frequency to the number of samples that exhibit the same log2ratio. The red line represents the cut-off (log2 ratio of 1.43) that corresponded to two standard deviations away from the mean.

DUP4 positive control (HG02554) was Illumina sequenced together with two DUP4 homozygotes and three DUP4 heterozygotes genomic DNA at high depth (50x). The same pattern of DUP4 that observed previously in Leffler *et al*. (2017) has shown from the analysis of sequence read depth across the glycophorin repeat region. DUP4 homozygotes show the expected increase to 4 copies and 6 copies in duplicated and triplicated regions, respectively (Algady *et al*., 2018). The sequence read depth of the DUP4 positive control (HG02554), and two homozygotes and the two heterozygotes from Nyamisati Tanzanian cohort Illumina sequenced at high depth were plotted in a 5kb window plot (Figure 5.15).

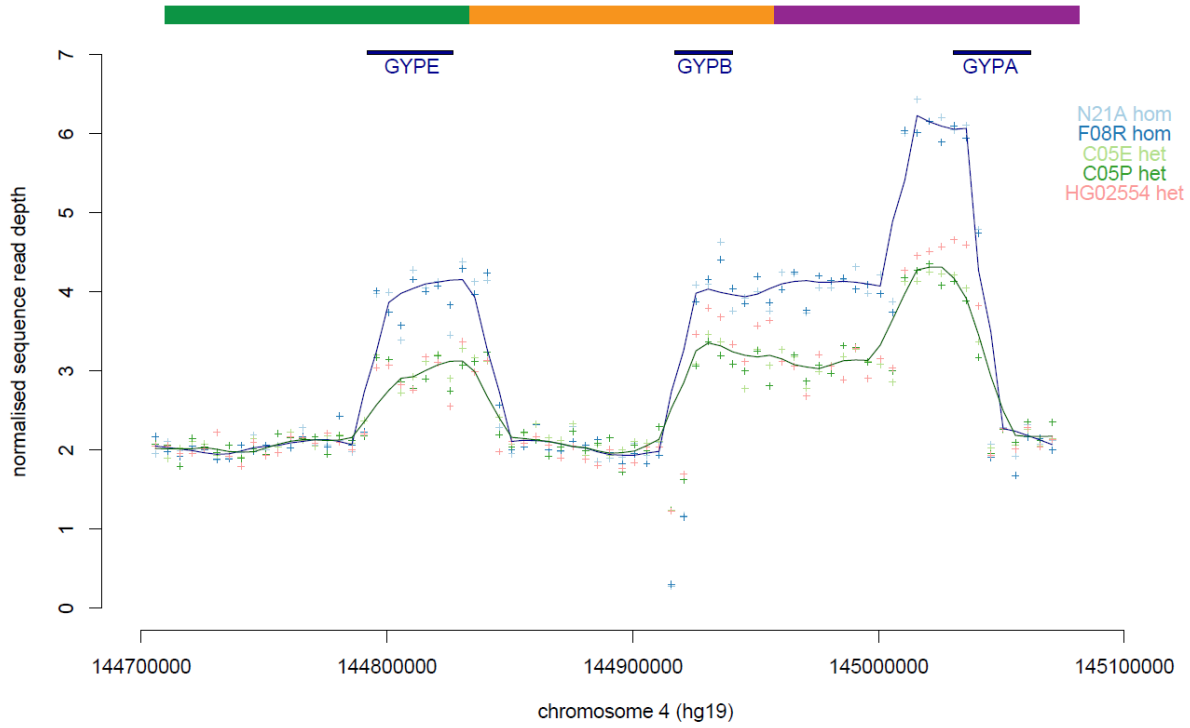



**Figure 5.15: Sequence read depth analysis of DUP4 homozygotes and heterozygotes.** Normalised sequence read depth of 5 kb windows five samples spanning the reference sequence glycophorin region. The blue Loess regression line (f=0.1) represents homozygotes and the green represents heterozygotes. Gene positions and repeats are shown in the top of the plot according to the reference.

The majority of samples were individuals from family groups, and unrelated individuals are needed to determine allele frequency and test for departure from Hardy-Weinberg equilibrium. A total of 348 samples from unrelated individuals were analysed using the DUP4 specific assay. The homozygous and heterozygous state was identified and the results show that only four samples were homozygous for the DUP4 allele, while the remaining 86 samples were heterozygous for the DUP4 allele (Table 5.14). Thus, the allele frequencies have been calculated for DUP4 variant and all of them are in Hardy-Weinberg equilibrium.

**Table 5.14: Table showing the number of individuals observed for each genotype.** The table details the number of unrelated individuals found for each specific genotype.

| Genotype | Observed number of individuals |
|---|---|
| **No DUP4** | 258 |
| **DUP4/WT** | 86 |
| **DUP4/DUP4** | 4 |
| **Total** | 348 |

| | |
|---|---|
| **DUP4 allele frequencies** | DUP4 – ((4x2) + 86)/ 348x2 = 0.135 |

### 5.4.2 Association of DUP4 and malarial phenotypes in the Tanzanian malaria cohort

The quantitative transmission disequilibrium test (QTDT) is a family-based analysis that uses parental information to test the association between phenotypes and alleles at a candidate locus. QTDT version 2.6.1 (Abecasis *et al*., 2000) was used to test the association between the quantitative epidemiological, parasite and serological phenotypes for malaria (Carpenter *et al*., 2009 and 2012) and the DUP4 allele genotype at the glycophorin genes locus. The sex and age of the each individual were included in the analysis as covariates. The QTDT output is a chi-squared value with one degree of freedom for each allele tested and only shows the significant p-values (see Chapter 2 for more details). However, total evidence of association was modelled including age and sex as covariates and environmental and polygenic heritability as additive major locus variance components (Algady *et al*., 2018). The QTDT program was used as it can use all the information in a pedigree to construct powerful tests of association that are robust in the presence of stratification. In addition, the Tanzanian cohort is suitable for QTDT analysis because it is a family-based cohort for which all clinical and non-clinical information is known for each individual. Moreover, because the main goal of testing the DUP4 allele in the Tanzanian cohort was to investigate whether there is any association between the DUP4 genotype and the malaria clinical phenotypes, the QTDT program was suitable as it tests the association between a genotype and a phenotype.

However, the Tanzanian cohort's epidemiological phenotypes were generated from non-selected 'total' population surveys carried out annually in March and April from 1993 to 1999, and the complete annual record of clinical malaria episodes documented in Nyamisati (Tanzania) between 1993 and 1999 by Rooth, who lived in the village since 1985 (Bereczky *et al*., 2007). The phenotypes used in the QTDT analysis are clinical episodes (EPI), parasitemia levels (PARA), haemoglobin levels (HB), corrected haemoglobin levels for age and sex (HBYBY) and corrected haemoglobin levels for age, sex and parasitemia levels (HBPYBY). All this clinical information data is known in the Tanzanian cohort for each individual and was used in previous studies for different purposes (Carpenter *et al*., 2009 and 2012). QTDT version 2.6.1 was used and applied on the 962 Tanzanian samples that have been tested for the DUP4 allele with their malaria phenotypes (Algady *et al*., 2018). The result was significant with the haemoglobin levels at a p-value of 0.0103, the haemoglobin levels corrected for age and sex at a p-value of 0.0078, and the haemoglobin levels corrected for age, sex and parasitemia levels at a p-value of 0.0054. However, the results for the clinical episodes and the parasitemia levels were not significant (Table 5.15).

**Table 5.15: Tests of the association between a glycophorin copy number haplotype (DUP4) and the quantitative malaria phenotypes.** The values presented are p-values for total and within-family tests of association using QTDT. Significant p-values are in bold.

| Testing trait with DUP4 | p-value |
|---|---|
| **EPI** (Clinical episodes) | 0.7184 |
| **PARA** (Parasitemia levels) | 0.3864 |
| **HB** (Uncorrected Haemoglobin levels) | **0.0103** |
| **HBYBY** (Hb levels corrected for age and sex) | **0.0078** |
| **HBPYBY** (Hb levels corrected for age, sex and parasitemia levels) | **0.0054** |

## 5.5 Discussion

The result shows a significant association between DUP4 genotype and the haemoglobin levels in the Tanzanian malaria patients. In general, malaria patients usually have low haemoglobin levels, which explain the anaemia in the malaria patients. The majority of malarial infections are associated with some degree of anaemia that often used also for diagnosis the malaria (Kai and Roberts, 2008), the severity of which depends on patient and parasite-specific characteristics. Malarial anaemia is capable of causing severe morbidity and mortality especially in children and pregnant women infected with *P. falciparum*. However, the malaria patients with the DUP4 allele show a higher level of haemoglobin than normal malaria patients. This may confirm that DUP4 is associated with resistance to severe *P. falciparum* malaria (Leffler *et al*., 2017). Interestingly, the association between DUP4 and haemoglobin levels became more significant and stronger when the haemoglobin levels were corrected for age, sex and parasitemia levels. These covariates are thus very important in the association study, as the p-value ($p \leq 0.05$) (from 0.0103 – 0.0054) was more significant and the association can be confirmed with increased confidence (Algady *et al*., 2018). This result matches with the finding of a recent study that was conducted 244 children with severe malaria they have found significant associations between risk for severe malaria overall and polymorphisms in 15 genes or locations, which most of them are related to red blood cells; they have found also that children with the Dantu blood group antigen (DUP4) have higher haemoglobin levels than other children, who have sever malaria without DUP4, whilst all of them have the same parasitemia levels (Ndila *et al*., 2018).

# Chapter 6: Discussion

This PhD thesis explored the structural variability of the glycophorin copy number variable regions and described the diversity and the characterization of the copy number variations of glycophorin regions across various global populations. It also assessed the linkage disequilibrium between the rs186873296 SNP and *GYPB* CNVs, studied the association of glycophorin CNVs with malaria susceptibility in an infant malaria cohort and explored the association of glycophorin CNVs with malaria phenotypes in a family-based malaria cohort.

## 6.1 Glycophorin variants sizes and breakpoints

Leffler *et al*. (2017) identified 16 different CNVs within the glycophorin regions. These variants were duplications, deletions of entire, or components of three main glycophorin genes, *GYPA*, *GYPB* and *GYPE* as well as complex CNVs. The authors identified glycophorin CNVs using a sequence read depth approach of all the 1000 Genome Project panel individuals. The results of the study revealed eight deletions and eight duplications that were confirmed in two or more individuals. *GYPB* was found to be the gene most affected by the CNVs, resulting in five different forms of deletion.

Glycophorin variants were identified (chapter 3) using high throughput sequencing data of selected 1000 Genomes Project samples. In total 12 variants were detected, analysed and confirmed using fibre-FISH at the Wellcome Sanger Institute, and their breakpoints confirmed using a junction fragment PCR assay, designed according to the breakpoint positions suggested by sequence read depth data. The results showed breakpoints for eight of the twelve glycophorin variants, including four deletions, two duplications, a complex variant (DUP5), and a gene conversion (*GYPE-B-E*). DUP29 was not identified by Leffler *et al*. (2017) and is a novel variant that has a *GYPE-B* hybrid gene. DUP29 is predicted to encodes a normal GPB protein and the duplicated GPB protein was missing the leader peptide. However, exon 1 of the *GYPE* was duplicated, which could cover the missing expressed leader peptide. However, *GYPE* is not expressed and there is no report confirming the existing of the GPE protein, probability because *GYPE* is a pseudogene, therefore, it is unlikely that DUP29 *GYPE-B* hybrid gene to express the duplicated part of the *GYPE* gene.

DUP5 contains two forms of CNVs, one triplication and two duplications; including a *GYPB-E-B* gene conversion. However, from the 1000 Genome project only one sample is a carrier for the DUP5 variant, this sample is from the Western Division in Gambia. Subsequent work by MSc student Eleanor Weyell has used Simons Diversity Project sequence data (Mallick *et*

*al*., 2016) and Gambian Genome Diversity Project sequence data in order to investigate the frequency of the glycophorin variants in these projects; Eleanor has shown that two individuals are DUP5 carriers from the Gambian Genome Diversity Project, while no DUP5 has been found in the Simons Diversity Project. Therefore, it is high likely that DUP5 is mainly found in West African populations.

Leffler *et al*. (2017) has identified all the DUP4 breakpoints and the DEL1 breakpoints. This thesis confirmed the breakpoints of DEL1 reported by Leffler *et al*. (2017). DUP4 is found primarily in East African populations and the DUP4 hybrid gene (*GYPB-A*) express the known Dantu NE+ blood group antigen (Algady *et al*., 2018).

In this thesis a novel gene conversion *GYPE-B-E* has been described and its breakpoints were identified. However, this gene conversion (*GYPE-B-E*) and the DUP5's gene conversion *GYPB-E-B* are not expressing any of the known MNS antigens; they might have their own antigens. In addition, the *GYPE-B* hybrid gene in DUP29 has not previously been reported to be one of the MNS antigens. Because of the majority of the known MNS antigens have different *GYPA* and *GYPB* gene conversion and hybrid genes (Storry *et al*., 2000 and Velliquette *et al*., 2008), while the most of our findings are variations in *GYPB*, *GYPE* and less of *GYPA* and that is why these gene conversions and DUP29 hybrid gene are might be responsible for new (not reported) MNS phenotype. However, *GYPE* is not detected at the red blood cell surface, which makes the prediction of these variant phenotypes are unclear. Moreover, although I could not identify the breakpoints of the glycophorin DUP2 duplication, it is expected from the sequence read depth data and the fibre-FISH result that DUP2 has a *GYPA-B* hybrid gene, which could be one of the MNS phenotypes, GP.Hil (Mi.V), GP.JL (Mi.XI) and Sat.

Apart from the gene conversion and DUP29, no fusion glycophorins were predicted to be generated from the eight variants for which breakpoints were identified. In general, compared to the blood group antigens, many of which were fusion products, a DNA analysis of the detected variants suggested that there was no variant related to the known blood group antigens. However, homozygous deletion of *GYPB* (exons 2 to 6) is responsible for null phenotype RBCs (S-s-U-), as demonstrated in previous studies (Blumenfeld and Huang, 1997; Reid, 2009). Therefore, it is expected that DEL1, 2, and 6 cause the S-s-U- phenotype, as these variants have full deletion of the *GYPB* gene, which means that these have to be homozygote or compound heterozygote.

In addition, the $M^k/M^k$ phenotype is a result of a full deletion of *GYPA* and *GYPB*, which leads to a resistant to malaria infection (Blumenfeld and Huang, 1997). However, of the variants analysed in this investigation, not one of the deletion variants covered either *GYPB* and *GYPA* genes together or *GYPA* alone. Therefore, future investigation could analyse more variants with a deletion of *GYPA* using our PRT assay by changing the test (*GYPA*) and the references (*GYPB* or *GYPE*) in order to expand the study of the relationship between glycophorins and malaria infection and other blood antigens, which could be expressed from the absence of the *GYPA* gene, such as the En(a-) RBCs antigen.

It would be interesting for future work to sequence and analyse the deletions and duplications that were not analysed and studied in this project, such as DEL4 and DUP6, in order to determine the exact breakpoints and design a simple specific assay for each of these variants and use these assays to test the frequencies of these variants in different populations, such as the Benin and the Tanzanian malaria cohorts and try to cover most of the glycophorin variants. Furthermore, it will be useful to designing a specific assay for the glycophorin variants that has complex structures, which could encode modified proteins with affected functions that might distract the RBC invasion process by the *P.falciparum* malaria parasite. These specific assays can be applied on a malaria cohort that expected to have a high frequency of a specific variant and then applied an association test between the glycophorin variant and susceptibility to malaria or malaria phenotypes of the cohort.

## 6.2: Novel genotyping strategies for glycophorin

One of the objectives of this thesis was to characterize and measure copy number variations across the entire human glycophorin gene cluster. In order to do this, a paralogue ratio test (PRT) strategy was designed to measure the diploid copy number across the glycophorin region. The PRT PCR was chosen to type the copy number of the glycophorin regions, because it is able to determine multiallelic copy number variations (mCNVs) and is an accurate and uses a small amount (5 ng) of genomic DNA (Aldhous *et al*., 2010).

The HapMap samples were typed using cis_PRT assays, which target the CNV regions in the glycophorin genes, to validate the PRT assay by comparison with NGS data. The quality of these assays was proven by the common CNV, as shown in the NGS data set detected by the PRT assays. The PRT results showed that the most common glycophorin variants in the HapMap samples were DEL1 and DEL2, which were found in the Yoruban population and DUP2, which was found in the Han Chinese and Northern and Western European populations

from Utah (CEU). PRT assays are robust in accurately using the CNVs of glycophorin to estimate the CNVs of glycophorin in large disease association (case-control) studies.

Therefore, the results showed that cis_PRT1 could be used to detect the most common glycophorin variants, such as DEL1, DEL2, and DUP2, and the cis_PRT2 assay could be used for purposes such as confirming the existence of deletion type 2, differentiating between deletion types 1 and 2, and detecting the gene conversion event that occurred in the *GYPE* (*E-B-E* gene conversion). In addition, the cis_PRT4 assay could be used to detect the gene conversion event *GYPA-B-A*.

The limitation in our PRT approach is that some variants cannot be detected, such as DEL6, DEL7, DUP3, DUP7, and DUP14. However, they can be detected using the amplicon mapping to the *GYPA* repeat as a reference instead of the amplicon mapping to the *GYPE* repeat, because most of the detected variants covered *GYPE* and *GYPB* rather than *GYPB* and *GYPA*, as predicted. This indicates that analysis of PRT data can be modified to detect different known variants. Therefore, these PRT assays can be used in future work to type the glycophorin copy numbers and to detect the frequencies of the glycophorin variants for any other cohort. In general, PRT assays cannot detect small CNVs because it is less likely that the test primers can located on the small deletion or duplication on the target region.

In addition, more accurate CN calling could be obtained if future work allowed the design of more PRT assays with suitable test and reference loci on the glycophorin regions. No any other method has been published or reported for specifically typing the copy number of the glycophorin regions.

My study aimed to design a simple PCR specific for the DUP4 variant using the Leffler *et al*. (2017) breakpoints in order to allow genotyping DUP4 of large cohorts of small DNA amounts. The assay was robust and specific for DUP4. Although, this complex variant is described as a *GYPB-A* hybrid, our assay was designed across another breakpoint between *GYPB* and *GYPE*. In addition, this assay can be used for any other cohort or even a single sample if necessary for genotyping DUP4 and apply any other association analysis in this new cohort, and this assay can be used as clinical genetics test in malaria cases as the doctor can evaluate the situation of the malaria patient.

## 6.3: Glycophorins CN and Malaria.

Many studies have discussed the relationship between variation in the glycophorin genes, which encode receptors for *P.falciparum*, and susceptibility to malaria (Band *et al*., 2015; Leffler *et al*., 2017; Ndila *et al*., 2018). In the current research, *GYPB* was found to be most affected by CNV, resulting in five different forms of deletion, which is consistent with the findings of Leffler *et al*. (2017). In this thesis, the relationship between the glycophorin CNVs and *P.falciparum* malaria infection in the cohort from Benin (infant malaria cohort) was assessed.

The results showed that most of the *GYPB* CNVs (four deletions and three duplications) were deletions (0 copy or 1 copy), and the common deletions are DEL1 and DEL2 variants. However, our PRTs assay has not detected any DUP4 in the Benin malaria cohort, which was unexpected at the beginning of the glycophorin copy number typing because we wanted to typing the copy number of the glycophorin in general rather than focusing in a certain variant.

The association between *GYPB* deletion and the time to first malarial infection and number of malarial infections was also investigated in the Benin cohort. There was no association between the *GYPB* deletion and these phenotypes as expected given Leffler *et al*. (2017) results. In addition, our study confirms the lack of linkage disequilibrium between the *GYPB* deletion variants and the rs186873296 SNP, consistent with previous results (Leffler *et al*., 2017).

The limitations of the Benin cohort study that conducted is that not all of the participants come to the clinic on time, which could affect the accuracy of the data because it is an infant cohort and the time is very crucial because the parasites density and the level of malaria could be affected with time, and some of them not come at all to the clinic for the following up in the middle of the study. These limitations lead to incomplete data for some of the samples (20 samples), since the analysis would be much better if all malaria information available for each DNA samples, because the DNA samples with no clinical information were excluded from the association analysis.

To investigate the relationship between DUP4 and haemoglobin levels in the malaria cohort individuals. It was suitable for this study because the cohort used (Tanzanian malaria cohort) was a family-based, total population cohort from a malaria-endemic region, and the program has demonstrated a proven ability to evaluate the relationship between genotypes and phenotypes in previous studies using the same Tanzanian malaria cohort samples (Carpenter

*et al*., 2009 and 2012). We have suggested that the individuals with the DUP4 allele possessed a higher level of haemoglobin than non-DUP4 carriers. This may confirm that DUP4 is associated with resistance to severe *P. falciparum* malaria as has been suggested by Leffler *et al*. (2017) because the level of haemoglobin in malaria patients is low, which is one of the malaria phenotypes, and when the haemoglobin level become higher than the usual malaria patients haemoglobin level the symptoms will be less severe than normal malaria. It is notable that the association between DUP4 and haemoglobin levels became strengthened when the haemoglobin levels were corrected for age, sex, and parasitemia levels. Thus, these covariates were very important in the study.

The Tanzanian cohort study is limited by that the data are not equally available for all variables at all time points. In addition, the study has another limitation, which is that DNA extractions were performed by different methods for the different years, which potentially could affect the sensitivity of detection and analysis.In the future the Tanzanian cohort can be used for calling the glycophorin copy number using our PRT assay in order to investigate the frequency of other variants in the Tanzanian cohort and could test the association between any other variant in the cohort and haemoglobin levels.

DUP4 is a novel complex structural variant in the glycophorin gene cluster that encodes the Dantu blood group antigen which confers a similar level of protection against severe malaria as HbS, but for which the mechanisms of protection remain unknown. However, an unpublished work for Kariuki *et al*. (2018) has been shown as a preprint article by biorxiv, the study used flow cytometry-based in vitro assays in order to investigate the impact of the DUP4 (Dantu) variant on the invasion RBC by the *P. falciparum* and the protein expression on the RBC membrane, they have used RBC samples from a genotyped cohort of Kenyan children. The study of Kariuki *et al*. (2018) has confirms a dramatic reduction in the ability of several *P. falciparum* strains to invade RBCs from DUP4 (Dantu) heterozygote and homozygotes respectively. In addition, they have seen a significant decline in the surface expression of GPA and GPB, while GPC showed a significant increase in expression on the surface of RBCs in the same cells. These observations that have been reported in the Kariuki *et al*. (2018) abstract regarding the reduction of invasion in carriers of the DUP4 variant allele suggests that this variant is high likely to give protection against malaria infection by significantly reducing parasite invasion into the RBC, perhaps due to alterations of the glycophorin receptors expression on the RBC membrane.

## 6.4: Future work

Understanding the LD between DUP4 and the rs186873296 SNP leads to find associations that may help to impute the variant through tagSNP. Therefore, genotyping of the SNP in the Tanzanian malaria cohort will help to confirm the linkage disequilibrium strength between the DUP4 glycophorin variant and the SNP that has been reported by Leffler *et al*. (2017) and Ndila *et al*. (2018). Sequencing of the Benin and Tanzanian cohorts will be recommended by a high-throughput sequencing technology, such as Illumina HiSeq. That will be useful to investigate any new glycophorin variant then identify its breakpoints and test the frequency of it in the cohort and test the association of this variant with the susceptibility to malaria and malaria phenotypes.

Moreover, In the light of a preprint article for Kariuki *et al*. (2018), which has confirmed that there is a significant decline in the surface expression of GPA and GPB, while GPC showed a significant increase in expression on the surface of RBCs in the same cells, DUP2 is expected to have a *GYPB-A* hybrid gene and this hybrid gene and this could affect the *GYPA* and *GYPB* expression, therefore, it would be recommended to find out the DUP2 breakpoints and try to predict the DUP2 possible protein, also use the flow cytometry-based in vitro assays in order to investigate the impact of the DUP2. Also, typing DUP2 in the Tanzanian cohort will be helpful to investigate if there is any association of the DUP2 glycophorin variant with the malaria phenotypes after genotyping the DUP2 in this family-based malaria cohort.

Malaria is a good example of an agent of natural selection in humans. Haemoglobin S (HbS) is the most common type of abnormal haemoglobin and the basis of sickle cell trait and sickle cell anaemia. HbS differs from normal haemoglobin (HbA) only by a single nucleotide mutation (A>T) of the *β-globin* gene, which replacing a valine amino acid by a glutamine in the 6th position of the beta chain of globin (Anstee, 2010). Sickle cell anaemia (SCD) is caused by homozygosity and is deleterious, and that it is at high frequency because of the heterozygous advantage of HbS allele against malaria. This demonstrates the strength of selection driven by malaria (Penman *et al*., 2009).

Human erythrocytes can evolve to eliminate or change their protein structure in order to avoid parasite invasion through these receptors. Malaria endemic regions have a considerable level of polymorphisms in *GYPA* and *GYPB* because it is a receptor for *P.falciparum* (Baldwin *et al*., 2015). Also, Miltenberger variants often cause RBCs to lose GPB in malaria endemic regions. For instance, Efe pygmies of the Ituri forest in Congo show 59% for the

gene frequency of S-s-U- (deletion of *GYPB* gene) (Mayer *et al*., 2009). Although, S-s-U- has been shown that it is not protective against malaria, the Efe example is maybe due to a positive selection of the deletion of *GYPB* gene.

I have shown that the DUP4 allele frequency in Nyamasati in Tanzania (East African country) result that we have investigated (13.5%), which is much higher than the DUP4 allele frequency that have been reported by Leffler *et al*. (2017) in different African populations from Gambia (0.0%), Malawi (3.9%) and Kenya (9.0%) and different populations from Tanzania (not Nyamasati) (Figure 6.1).

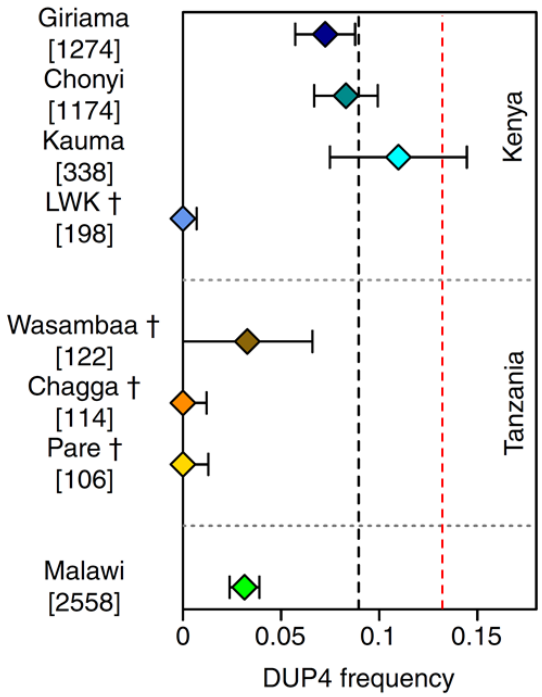



**Figure 6.1: Figure that shows the allele frequency of DUP4 in different east African populations.** The black dotted vertical line indicates the allele frequency of DUP4 in Kenyan controls in general. The red dotted vertical line indicates to the allele frequency of DUP4 in Nyamasati Tanzanian population. The figure modified from Leffler *et al*. (2017).

That indicates that DUP4 is more common in east African populations than the West African populations as Gambia shows no DUP4 alleles in both controls and cases in the Leffler *et al*. (2017) study. However, genotyping the DUP4 in the Benin malaria cohort (already available in the lab) will be interesting to confirm that DUP4 is more prevalent in east African countries that the west African countries.

Moreover, Leffler *et al*. (2017) has used a haplotype-based technique (after excluding other glycophorin variants within the region) in order estimate the relative age of the DUP4 allele

and screening loci for signals of recent positive selection by computing outward the extended haplotype homozygosity (EHH) from the glycophorin region for DUP4 haplotypes and non-DUP4 haplotypes in Kenya, which gives proof that DUP4 is carried on an extended haplotype that may have raised to its current frequency in Kenya relatively recently. Because that the positive selection leads to a rapid rise in frequency of a variant in a relatively few generations will preserve the original haplotype structure (core haplotype), since the number of recombination events would be limited. Therefore, the extended haplotype homozygosity can be computed outward from the glycophorin region for DUP4 haplotypes and non-DUP4 haplotypes in our family-based Tanzanian cohort after having the phased variant data and identify core haplotypes of the cohort using the International HapMap Project recombination rate for example.

In addition, genotyping any other east African malaria cohort, such as Mozambique, Democratic Republic of the Congo and Uganda, which has high prevalence of *P.falciparum* parasite (Bhatt *et al*., 2015) and investigate the allele frequency of the variant and compare it with other allele frequencies with other samples in order to investigate if there is any evidence supporting this positive selection. Therefore, mapping the allele frequency of DUP4 across additional populations could help clarify the nature of selection. However, many techniques for detecting positive selection rely on accurate phasing (Sabeti *et al*., 2007). Identification of the haplotypic background of a given variant, will also help to distinguish single versus recurrent deletion/amplification events and will also give an idea about the age of the variant (from the size of the extended haplotype containing the CNV).

# Bibliography

'A comprehensive genetic linkage map of the human genome. NIH/CEPH Collaborative Mapping Group', (1992) *Science (New York, N.Y.),* 258(5079), pp. 148-162.

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010) 'A map of human genome variation from population-scale sequencing', *Nature,* 467(7319), pp. 1061-1073.

Abecasis, G.R., Cardon, L.R. and Cookson, W.O. (2000) 'A general test of association for quantitative traits in nuclear families', *American Journal of Human Genetics,* 66(1), pp. 279-292.

Abu Bakar, S., Hollox, E.J. and Armour, J.A. (2009) 'Allelic recombination between distinct genomic locations generates copy number diversity in human beta-defensins', *Proceedings of the National Academy of Sciences of the United States of America,* 106(3), pp. 853-858.

Ahn, J.W., Bint, S., Bergbaum, A., Mann, K., Hall, R.P. and Ogilvie, C.M. (2013) 'Array CGH as a first line diagnostic test in place of karyotyping for postnatal referrals - results from four years' clinical application for over 8,700 patients', *Molecular cytogenetics,* 6(1), pp. 16-8166-6-16.

Ahn, K., Gotay, N., Andersen, T.M., Anvari, A.A., Gochman, P., Lee, Y., Sanders, S., Guha, S., Darvasi, A., Glessner, J.T., Hakonarson, H., Lencz, T., State, M.W., Shugart, Y.Y. and Rapoport, J.L. (2014) 'High rate of disease-related copy number variations in childhood onset schizophrenia', *Molecular psychiatry,* 19(5), pp. 568-572.

Aigner, J., Villatoro, S., Rabionet, R., Roquer, J., Jimenez-Conde, J., Marti, E. and Estivill, X. (2013) 'A common 56-kilobase deletion in a primate-specific segmental duplication creates a novel butyrophilin-like protein', *BMC genetics,* 14, pp. 61-2156-14-61.

Aklillu, E., Odenthal-Hesse, L., Bowdrey, J., Habtewold, A., Ngaimisi, E., Yimer, G., Amogne, W., Mugusi, S., Minzi, O., Makonnen, E., Janabi, M., Mugusi, F., Aderaye, G., Hardwick, R., Fu, B., Viskaduraki, M., Yang, F. and Hollox, E.J. (2013) 'CCL3L1 copy number, HIV load, and immune reconstitution in sub-Saharan Africans', *BMC infectious diseases,* 13, pp. 536-2334-13-536.

Aldhous, M.C., Abu Bakar, S., Prescott, N.J., Palla, R., Soo, K., Mansfield, J.C., Mathew, C.G., Satsangi, J. and Armour, J.A. (2010) 'Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease', *Human molecular genetics,* 19(24), pp. 4930-4938.

Algady, W., Louzada, S., Carpenter, D., Brajer, P., Farnert, A., Rooth, I., Ngasala, B., Yang, F., Shaw, M.A. and Hollox, E.J. (2018) 'The Malaria-Protective Human Glycophorin Structural Variant DUP4 Shows Somatic Mosaicism and Association with Hemoglobin Levels', *American Journal of Human Genetics,* 103(5), pp. 769-776.

Anstee, D.J. (2010) 'The relationship between blood groups and disease', *Blood,* 115(23), pp. 4635-4643.

Aoki, T. (2017) 'A Comprehensive Review of Our Current Understanding of Red Blood Cell (RBC) Glycoproteins', *Membranes,* 7(4), pp. 10.3390/membranes7040056.

Arlt, M.F., Wilson, T.E. and Glover, T.W. (2012) 'Replication stress and mechanisms of CNV formation', *Current opinion in genetics & development,* 22(3), pp. 204-210.

Armour, J.A., Palla, R., Zeeuwen, P.L., den Heijer, M., Schalkwijk, J. and Hollox, E.J. (2007) 'Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats', *Nucleic acids research,* 35(3), pp. e19.

Armour, J.A., Sismani, C., Patsalis, P.C. and Cross, G. (2000) 'Measurement of locus copy number by hybridisation with amplifiable probes', *Nucleic acids research,* 28(2), pp. 605-609.

Avent, N.D. and Reid, M.E. (2000) 'The Rh blood group system: a review', *Blood,* 95(2), pp. 375-387.

Baldwin, M.R., Li, X., Hanada, T., Liu, S.C. and Chishti, A.H. (2015) 'Merozoite surface protein 1 recognition of host glycophorin A mediates malaria parasite invasion of red blood cells', *Blood,* 125(17), pp. 2704-2711.

Ballantyne, K.N., van Oorschot, R.A. and Mitchell, R.J. (2008) 'Locked nucleic acids in PCR primers increase sensitivity and performance', *Genomics,* 91(3), pp. 301-305.

Band, G., Le, Q.S., Jostins, L., Pirinen, M., Kivinen, K., Jallow, M., Sisay-Joof, F., Bojang, K., Pinder, M., Sirugo, G., Conway, D.J., Nyirongo, V., Kachala, D., Molyneux, M., Taylor, T., Ndila, C., Peshu, N., Marsh, K., Williams, T.N., Alcock, D., Andrews, R., Edkins, S., Gray, E., Hubbart, C., Jeffreys, A., Rowlands, K., Schuldt, K., Clark, T.G., Small, K.S., Teo, Y.Y., Kwiatkowski, D.P., Rockett, K.A., Barrett, J.C., Spencer, C.C., Malaria Genomic Epidemiology Network and Malaria Genomic Epidemiological Network (2013) 'Imputation-based meta-analysis of severe malaria in three African populations', *PLoS genetics,* 9(5), pp. e1003509.

Baris, H.N., Tan, W.H., Kimonis, V.E. and Irons, M.B. (2007) 'Diagnostic utility of array-based comparative genomic hybridization in a clinical setting', *American journal of medical genetics.Part A,* 143A(21), pp. 2523-2533.

Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D. and Hurles, M.E. (2008) 'A robust statistical method for case-control association testing with copy number variation', *Nature genetics*, 40(10), pp. 1245-1252.

Bassuk, A.G., Geraghty, E., Wu, S., Mullen, S.A., Berkovic, S.F., Scheffer, I.E. and Mefford, H.C. (2013) 'Deletions of 16p11.2 and 19p13.2 in a family with intellectual disability and generalized epilepsy', *American journal of medical genetics.Part A*, 161A(7), pp. 1722-1725.

Batstone-Cunningham, R.L., Hardy, R.E., Daman, M.E. and Dill, K. (1983) 'Possible role of the carbohydrate residues in the display of the MN blood group determinants by glycophorin A', *Biochimica et biophysica acta*, 746(1-2), pp. 1-7.

Baum, J., Ward, R.H. and Conway, D.J. (2002) 'Natural selection on the erythrocyte surface', *Molecular biology and evolution*, 19(3), pp. 223-229.

Bensimon, A., Simon, A., Chiffaudel, A., Croquette, V., Heslot, F. and Bensimon, D. (1994) 'Alignment and sensitive detection of DNA by a moving interface', *Science (New York, N.Y.)*, 265(5181), pp. 2096-2098.

Bereczky, S., Liljander, A., Rooth, I., Faraja, L., Granath, F., Montgomery, S.M. and Farnert, A. (2007) 'Multiclonal asymptomatic Plasmodium falciparum infections predict a reduced risk of malaria disease in a Tanzanian population', *Microbes and Infection*, 9(1), pp. 103-110.

Bhardwaj, U., Zhang, Y.H. and McCabe, E.R. (2003) 'Neonatal hemoglobinopathy screening: molecular genetic technologies', *Molecular genetics and metabolism*, 80(1-2), pp. 129-137.

Bhatt, S., Weiss, D.J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C.L., Henry, A., Eckhoff, P.A., Wenger, E.A., Briet, O., Penny, M.A., Smith, T.A., Bennett, A., Yukich, J., Eisele, T.P., Griffin, J.T., Fergus, C.A., Lynch, M., Lindgren, F., Cohen, J.M., Murray, C.L.J., Smith, D.L., Hay, S.I., Cibulskis, R.E. and Gething, P.W. (2015) 'The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015', *Nature*, 526(7572), pp. 207-211.

Blumenfeld, O.O. and Huang, C.H. (1997) 'Molecular genetics of glycophorin MNS variants', *Transfusion clinique et biologique : journal de la Societe francaise de transfusion sanguine*, 4(4), pp. 357-365.

Blumenfeld, O.O. and Huang, C.H. (1995) 'Molecular genetics of the glycophorin gene family, the antigens for MNSs blood groups: multiple gene rearrangements and modulation of splice site usage result in extensive diversification', *Human mutation*, 6(3), pp. 199-209.

Blumenfeld, O.O., Smith, A.J. and Moulds, J.J. (1987) 'Membrane glycophorins of Dantu blood group erythrocytes', *The Journal of biological chemistry*, 262(24), pp. 11864-11870.

Broad Institute (2015) 'genome STRiP overview', [online] Available from: http://www.broadinstitute.org/software/genomestrip/ (Accessed 14 September 2018).

Burness, A.T. and Pardoe, I.U. (1983) 'A sialoglycopeptide from human erythrocytes with receptor-like properties for encephalomyocarditis and influenza viruses', *The Journal of general virology*, 64(Pt 5), pp. 1137-1148.

Cahan, P., Li, Y., Izumi, M. and Graubert, T.A. (2009) 'The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells', *Nature genetics*, 41(4), pp. 430-437.

Campbell, C.D., Sampas, N., Tsalenko, A., Sudmant, P.H., Kidd, J.M., Malig, M., Vu, T.H., Vives, L., Tsang, P., Bruhn, L. and Eichler, E.E. (2011) 'Population-genetic properties of differentiated human copy-number polymorphisms', *American Journal of Human Genetics*, 88(3), pp. 317-332.

Cantsilieris, S., Baird, P.N. and White, S.J. (2013) 'Molecular methods for genotyping complex copy number polymorphisms', *Genomics*, 101(2), pp. 86-93.

Cantsilieris, S. and White, S.J. (2013) 'Correlating multiallelic copy number polymorphisms with disease susceptibility', *Human mutation*, 34(1), pp. 1-13.

Cantsilieris, S., White, S.J., Richardson, A.J., Guymer, R.H. and Baird, P.N. (2012) 'Comprehensive analysis of Copy Number Variation of genes at chromosome 1 and 10 loci associated with late age related macular degeneration', *PloS one*, 7(4), pp. e35255.

Carpenter, D., Abushama, H., Bereczky, S., Farnert, A., Rooth, I., Troye-Blomberg, M., Quinnell, R.J. and Shaw, M.A. (2007) 'Immunogenetic control of antibody responsiveness in a malaria endemic area', *Human immunology*, 68(3), pp. 165-169.

Carpenter, D., Farnert, A., Rooth, I., Armour, J.A. and Shaw, M.A. (2012) 'CCL3L1 copy number and susceptibility to malaria', *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 12(5), pp. 1147-1154.

Carpenter, D., Rooth, I., Farnert, A., Abushama, H., Quinnell, R.J. and Shaw, M.A. (2009) 'Genetics of susceptibility to malaria related phenotypes', *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 9(1), pp. 97-103.

Carpenter, D., Walker, S., Prescott, N., Schalkwijk, J. and Armour, J.A. (2011) 'Accuracy and differential bias in copy number measurement of CCL3L1 in association studies with three auto-immune disorders', *BMC genomics*, 12, pp. 418-2164-12-418.

Cartron, J.P., Le Van Kim, C. and Colin, Y. (1993) 'Glycophorin C and related glycoproteins: structure, function, and regulation', *Seminars in hematology*, 30(2), pp. 152-168.

Cartron, J.P. and Rahuel, C. (1992) 'Human erythrocyte glycophorins: protein and gene structure analyses', *Transfusion medicine reviews*, 6(2), pp. 63-92.

Cartron, J.P., Rouger P. and Eds. Molecular Basis of Human Blood Group Antigens (Blood Cell Biochemistry Series, Springer, 1995), vol. 6.

Carvalho, C.M. and Lupski, J.R. (2016) 'Mechanisms underlying structural variant formation in genomic disorders', *Nature reviews.Genetics*, 17(4), pp. 224-238.

Cavalli-Sforza, L.L. (2005) 'The Human Genome Diversity Project: past, present and future', *Nature reviews.Genetics*, 6(4), pp. 333-340.

Chasis, J.A., Reid, M.E., Jensen, R.H. and Mohandas, N. (1988) 'Signal transduction by glycophorin A: role of extracellular and cytoplasmic domains in a modulatable process', *The Journal of cell biology*, 107(4), pp. 1351-1357.

Chen, J.M., Cooper, D.N., Chuzhanova, N., Ferec, C. and Patrinos, G.P. (2007) 'Gene conversion: mechanisms, evolution and human disease', *Nature reviews.Genetics*, 8(10), pp. 762-775.

Chen, Q., Book, M., Fang, X., Hoeft, A. and Stuber, F. (2006) 'Screening of copy number polymorphisms in human beta-defensin genes using modified real-time quantitative PCR', *Journal of immunological methods*, 308(1-2), pp. 231-240.

Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.C., Agarwala, R., McLaren, W.M., Ritchie, G.R., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., Eichler, E.E., Weinstock, G., Mardis, E.R., Wilson, R.K., Howe, K., Flicek, P. and Hubbard, T. (2011) 'Modernizing reference genome assemblies', *PLoS biology*, 9(7), pp. e1001091.

Conrad, D.F., Bird, C., Blackburne, B., Lindsay, S., Mamanova, L., Lee, C., Turner, D.J. and Hurles, M.E. (2010) 'Mutation spectrum revealed by breakpoint sequencing of human germline CNVs', *Nature genetics*, 42(5), pp. 385-391.

Conrad, D.F. and Hurles, M.E. (2007) 'The population genetics of structural variation', *Nature genetics,* 39(7 Suppl), pp. S30-6.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C.H., Kristiansson, K., Macarthur, D.G., Macdonald, J.R., Onyiah, I., Pang, A.W., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Wellcome Trust Case Control Consortium, Tyler-Smith, C., Carter, N.P., Lee, C., Scherer, S.W. and Hurles, M.E. (2010) 'Origins and functional impact of copy number variation in the human genome', *Nature,* 464(7289), pp. 704-712.

Dahr, W. (1992) 'Miltenberger subsystem of the MNSs blood group system. Review and outlook', *Vox sanguinis,* 62(3), pp. 129-135.

Denomme, G.A. (2004) 'The structure and function of the molecules that carry human red blood cell and platelet antigens', *Transfusion medicine reviews,* 18(3), pp. 203-231.

Deutsch, S., Choudhury, U., Merla, G., Howald, C., Sylvan, A. and Antonarakis, S.E. (2004) 'Detection of aneuploidies by paralogous sequence quantification', *Journal of medical genetics,* 41(12), pp. 908-915.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C. and Gibbs, R.A. (2012) 'Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology', *PloS one,* 7(11), pp. e47768.

Ersfeld, K. (2004) 'Fiber-FISH: fluorescence in situ hybridization on stretched DNA', *Methods in molecular biology (Clifton, N.J.),* 270, pp. 395-402.

Everitt, B. (2004) 'Cluster analysis is a generic term for a wide range of numerical methods for examining data', *Statistical methods in medical research,* 13(5), pp. 343-345.

Farnert, A., Yman, V., Homann, M.V., Wandell, G., Mhoja, L., Johansson, M., Jesaja, S., Sandlund, J., Tanabe, K., Hammar, U., Bottai, M., Premji, Z.G., Bjorkman, A. and Rooth, I. (2014) 'Epidemiology of malaria in a village in the Rufiji River Delta, Tanzania: declining transmission over 25 years revealed by different parasitological metrics', *Malaria journal,* 13, pp. 459-2875-13-459.

Fellermann, K., Stange, D.E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C.L., Reinisch, W., Teml, A., Schwab, M., Lichter, P., Radlwimmer, B. and Stange, E.F. (2006) 'A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon', *American Journal of Human Genetics,* 79(3), pp. 439-448.

Fernando, M.M., Boteva, L., Morris, D.L., Zhou, B., Wu, Y.L., Lokki, M.L., Yu, C.Y., Rioux, J.D., Hollox, E.J. and Vyse, T.J. (2010) 'Assessment of complement C4 gene copy number using the paralog ratio test', *Human mutation,* 31(7), pp. 866-874.

Feuk, L., Carson, A.R. and Scherer, S.W. (2006) 'Structural variation in the human genome', *Nature reviews.Genetics,* 7(2), pp. 85-97.

Field, S.P., Hempelmann, E., Mendelow, B.V. and Fleming, A.F. (1994) 'Glycophorin variants and Plasmodium falciparum: protective effect of the Dantu phenotype in vitro', *Human genetics,* 93(2), pp. 148-150.

Flores, M., Morales, L., Gonzaga-Jauregui, C., Dominguez-Vidana, R., Zepeda, C., Yanez, O., Gutierrez, M., Lemus, T., Valle, D., Avila, M.C., Blanco, D., Medina-Ruiz, S., Meza, K., Ayala, E., Garcia, D., Bustos, P., Gonzalez, V., Girard, L., Tusie-Luna, T., Davila, G. and Palacios, R. (2007) 'Recurrent DNA inversion rearrangements in the human genome', *Proceedings of the National Academy of Sciences of the United States of America,* 104(15), pp. 6099-6106.

Fode, P., Jespersgaard, C., Hardwick, R.J., Bogle, H., Theisen, M., Dodoo, D., Lenicek, M., Vitek, L., Vieira, A., Freitas, J., Andersen, P.S. and Hollox, E.J. (2011) 'Determination of beta-defensin genomic copy number in different populations: a comparison of three methods', *PloS one,* 6(2), pp. e16768.

Frank, B., Bermejo, J.L., Hemminki, K., Sutter, C., Wappenschmidt, B., Meindl, A., Kiechle-Bahat, M., Bugert, P., Schmutzler, R.K., Bartram, C.R. and Burwinkel, B. (2007) 'Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk', *Carcinogenesis,* 28(7), pp. 1442-1445.

Frazer, K.A., Murray, S.S., Schork, N.J. and Topol, E.J. (2009) 'Human genetic variation and its contribution to complex traits', *Nature reviews.Genetics,* 10(4), pp. 241-251.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., Carter, N.P., Scherer, S.W. and Lee, C. (2006) 'Copy number variation: new insights in genome diversity', *Genome research,* 16(8), pp. 949-961.

Gething, P.W., Patil, A.P., Smith, D.L., Guerra, C.A., Elyazar, I.R., Johnston, G.L., Tatem, A.J. and Hay, S.I. (2011) 'A new world malaria map: Plasmodium falciparum endemicity in 2010', *Malaria journal,* 10, pp. 378-2875-10-378.

Ghanem, N., Uring-Lambert, B., Abbal, M., Hauptmann, G., Lefranc, M.P. and Lefranc, G. (1988) 'Polymorphism of MHC class III genes: definition of restriction fragment linkage groups and evidence for frequent deletions and duplications', *Human genetics,* 79(3), pp. 209-218.

Gillet-Markowska, A., Richard, H., Fischer, G. and Lafontaine, I. (2015) 'Ulysses: accurate detection of low-frequency structural variations in large insert-size sequencing libraries', *Bioinformatics (Oxford, England),* 31(6), pp. 801-808.

Girirajan, S., Campbell, C.D. and Eichler, E.E. (2011) 'Human copy number variation and complex genetic disease', *Annual Review of Genetics,* 45, pp. 203-226.

Goossens, M., Dozy, A.M., Embury, S.H., Zachariades, Z., Hadjiminas, M.G., Stamatoyannopoulos, G. and Kan, Y.W. (1980) 'Triplicated alpha-globin loci in humans', *Proceedings of the National Academy of Sciences of the United States of America,* 77(1), pp. 518-521.

Gribble, S.M., Wiseman, F.K., Clayton, S., Prigmore, E., Langley, E., Yang, F., Maguire, S., Fu, B., Rajan, D., Sheppard, O., Scott, C., Hauser, H., Stephens, P.J., Stebbings, L.A., Ng, B.L., Fitzgerald, T., Quail, M.A., Banerjee, R., Rothkamm, K., Tybulewicz, V.L., Fisher, E.M. and Carter, N.P. (2013) 'Massively parallel sequencing reveals the complex structure of an irradiated human chromosome on a mouse background in the Tc1 model of Down syndrome', *PloS one,* 8(4), pp. e60482.

Griffin, J.T., Ferguson, N.M. and Ghani, A.C. (2014) 'Estimates of the changing age-burden of Plasmodium falciparum malaria disease in sub-Saharan Africa', *Nature communications,* 5, pp. 3136.

Groth, M., Szafranski, K., Taudien, S., Huse, K., Mueller, O., Rosenstiel, P., Nygren, A.O., Schreiber, S., Birkenmeier, G. and Platzer, M. (2008) 'High-resolution mapping of the 8p23.1 beta-defensin cluster reveals strictly concordant copy number variation of all genes', *Human mutation,* 29(10), pp. 1247-1254.

Handsaker, R.E., Korn, J.M., Nemesh, J. and McCarroll, S.A. (2011) 'Discovery and genotyping of genome structural polymorphism by sequencing on a population scale', *Nature genetics,* 43(3), pp. 269-276.

Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., Kashin, S., Boettger, L.M. and McCarroll, S.A. (2015) 'Large multiallelic copy number variations in humans', *Nature genetics,* 47(3), pp. 296-303.

Hardwick, R.J., Amogne, W., Mugusi, S., Yimer, G., Ngaimisi, E., Habtewold, A., Minzi, O., Makonnen, E., Janabi, M., Machado, L.R., Viskaduraki, M., Mugusi, F., Aderaye, G., Lindquist, L., Hollox, E.J. and Aklillu, E. (2012) 'beta-defensin genomic copy number is associated with HIV load and immune reconstitution in sub-saharan Africans', *The Journal of infectious diseases,* 206(7), pp. 1012-1019.

Hardwick, R.J., Machado, L.R., Zuccherato, L.W., Antolinos, S., Xue, Y., Shawa, N., Gilman, R.H., Cabrera, L., Berg, D.E., Tyler-Smith, C., Kelly, P., Tarazona-Santos, E. and Hollox, E.J. (2011) 'A worldwide analysis of beta-defensin copy number variation

suggests recent selection of a high-expressing DEFB103 gene copy in East Asia', *Human mutation,* 32(7), pp. 743-750.

Hardwick, R.J., Menard, A., Sironi, M., Milet, J., Garcia, A., Sese, C., Yang, F., Fu, B., Courtin, D. and Hollox, E.J. (2014) 'Haptoglobin (HP) and Haptoglobin-related protein (HPR) copy number variation, natural selection, and trypanosomiasis', *Human genetics,* 133(1), pp. 69-83.

Hastings, P.J., Ira, G. and Lupski, J.R. (2009) 'A microhomology-mediated break-induced replication model for the origin of human copy number variation', *PLoS genetics,* 5(1), pp. e1000327.

Henrichsen, C.N., Vinckenbosch, N., Zollner, S., Chaignat, E., Pradervand, S., Schutz, F., Ruedi, M., Kaessmann, H. and Reymond, A. (2009) 'Segmental copy number variation shapes tissue transcriptomes', *Nature genetics,* 41(4), pp. 424-429.

Higgs, D.R., Vickers, M.A., Wilkie, A.O., Pretorius, I.M., Jarman, A.P. and Weatherall, D.J. (1989) 'A review of the molecular genetics of the human alpha-globin gene cluster', *Blood,* 73(5), pp. 1081-1104.

Hills, A., Ahn, J.W., Donaghue, C., Thomas, H., Mann, K. and Ogilvie, C.M. (2010) 'MLPA for confirmation of array CGH results and determination of inheritance', *Molecular cytogenetics,* 3, pp. 19-8166-3-19.

Hoher, G., Fiegenbaum, M. and Almeida, S. (2018) 'Molecular basis of the Duffy blood group system', *Blood transfusion = Trasfusione del sangue,* 16(1), pp. 93-100.

Hollox, E.J. (2012) 'The challenges of studying complex and dynamic regions of the human genome', *Methods in molecular biology (Clifton, N.J.),* 838, pp. 187-207.

Hollox, E.J. (2010) 'Beta-defensins and Crohn's disease: confusion from counting copies', *The American Journal of Gastroenterology,* 105(2), pp. 360-362.

Hollox, E.J. (2008) 'Copy number variation of beta-defensins and relevance to disease', *Cytogenetic and genome research,* 123(1-4), pp. 148-155.

Hollox, E.J., Akrami, S.M. and Armour, J.A. (2002) 'DNA copy number analysis by MAPH: molecular diagnostic applications', *Expert review of molecular diagnostics,* 2(4), pp. 370-378.

Hollox, E.J., Armour, J.A. and Barber, J.C. (2003) 'Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster', *American Journal of Human Genetics,* 73(3), pp. 591-600.

Hollox, E.J., Barber, J.C., Brookes, A.J. and Armour, J.A. (2008) 'Defensins and the dynamic genome: what we can learn from structural variation at human chromosome band 8p23.1', *Genome research,* 18(11), pp. 1686-1697.

Hollox, E.J., Davies, J., Griesenbach, U., Burgess, J., Alton, E.W. and Armour, J.A. (2005) 'Beta-defensin genomic copy number is not a modifier locus for cystic fibrosis', *Journal of negative results in biomedicine,* 4, pp. 9-5751-4-9.

Hollox, E.J. and Hoh, B.P. (2014) 'Human gene copy number variation and infectious disease', *Human genetics,* 133(10), pp. 1217-1233.

Hollox, E.J., Huffmeier, U., Zeeuwen, P.L., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P.C., Traupe, H., de Jongh, G., den Heijer, M., Reis, A., Armour, J.A. and Schalkwijk, J. (2008) 'Psoriasis is associated with increased beta-defensin genomic copy number', *Nature genetics,* 40(1), pp. 23-25.

Hsu, K., Chi, N., Gucek, M., Van Eyk, J.E., Cole, R.N., Lin, M. and Foster, D.B. (2009) 'Miltenberger blood group antigen type III (Mi.III) enhances the expression of band 3', *Blood,* 114(9), pp. 1919-1928.

Huang, C.H. and Blumenfeld, O.O. (1991) 'Multiple origins of the human glycophorin Sta gene. Identification of hot spots for independent unequal homologous recombinations', *The Journal of biological chemistry,* 266(34), pp. 23306-23314.

Huang, C.H. and Blumenfeld, O.O. (1988) 'Characterization of a genomic hybrid specifying the human erythrocyte antigen Dantu: Dantu gene is duplicated and linked to a delta glycophorin gene deletion', *Proceedings of the National Academy of Sciences of the United States of America,* 85(24), pp. 9640-9644.

Huang, C.H., Reid, M.E., Xie, S.S. and Blumenfeld, O.O. (1996) 'Human red blood cell Wright antigens: a genetic and evolutionary perspective on glycophorin A-band 3 interaction', *Blood,* 87(9), pp. 3942-3947.

Huang, T., Zhuge, J. and Zhang, W.W. (2013) 'Sensitive detection of BRAF V600E mutation by Amplification Refractory Mutation System (ARMS)-PCR', *Biomarker research,* 1(1), pp. 3-7771-1-3.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) 'Detection of large-scale variation in the human genome', *Nature genetics,* 36(9), pp. 949-951.

Ihaka, R. and Gentleman, R. (1996) 'R: A Language for Data Analysis and Graphics', *Journal of Computational and Graphical Statistics,* 5(3), pp. 299-314.

Inoue, K. and Lupski, J.R. (2002) 'Molecular mechanisms for genomic disorders', *Annual review of genomics and human genetics*, 3, pp. 199-242.

International HapMap Consortium (2003) 'The International HapMap Project', *Nature*, 426(6968), pp. 789-796.

International Human Genome Sequencing Consortium (2004) 'Finishing the euchromatic sequence of the human genome', *Nature*, 431(7011), pp. 931-945.

Ira, G., Malkova, A., Liberi, G., Foiani, M. and Haber, J.E. (2003) 'Srs2 and Sgs1-Top3 suppress crossovers during double-strand break repair in yeast', *Cell*, 115(4), pp. 401-411.

Jaillard, S., Drunat, S., Bendavid, C., Aboura, A., Etcheverry, A., Journel, H., Delahaye, A., Pasquier, L., Bonneau, D., Toutain, A., Burglen, L., Guichet, A., Pipiras, E., Gilbert-Dussardier, B., Benzacken, B., Martin-Coignard, D., Henry, C., David, A., Lucas, J., Mosser, J., David, V., Odent, S., Verloes, A. and Dubourg, C. (2010) 'Identification of gene copy number variations in patients with mental retardation using array-CGH: Novel syndromes in a large French series', *European journal of medical genetics*, 53(2), pp. 66-75.

Jallow, M., Teo, Y.Y., Small, K.S., Rockett, K.A., Deloukas, P., Clark, T.G., Kivinen, K., Bojang, K.A., Conway, D.J., Pinder, M., Sirugo, G., Sisay-Joof, F., Usen, S., Auburn, S., Bumpstead, S.J., Campino, S., Coffey, A., Dunham, A., Fry, A.E., Green, A., Gwilliam, R., Hunt, S.E., Inouye, M., Jeffreys, A.E., Mendy, A., Palotie, A., Potter, S., Ragoussis, J., Rogers, J., Rowlands, K., Somaskantharajah, E., Whittaker, P., Widden, C., Donnelly, P., Howie, B., Marchini, J., Morris, A., SanJoaquin, M., Achidi, E.A., Agbenyega, T., Allen, A., Amodu, O., Corran, P., Djimde, A., Dolo, A., Doumbo, O.K., Drakeley, C., Dunstan, S., Evans, J., Farrar, J., Fernando, D., Hien, T.T., Horstmann, R.D., Ibrahim, M., Karunaweera, N., Kokwaro, G., Koram, K.A., Lemnge, M., Makani, J., Marsh, K., Michon, P., Modiano, D., Molyneux, M.E., Mueller, I., Parker, M., Peshu, N., Plowe, C.V., Puijalon, O., Reeder, J., Reyburn, H., Riley, E.M., Sakuntabhai, A., Singhasivanon, P., Sirima, S., Tall, A., Taylor, T.E., Thera, M., Troye-Blomberg, M., Williams, T.N., Wilson, M., Kwiatkowski, D.P., Wellcome Trust Case Control Consortium and Malaria Genomic Epidemiology Network (2009) 'Genome-wide and fine-resolution association analysis of malaria in West Africa', *Nature genetics*, 41(6), pp. 657-665.

James, C.P., Bajaj-Elliott, M., Abujaber, R., Forya, F., Klein, N., David, A.L., Hollox, E.J. and Peebles, D.M. (2018) 'Human beta defensin (HBD) gene copy number affects HBD2 protein levels: impact on cervical bactericidal immunity in pregnancy', *European journal of human genetics : EJHG*, 26(3), pp. 434-439.

Jansen, P.A., Rodijk-Olthuis, D., Hollox, E.J., Kamsteeg, M., Tjabringa, G.S., de Jongh, G.J., van Vlijmen-Willems, I.M., Bergboer, J.G., van Rossum, M.M., de Jong, E.M., den

Heijer, M., Evers, A.W., Bergers, M., Armour, J.A., Zeeuwen, P.L. and Schalkwijk, J. (2009) 'Beta-defensin-2 protein is a serum biomarker for disease activity in psoriasis and reaches biologically relevant concentrations in lesional skin', *PloS one,* 4(3), pp. e4725.

Janssen, B., Hartmann, C., Scholz, V., Jauch, A. and Zschocke, J. (2005) 'MLPA analysis for the detection of deletions, duplications and complex rearrangements in the dystrophin gene: potential and pitfalls', *Neurogenetics,* 6(1), pp. 29-35.

Jiang, Y., Wang, Y. and Brudno, M. (2012) 'PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants', *Bioinformatics (Oxford, England),* 28(20), pp. 2576-2583.

Jobling, M., Hollox E., Hurles M., Kivisild T., Tyler-Smith C. (2014). Human Evolutionary Genetics. New York and London: Garland Science. p43-49-75-76-77.

Johnson, M.P., Haupt, L.M. and Griffiths, L.R. (2004) 'Locked nucleic acid (LNA) single nucleotide polymorphism (SNP) genotype analysis and validation using real-time PCR', *Nucleic acids research,* 32(6), pp. e55.

Kai, O.K. and Roberts, D.J. (2008) 'The pathophysiology of malarial anaemia: where have all the red cells gone?', *BMC medicine,* 6, pp. 24-7015-6-24.

Kariuki, S.N., Marin-Menendez, A., Introini, V., Ravenhill, B.J., Lin, Y., Macharia, A., Makale, J., Tendwa, M., Nyamu, W., Kotar, J., Carrasquilla, M., Rowe, J.A., Rockett, K., Kwiatkowski, D., Weekes, M.P., Cicuta, P., Williams, T.N. and Rayner, J.C. (2018) 'Red blood cell tension controls Plasmodium falciparum invasion and protects against severe malaria in the Dantu blood group', *bioRxiv,* .

Katoh, K. and Standley, D.M. (2013) 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability', *Molecular biology and evolution,* 30(4), pp. 772-780.

Kauppinen, S., Nielsen, P.S., Mouritzen, P., Nielsen, A.T., Vissing, H., Møller, S. and Ramsing, N.B. (2003) LNA microarrays in genomics. *PharmaGenomics*, 3, pp. 24–34.

Kauppinen, S., Vester, B. and Wengel, J. (2005) 'Locked nucleic acid (LNA): High affinity targeting of RNA for diagnostics and therapeutics', *Drug discovery today.Technologies,* 2(3), pp. 287-290.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) 'The human genome browser at UCSC', *Genome research,* 12(6), pp. 996-1006.

Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.K. and Eichler, E.E. (2010) 'A human genome structural variation sequencing resource reveals insights into mutational mechanisms', *Cell*, 143(5), pp. 837-847.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., Thomas, J.K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N.A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L.D., Patil, A.H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S.K., Marimuthu, A., Sathe, G.J., Chavan, S., Datta, K.K., Subbannayya, Y., Sahu, A., Yelamanchi, S.D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K.R., Syed, N., Goel, R., Khan, A.A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.C., Zhong, J., Wu, X., Shaw, P.G., Freed, D., Zahari, M.S., Mukherjee, K.K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C.J., Shankar, S.K., Satishchandra, P., Schroeder, J.T., Sirdeshmukh, R., Maitra, A., Leach, S.D., Drake, C.G., Halushka, M.K., Prasad, T.S., Hruban, R.H., Kerr, C.L., Bader, G.D., Iacobuzio-Donahue, C.A., Gowda, H. and Pandey, A. (2014) 'A draft map of the human proteome', *Nature*, 509(7502), pp. 575-581.

Ko, W.Y., Kaercher, K.A., Giombini, E., Marcatili, P., Froment, A., Ibrahim, M., Lema, G., Nyambo, T.B., Omar, S.A., Wambebe, C., Ranciaro, A., Hirbo, J.B. and Tishkoff, S.A. (2011) 'Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations', *American Journal of Human Genetics*, 88(6), pp. 741-754.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., Taillon, B.E., Chen, Z., Tanzer, A., Saunders, A.C., Chi, J., Yang, F., Carter, N.P., Hurles, M.E., Weissman, S.M., Harkins, T.T., Gerstein, M.B., Egholm, M. and Snyder, M. (2007) 'Paired-end mapping reveals extensive structural variation in the human genome', *Science (New York, N.Y.)*, 318(5849), pp. 420-426.

Korbel, J.O., Urban, A.E., Grubert, F., Du, J., Royce, T.E., Starr, P., Zhong, G., Emanuel, B.S., Weissman, S.M., Snyder, M. and Gerstein, M.B. (2007) 'Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome', *Proceedings of the National Academy of Sciences of the United States of America*, 104(24), pp. 10110-10115.

Koshkin, A.A., Singh, S.K., Nielsen, P., Rajwanshi, V.K., Kumar, R., Meldgaard, M., Olsen, C.E. and Wengel, J. (eds.) (1998) *LNA (Locked Nucleic Acids): Synthesis of the adenine, cytosine, guanine, 5-methylcytosine, thymine and uracil bicyclonucleoside monomers, oligomerisation, and unprecedented nucleic acid recognition.*

Kucukkilic, E., Brookes, K., Barber, I., Guetta-Baranes, T., ARUK Consortium, Morgan, K. and Hollox, E.J. (2018) 'Complement receptor 1 gene (CR1) intragenic duplication and risk of Alzheimer's disease', *Human genetics,* 137(4), pp. 305-314.

Kudo, S. and Fukuda, M. (1990) 'Identification of a novel human glycophorin, glycophorin E, by isolation of genomic clones and complementary DNA clones utilizing polymerase chain reaction', *The Journal of biological chemistry,* 265(2), pp. 1102-1110.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007) 'Clustal W and Clustal X version 2.0', *Bioinformatics (Oxford, England),* 23(21), pp. 2947-2948.

Larsen, M.H., Thorner, L.W., Zinyama, R., Amstrup, J., Kallestrup, P., Gerstoft, J., Gomo, E., Erikstrup, C. and Ullum, H. (2012) 'CCL3L gene copy number and survival in an HIV-1 infected Zimbabwean population', *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases,* 12(5), pp. 1087-1093.

Latorra, D., Campbell, K., Wolter, A. and Hurley, J.M. (2003) 'Enhanced allele-specific PCR discrimination in SNP genotyping using 3' locked nucleic acid (LNA) primers', *Human mutation,* 22(1), pp. 79-85.

Latorra, D., Hopkins, D., Campbell, K. and Hurley, J.M. (2003) 'Multiplex allele-specific PCR with optimized locked nucleic acid primers', *BioTechniques,* 34(6), pp. 1150-2, 1154, 1158.

Layer, R.M., Chiang, C., Quinlan, A.R. and Hall, I.M. (2014) 'LUMPY: a probabilistic framework for structural variant discovery', *Genome biology,* 15(6), pp. R84-2014-15-6-r84.

Le Port, A., Cottrell, G., Martin-Prevel, Y., Migot-Nabias, F., Cot, M. and Garcia, A. (2012) 'First malaria infections in a cohort of infants in Benin: biological, environmental and genetic determinants. Description of the study site, population methods and preliminary results', *BMJ open,* 2(2), pp. e000342-2011-000342. Print 2012.

Lee, J.A., Carvalho, C.M. and Lupski, J.R. (2007) 'A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders', *Cell,* 131(7), pp. 1235-1247.

Lee, P.Y., Costumbrado, J., Hsu, C.Y. and Kim, Y.H. (2012) 'Agarose gel electrophoresis for the separation of DNA fragments', *Journal of visualized experiments: JoVE,* (62). pii: 3923. doi(62), pp. 10.3791/3923.

Leffler, E.M., Band, G., Busby, G.B.J., Kivinen, K., Le, Q.S., Clarke, G.M., Bojang, K.A., Conway, D.J., Jallow, M., Sisay-Joof, F., Bougouma, E.C., Mangano, V.D., Modiano, D., Sirima, S.B., Achidi, E., Apinjoh, T.O., Marsh, K., Ndila, C.M., Peshu, N., Williams, T.N., Drakeley, C., Manjurano, A., Reyburn, H., Riley, E., Kachala, D., Molyneux, M., Nyirongo, V., Taylor, T., Thornton, N., Tilley, L., Grimsley, S., Drury, E., Stalker, J., Cornelius, V., Hubbart, C., Jeffreys, A.E., Rowlands, K., Rockett, K.A., Spencer, C.C.A., Kwiatkowski, D.P. and Malaria Genomic Epidemiology Network (2017) 'Resistance to malaria through structural variation of red blood cell invasion receptors', *Science (New York, N.Y.)*, 356(6343), pp. 10.1126/science.aam6393. Epub 2017 May 18.

Li, H. (2011) 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', *Bioinformatics (Oxford, England)*, 27(21), pp. 2987-2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics (Oxford, England)*, 25(16), pp. 2078-2079.

Li, W. and Olivier, M. (2013) 'Current analysis platforms and methods for detecting copy number variation', *Physiological genomics*, 45(1), pp. 1-16.

Li, X., Marinkovic, M., Russo, C., McKnight, C.J., Coetzer, T.L. and Chishti, A.H. (2012) 'Identification of a specific region of Plasmodium falciparum EBL-1 that binds to host receptor glycophorin B and inhibits merozoite invasion in human red blood cells', *Molecular and biochemical parasitology*, 183(1), pp. 23-31.

Lindahl, M. and Wadstrom, T. (1984) 'K99 surface haemagglutinin of enterotoxigenic E. coli recognize terminal N-acetylgalactosamine and sialic acid residues of glycophorin and other complex glycoconjugates', *Veterinary microbiology*, 9(3), pp. 249-257.

Linzmeier, R.M. and Ganz, T. (2005) 'Human defensin gene copy number polymorphisms: comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23', *Genomics*, 86(4), pp. 423-430.

Liu, S., Yao, L., Ding, D. and Zhu, H. (2010) 'CCL3L1 copy number variation and susceptibility to HIV-1 infection: a meta-analysis', *PloS one*, 5(12), pp. e15778.

Lomas-Francis, C. (2011) 'Miltenberger phenotypes are glycophorin variants: a review', *ISBT Science Series*, 6(2), pp. 296-301.

Lupski, J.R. and Stankiewicz, P. (2005) 'Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes', *PLoS genetics*, 1(6), pp. e49.

MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L. and Scherer, S.W. (2014) 'The Database of Genomic Variants: a curated collection of structural variation in the human genome', *Nucleic acids research*, 42(Database issue), pp. D986-92.

Machado, L.R., Bowdrey, J., Ngaimisi, E., Habtewold, A., Minzi, O., Makonnen, E., Yimer, G., Amogne, W., Mugusi, S., Janabi, M., Aderaye, G., Mugusi, F., Viskaduraki, M., Aklillu, E. and Hollox, E.J. (2013) 'Copy number variation of Fc gamma receptor genes in HIV-infected and HIV-tuberculosis co-infected individuals in sub-Saharan Africa', *PloS one*, 8(11), pp. e78165.

MacKenzie, K.R., Prestegard, J.H. and Engelman, D.M. (1997) 'A transmembrane helix dimer: structure and implications', *Science (New York, N.Y.)*, 276(5309), pp. 131-133.

Malaria Genomic Epidemiology Network, Band, G., Rockett, K.A., Spencer, C.C. and Kwiatkowski, D.P. (2015) 'A novel locus of resistance to severe malaria in a region of ancient balancing selection', *Nature*, 526(7572), pp. 253-257.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Jorde, L.B., Tishkoff, S.A., Watkins, W.S., Metspalu, M., Dryomov, S., Sukernik, R., Singh, L., Thangaraj, K., Paabo, S., Kelso, J., Patterson, N. and Reich, D. (2016) 'The Simons Genome Diversity Project: 300 genomes from 142 diverse populations', *Nature*, 538(7624), pp. 201-206.

Marchesi, V.T., Furthmayr, H. and Tomita, M. (1976) 'The red cell membrane', *Annual Review of Biochemistry*, 45, pp. 667-698.

Marshall, C.R. and Scherer, S.W. (2012) 'Detection and characterization of copy number variation in autism spectrum disorder', *Methods in molecular biology (Clifton, N.J.)*, 838, pp. 115-135.

Mayer, D.C., Cofie, J., Jiang, L., Hartl, D.L., Tracy, E., Kabat, J., Mendoza, L.H. and Miller, L.H. (2009) 'Glycophorin B is the erythrocyte receptor of Plasmodium falciparum erythrocyte-binding ligand, EBL-1', *Proceedings of the National Academy of Sciences of the United States of America*, 106(13), pp. 5348-5352.

McCarroll, S.A. and Altshuler, D.M. (2007) 'Copy-number variation and association studies of human disease', *Nature genetics*, 39(7 Suppl), pp. S37-42.

Medvedev, P., Stanciu, M. and Brudno, M. (2009) 'Computational methods for discovering structural variation with next-generation sequencing', *Nature methods*, 6(11 Suppl), pp. S13-20.

Mefford, H.C., Muhle, H., Ostertag, P., von Spiczak, S., Buysse, K., Baker, C., Franke, A., Malafosse, A., Genton, P., Thomas, P., Gurnett, C.A., Schreiber, S., Bassuk, A.G., Guipponi, M., Stephani, U., Helbig, I. and Eichler, E.E. (2010) 'Genome-wide copy

number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies', *PLoS genetics*, 6(5), pp. e1000962.

Merckmillipore.com. (2018). Amicon Ultra-0.5 mL Centrifugal Filters for DNA and Protein Purification and Concentration - Sample Prep Centrifugal Filter Units. [online] Available at: http://www.merckmillipore.com/GB/en/product/Amicon-Ultra-0.5%C2%A0mL-Centrifugal-Filters-for-DNA-and-Protein-Purification-and-Concentration,MM_NF-C82301?ReferrerURL=https%3A%2F%2F%2F (Accessed 19 April. 2018).

Michalet, X., Ekong, R., Fougerousse, F., Rousseaux, S., Schurra, C., Hornigold, N., van Slegtenhorst, M., Wolfe, J., Povey, S., Beckmann, J.S. and Bensimon, A. (1997) 'Dynamic molecular combing: stretching the whole human genome for high-resolution studies', *Science (New York, N.Y.)*, 277(5331), pp. 1518-1523.

Milet, J., Nuel, G., Watier, L., Courtin, D., Slaoui, Y., Senghor, P., Migot-Nabias, F., Gaye, O. and Garcia, A. (2010) 'Genome wide linkage study, using a 250K SNP map, of Plasmodium falciparum infection and mild malaria attack in a Senegalese population', *PloS one*, 5(7), pp. e11616.

Montavon, T., Thevenet, L. and Duboule, D. (2012) 'Impact of copy number variations (CNVs) on long-range gene regulation at the HoxD locus', *Proceedings of the National Academy of Sciences of the United States of America*, 109(50), pp. 20204-20211.

Murayama, J.I., Tomita, M. and Hamada, A. (1982) 'Primary structure of horse erythrocyte glycophorin HA. Its amino acid sequence has a unique homology with those of human and porcine erythrocyte glycophorins', *The Journal of membrane biology*, 64(3), pp. 205-215.

Murray, C.J., Rosenfeld, L.C., Lim, S.S., Andrews, K.G., Foreman, K.J., Haring, D., Fullman, N., Naghavi, M., Lozano, R. and Lopez, A.D. (2012) 'Global malaria mortality between 1980 and 2010: a systematic analysis', *Lancet (London, England)*, 379(9814), pp. 413-431.

Murray, J.C., Buetow, K.H., Weber, J.L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., Quillen, J., Sheffield, V.C., Sunden, S. and Duyk, G.M. (1994) 'A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC)', *Science (New York, N.Y.)*, 265(5181), pp. 2049-2054.

Ndila, C.M., Uyoga, S., Macharia, A.W., Nyutu, G., Peshu, N., Ojal, J., Shebe, M., Awuondo, K.O., Mturi, N., Tsofa, B., Sepulveda, N., Clark, T.G., Band, G., Clarke, G., Rowlands, K., Hubbart, C., Jeffreys, A., Kariuki, S., Marsh, K., Mackinnon, M., Maitland, K., Kwiatkowski, D.P., Rockett, K.A., Williams, T.N. and MalariaGEN Consortium (2018) 'Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study', *The Lancet.Haematology*, 5(8), pp. e333-e345.

Nigg, E.A., Gahmberg, C.G. and Cherry, R.J. (1980) 'Rotational diffusion of band 3 proteins in membranes from En(a-) and neuraminidase-treated normal human erythrocytes', *Biochimica et biophysica acta,* 600(3), pp. 636-642.

Old, J.M. (2003) 'Screening and genetic diagnosis of haemoglobin disorders', *Blood reviews,* 17(1), pp. 43-53.

Ouahchi, K., Lindeman, N. and Lee, C. (2006) 'Copy number variants and pharmacogenomics', *Pharmacogenomics,* 7(1), pp. 25-29.

Palacajornsuk, P. (2006) 'Review: molecular basis of MNS blood group variants', *Immunohematology / American Red Cross,* 22(4), pp. 171-182.

Pavone, B.G., Billman, R., Bryant, J., Sniecinski, I. and Issitt, P.D. (1981) 'An auto-anti-Ena, inhibitable by MN sialoglycoprotein', *Transfusion,* 21(1), pp. 25-31.

Pedersen, K., Wiechec, E., Madsen, B.E., Overgaard, J. and Hansen, L.L. (2010) 'A simple way to evaluate self-designed probes for tumor specific Multiplex Ligation-dependent Probe Amplification (MLPA)', *BMC research notes,* 3, pp. 179-0500-3-179.

Penman, B.S., Pybus, O.G., Weatherall, D.J. and Gupta, S. (2009) 'Epistatic interactions between genetic disorders of hemoglobin can explain why the sickle-cell gene is uncommon in the Mediterranean', *Proceedings of the National Academy of Sciences of the United States of America,* 106(50), pp. 21242-21246.

Perry, G.H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C.W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N.A., Park, H.S., Kim, J.I., Seo, J.S., Yakhini, Z., Laderman, S., Bruhn, L. and Lee, C. (2008) 'The fine-scale and complex architecture of human copy-number variation', *American Journal of Human Genetics,* 82(3), pp. 685-695.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., Carter, N.P., Lee, C. and Stone, A.C. (2007) 'Diet and the evolution of human amylase gene copy number variation', *Nature genetics,* 39(10), pp. 1256-1260.

Piotrowski, A., Bruder, C.E., Andersson, R., Diaz de Stahl, T., Menzel, U., Sandgren, J., Poplawski, A., von Tell, D., Crasto, C., Bogdan, A., Bartoszewski, R., Bebok, Z., Krzyzanowski, M., Jankowski, Z., Partridge, E.C., Komorowski, J. and Dumanski, J.P. (2008) 'Somatic mosaicism for copy number variation in differentiated human tissues', *Human mutation,* 29(9), pp. 1118-1124.

Pirooznia, M., Goes, F.S. and Zandi, P.P. (2015) 'Whole-genome CNV analysis: advances in computational approaches', *Frontiers in genetics,* 6, pp. 138.

Polan, M.B., Pastore, M.T., Steingass, K., Hashimoto, S., Thrush, D.L., Pyatt, R., Reshmi, S., Gastier-Foster, J.M., Astbury, C. and McBride, K.L. (2014) 'Neurodevelopmental disorders among individuals with duplication of 4p13 to 4p12 containing a GABAA receptor subunit gene cluster', *European journal of human genetics: EJHG,* 22(1), pp. 105-109.

Polley, S., Louzada, S., Forni, D., Sironi, M., Balaskas, T., Hains, D.S., Yang, F. and Hollox, E.J. (2015) 'Evolution of the rapidly mutating human salivary agglutinin gene (DMBT1) and population subsistence strategy', *Proceedings of the National Academy of Sciences of the United States of America,* 112(16), pp. 5105-5110.

Postoway, N., Anstee, D.J., Wortman, M. and Garratty, G. (1988) 'A severe transfusion reaction associated with anti-EnaTS in a patient with an abnormal alpha-like red cell sialoglycoprotein', *Transfusion,* 28(1), pp. 77-80.

Raap, A.K. (1998) 'Advances in fluorescence in situ hybridization', *Mutation research,* 400(1-2), pp. 287-298.

Ravenhall, M., Campino, S., Sepulveda, N., Manjurano, A., Nadjm, B., Mtove, G., Wangai, H., Maxwell, C., Olomi, R., Reyburn, H., Drakeley, C.J., Riley, E.M., Clark, T.G. and MalariaGEN (2018) 'Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania', *PLoS genetics,* 14(1), pp. e1007172.

Rearden, A., Magnet, A., Kudo, S. and Fukuda, M. (1993) 'Glycophorin B and glycophorin E genes arose from the glycophorin A ancestral gene via two duplications during primate evolution', *The Journal of biological chemistry,* 268(3), pp. 2260-2267.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., F., Zhang, Conrad, D.F., Estivill, X., Tyler-Smith, C., Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W. and Hurles, M.E. (2006) 'Global variation in copy number in the human genome', *Nature,* 444(7118), pp. 444-454.

Reid, M.E. (2009) 'MNS blood group system: a review', *Immunohematology / American Red Cross,* 25(3), pp. 95-101.

Rooth, I. (1992). Malaria morbidity and control in a Tanzanian village, Doctoral thesis. Karolinska Institutet, Stockholm.

Rouillac, C., Colin, Y., Hughes-Jones, N.C., Beolet, M., D'Ambrosio, A.M., Cartron, J.P. and Le Van Kim, C. (1995) 'Transcript analysis of D category phenotypes predicts hybrid Rh D-CE-D proteins associated with alteration of D epitopes', *Blood,* 85(10), pp. 2937-2944.

Satchwell, T.J. (2016) 'Erythrocyte invasion receptors for Plasmodium falciparum: new and old', *Transfusion medicine (Oxford, England),* 26(2), pp. 77-88.

Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S.A., Handsaker, R.E., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Daly, M.J., Carroll, M.C., Stevens, B. and McCarroll, S.A. (2016) 'Schizophrenia risk from complex variation of complement component 4', *Nature*, 530(7589), pp. 177-183.

Schauer, R. (1982) 'Chemistry, metabolism, and biological functions of sialic acids', *Advances in Carbohydrate Chemistry and Biochemistry*, 40, pp. 131-234.

Scherer, S.W., Lee, C., Birney, E., Altshuler, D.M., Eichler, E.E., Carter, N.P., Hurles, M.E. and Feuk, L. (2007) 'Challenges and standards in integrating surveys of structural variation', *Nature genetics*, 39(7 Suppl), pp. S7-15.

Schneider, C.A., Rasband, W.S. and Eliceiri, K.W. (2012) 'NIH Image to ImageJ: 25 years of image analysis', *Nature methods*, 9(7), pp. 671-675.

Schneider, G.F. and Dekker, C. (2012) 'DNA sequencing with nanopores', *Nature biotechnology*, 30(4), pp. 326-328.

Schouten, J.P., McElgunn, C.J., Waaijer, R., Zwijnenburg, D., Diepvens, F. and Pals, G. (2002) 'Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification', *Nucleic acids research*, 30(12), pp. e57.

Schulte, T.H. and Marchesi, V.T. (1979) 'Conformation of human erythrocyte glycophorin A and its constituent peptides', *Biochemistry*, 18(2), pp. 275-280.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T.C., Trask, B., Patterson, N., Zetterberg, A. and Wigler, M. (2004) 'Large-scale copy number polymorphism in the human genome', *Science (New York, N.Y.)*, 305(5683), pp. 525-528.

Shendure, J. and Ji, H. (2008) 'Next-generation DNA sequencing', *Nature biotechnology*, 26(10), pp. 1135-1145.

Shlien, A. and Malkin, D. (2009) 'Copy number variations and cancer', *Genome medicine*, 1(6), pp. 62.

Sim, B.K., Chitnis, C.E., Wasniowska, K., Hadley, T.J. and Miller, L.H. (1994) 'Receptor and ligand domains for invasion of erythrocytes by Plasmodium falciparum', *Science (New York, N.Y.)*, 264(5167), pp. 1941-1944.

Stankiewicz, P. and Lupski, J.R. (2002) 'Genome architecture, rearrangements and genomic disorders', *Trends in genetics : TIG*, 18(2), pp. 74-82.

Stevens, E.L., Heckenberg, G., Baugher, J.D., Roberson, E.D., Downey, T.J. and Pevsner, J. (2012) 'Consanguinity in Centre d'Etude du Polymorphisme Humain (CEPH) pedigrees', *European journal of human genetics : EJHG*, 20(6), pp. 657-667.

Storry, J.R., Poole, J., Condon, J. and Reid, M.E. (2000) 'Identification of a novel hybrid glycophorin gene encoding GP.Hop', *Transfusion*, 40(5), pp. 560-565.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S.W., Tavare, S., Deloukas, P., Hurles, M.E. and Dermitzakis, E.T. (2007) 'Relative impact of nucleotide and copy number variation on gene expression phenotypes', *Science (New York, N.Y.)*, 315(5813), pp. 848-853.

Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., 1000 Genomes Project and Eichler, E.E. (2010) 'Diversity of human copy number variation and multicopy genes', *Science (New York, N.Y.)*, 330(6004), pp. 641-646.

Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., Jorde, L.B., Posukh, O.L., Sahakyan, H., Watkins, W.S., Yepiskoposyan, L., Abdullah, M.S., Bravi, C.M., Capelli, C., Hervig, T., E., Klitz, W., Winkler, C., Labuda, D., Metspalu, M., Tishkoff, S.A., Dryomov, S., Sukernik, R., Patterson, N., Reich, D. and Eichler, E.E. (2015) 'Global diversity, population stratification, and selection of human copy-number variation', *Science (New York, N.Y.)*, 349(6253), pp. aab3761.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H., Konkel, M.K., Malhotra, A., Stutz, A.M., Shi, X., Casale, F.P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M.J.P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H.Y.K., Mu, X.J., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Zichner, T., Sebat, J., Batzer, M.A., McCarroll, S.A., 1000 Genomes Project Consortium, Mills, R.E., Gerstein, M.B., Bashir, A., Stegle, O., Devine, S.E., Lee, C., Eichler, E.E. and Korbel, J.O. (2015) 'An integrated map of structural variation in 2,504 human genomes', *Nature*, 526(7571), pp. 75-81.

Szumilas, M. (2010) 'Explaining odds ratios', *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent*, 19(3), pp. 227-229.

Tabelow, K., Clayden, J.D., de Micheaux, P.L., Polzehl, J., Schmid, V.J. and Whitcher, B. (2011) 'Image analysis and statistical inference in neuroimaging with R', *NeuroImage*, 55(4), pp. 1686-1693.

Tako, E.A., Zhou, A., Lohoue, J., Leke, R., Taylor, D.W. and Leke, R.F. (2005) 'Risk factors for placental malaria and its effect on pregnancy outcome in Yaounde, Cameroon', *The American Journal of Tropical Medicine and Hygiene*, 72(3), pp. 236-242.

Tanner, M.J. (1993) 'The major integral proteins of the human red cell', *Bailliere's Clinical Haematology,* 6(2), pp. 333-356.

Tate, C.G. and Tanner, M.J. (1988) 'Isolation of cDNA clones for human erythrocyte membrane sialoglycoproteins alpha and delta', *The Biochemical journal,* 254(3), pp. 743-750.

Tayyab, S. and Qasim, M.A. (1988) 'Biochemistry and roles of glycophorin A', *Biochemical education,* 16(2), pp. 63-66.

Teo, S.M., Pawitan, Y., Ku, C.S., Chia, K.S. and Salim, A. (2012) 'Statistical challenges associated with detecting copy number variations with next-generation sequencing', *Bioinformatics (Oxford, England),* 28(21), pp. 2711-2718.

Timmann, C., Thye, T., Vens, M., Evans, J., May, J., Ehmen, C., Sievertsen, J., Muntau, B., Ruge, G., Loag, W., Ansong, D., Antwi, S., Asafo-Adjei, E., Nguah, S.B., Kwakye, K.O., Akoto, A.O., Sylverken, J., Brendel, M., Schuldt, K., Loley, C., Franke, A., Meyer, C.G., Agbenyega, T., Ziegler, A. and Horstmann, R.D. (2012) 'Genome-wide association study indicates two novel resistance loci for severe malaria', *Nature,* 489(7416), pp. 443-446.

Ugozzoli, L.A., Latorra, D., Puckett, R., Arar, K. and Hamby, K. (2004) 'Real-time genotyping with oligonucleotide probes containing locked nucleic acids', *Analytical Biochemistry,* 324(1), pp. 143-152.

Uneke, C.J. (2007) 'Impact of placental Plasmodium falciparum malaria on pregnancy and perinatal outcome in sub-Saharan Africa: I: introduction to placental malaria', *The Yale journal of biology and medicine,* 80(2), pp. 39-50.

Veal, C.D., Xu, H., Reekie, K., Free, R., Hardwick, R.J., McVey, D., Brookes, A.J., Hollox, E.J. and Talbot, C.J. (2013) 'Automated design of paralogue ratio test assays for the accurate and rapid typing of copy number variation', *Bioinformatics (Oxford, England),* 29(16), pp. 1997-2003.

Velliquette, R.W., Palacajornsuk, P., Hue-Roye, K., Lindgren, S., Ilstrup, S., Green, C., Lomas-Francis, C. and Reid, M.E. (2008) 'Novel GYP(A-B-A) hybrid gene in a DANE+ person who made an antibody to a high-prevalence MNS antigen', *Transfusion,* 48(12), pp. 2618-2623.

Volik, S., Raphael, B.J., Huang, G., Stratton, M.R., Bignel, G., Murnane, J., Brebner, J.H., Bajsarowicz, K., Paris, P.L., Tao, Q., Kowbel, D., Lapuk, A., Shagin, D.A., Shagina, I.A., Gray, J.W., Cheng, J.F., de Jong, P.J., Pevzner, P. and Collins, C. (2006) 'Decoding the fine-scale structure of a breast cancer genome and transcriptome', *Genome research*, 16(3), pp. 394-404.

Wain, L.V., Armour, J.A. and Tobin, M.D. (2009) 'Genomic copy number variation, human health, and disease', *Lancet (London, England)*, 374(9686), pp. 340-350.

Wain, L.V., Odenthal-Hesse, L., Abujaber, R., Sayers, I., Beardsmore, C., Gaillard, E.A., Chappell, S., Dogaru, C.M., McKeever, T., Guetta-Baranes, T., Kalsheker, N., Kuehni, C.E., Hall, I.P., Tobin, M.D. and Hollox, E.J. (2014) 'Copy number variation of the beta-defensin genes in europeans: no supporting evidence for association with lung function, chronic obstructive pulmonary disease or asthma', *PloS one*, 9(1), pp. e84192.

Walker, S., Janyakhantikul, S. and Armour, J.A. (2009) 'Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs', *Genomics*, 93(1), pp. 98-103.

Wang, H.Y., Tang, H., Shen, C.K. and Wu, C.I. (2003) 'Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites', *Molecular biology and evolution*, 20(11), pp. 1795-1804.

Wassmer, S.C. and Carlton, J.M. (2016) 'Glycophorins, Blood Groups, and Protection from Severe Malaria', *Trends in parasitology*, 32(1), pp. 5-7.

Weiss, M.M., Hermsen, M.A., Meijer, G.A., van Grieken, N.C., Baak, J.P., Kuipers, E.J. and van Diest, P.J. (1999) 'Comparative genomic hybridisation', *Molecular pathology: MP*, 52(5), pp. 243-251.

Wellcome Trust Case Control Consortium, Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., Giannoulatou, E., Holmes, C., Marchini, J.L., Stirrups, K., Tobin, M.D., Wain, L.V., Yau, C., Aerts, J., Ahmad, T., Andrews, T.D., Arbury, H., Attwood, A., Auton, A., Ball, S.G., Balmforth, A.J., Barrett, J.C., Barroso, I., Barton, A., Bennett, A.J., Bhaskar, S., Blaszczyk, K., G., Eyre, S., Farmer, A., Ferrier, I.N. and Donnelly, P. (2010) 'Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls', *Nature*, 464(7289), pp. 713-720.

WHO (2016a), 'Malaria in children under five', [online] Available from: http://www.who.int/malaria/areas/high_risk_groups/children/en/ (Accessed 23 April 2017).

WHO (2016b), 'Malaria in infants', [online] Available from: http://www.who.int/malaria/areas/high_risk_groups/infants/en/ (Accessed 23 April 2017).

Wong, W.W., Cahill, J.M., Rosen, M.D., Kennedy, C.A., Bonaccio, E.T., Morris, M.J., Wilson, J.G., Klickstein, L.B. and Fearon, D.T. (1989) 'Structure of the human CR1 gene. Molecular basis of the structural and quantitative polymorphisms and identification of a new CR1-like allele', *The Journal of experimental medicine,* 169(3), pp. 847-863.

Wright, G.J. and Rayner, J.C. (2014) 'Plasmodium falciparum erythrocyte invasion: combining function with immune evasion', *PLoS pathogens,* 10(3), pp. e1003943.

Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J. (2009) 'Sensitive and accurate detection of copy number variants using read depth of coverage', *Genome research,* 19(9), pp. 1586-1592.

You, Y., Moreira, B.G., Behlke, M.A. and Owczarzy, R. (2006) 'Design of LNA probes that improve mismatch discrimination', *Nucleic acids research,* 34(8), pp. e60.

Zarrei, M., MacDonald, J.R., Merico, D. and Scherer, S.W. (2015) 'A copy number variation map of the human genome', *Nature reviews.Genetics,* 16(3), pp. 172-183.

Zeitouni, B., Boeva, V., Janoueix-Lerosey, I., Loeillet, S., Legoix-ne, P., Nicolas, A., Delattre, O. and Barillot, E. (2010) 'SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data', *Bioinformatics (Oxford, England),* 26(15), pp. 1895-1896.

Zhang, Z.D., Du, J., Lam, H., Abyzov, A., Urban, A.E., Snyder, M. and Gerstein, M. (2011) 'Identification of genomic indels and structural variations using split reads', *BMC genomics,* 12, pp. 375-2164-12-375.

Zhang, F., Gu, W., Hurles, M.E. and Lupski, J.R. (2009) 'Copy number variation in human health, disease, and evolution', *Annual review of genomics and human genetics,* 10, pp. 451-481.

Zhang, F., Khajavi, M., Connolly, A.M., Towne, C.F., Batish, S.D. and Lupski, J.R. (2009) 'The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans', *Nature genetics,* 41(7), pp. 849-853.

Zhao, M., Wang, Q., Wang, Q., Jia, P. and Zhao, Z. (2013) 'Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives', *BMC bioinformatics,* 14 Suppl 11, pp. S1-2105-14-S11-S1. Epub 2013 Sep 13.

# Appendices

## Appendix 1: General reagents

### 10X Low dNTPs PCR Mix

10X Ld PCR buffer was used for the Paralogue Ratio Test method. The buffer contained a final concentration of 50mM Tris-HCl (pH8.8), 12.5mM ammonium sulphate, 1.4mM magnesium chloride, 7.5mM 2-mercaptoethanol, 125µg/mL non-acetylated Bovine Serum Albumin (BSA) (Ambion®, Thermo Scientific) and 200µM of each dNTP.

### 1.11X dNTPs buffer

11.1x dNTPs mix was used for the long-PCR method. The buffer contained a final concentration of 45mM Tris-HCl (pH 8.8), 11 mM ammonium sulphate, 4.5mM magnesium chloride, 6.7 mM 2-mercaptoethanol, 4.4 mM EDTA (pH 8.0), 113 µg/mL non-acetylated Bovine Serum Albumin (BSA) (Ambion®, Thermo Scientific) and 1 mM of each dNTP (Promega).

### 1x Tris-Borate-EDTA Buffer (TBE Buffer)

TBE (Tris/Borate/EDTA)

This buffer was used in agarose gel electrophoresis and was made from 40mM Tris-HCl (pH8.3), 4mM Boric acid and 1mM EDTA.

### 1x Tris-EDTA Buffer (TEA Buffer)

TAE (Tris/Acetic acid/EDTA).

1x TAE Buffer was composed of 10mM Tris-HCl (pH8.0), which maintains the pH of the solution and 1mM EDTA, a chelator of metal ions which helps protect DNA and RNA from enzymatic degradation.

**Appendix 2A: The annealing temperature for each variant-specific Long-PCR assay.**

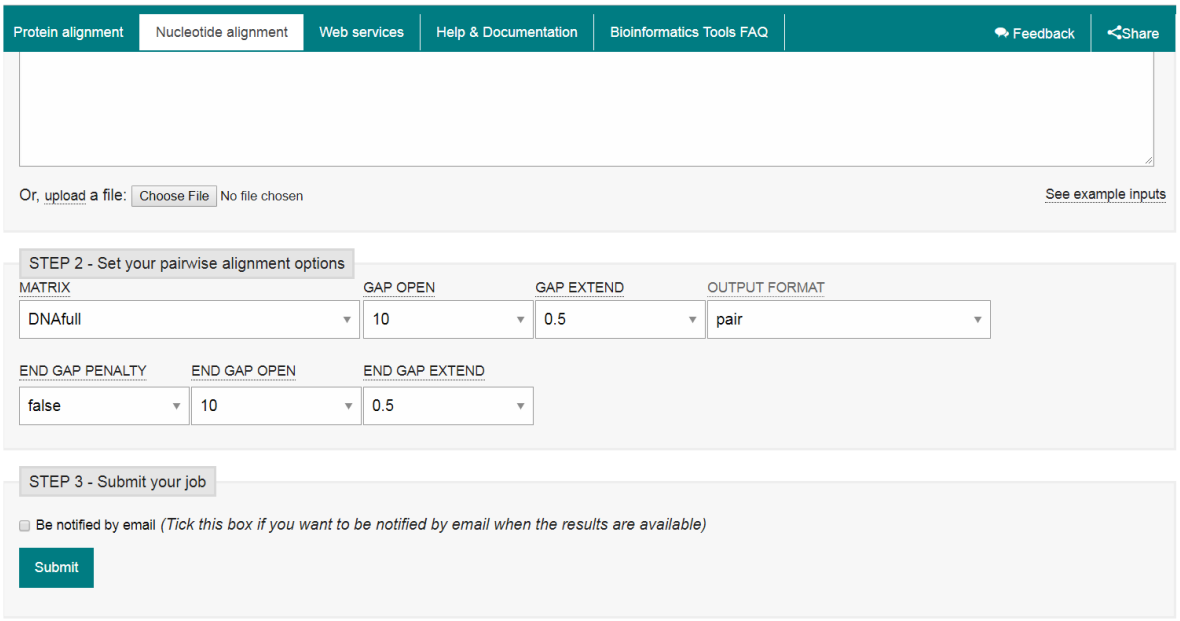
| Specific assay | Annealing temperature |
|---|---|
| **DEL1** | 65°C |
| **DEL2** | 65°C |
| **DEL6** | 63°C |
| **DEL7** | 65°C |
| **DUP2** | 64°C |
| **DUP3** | 63.5°C |
| **Specific_*GYPE*** | 63°C |
| **DUP5** | 65-67°C |
| **DUP7** | 64°C |
| **DUP14** | 65°C |
| **DUP29** | 64°C |
| **DUP24** | Cannot be optimised |
| ***GYPE-B-E* Gene conversion** | 65°C |

# Appendix 2B: Table of all the sequencing primers used throughout this work

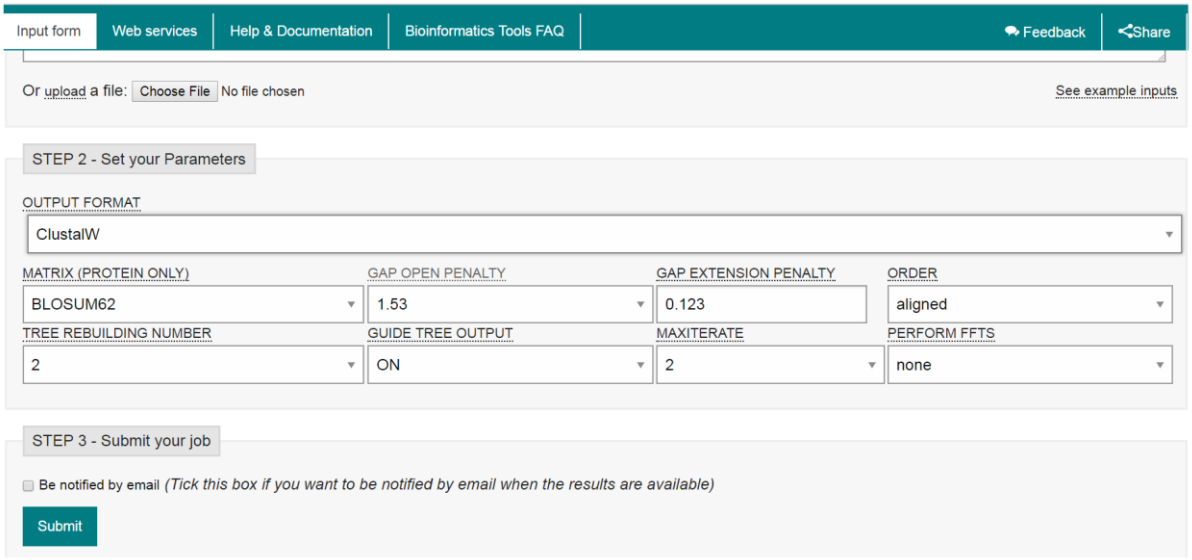
| Primer name | Primer sequence 5' to 3' | Primer name | Primer sequence 5' to 3' |
|---|---|---|---|
| GYP_del_A_seq_F1 | GTAAAGATGAATTTAGCCCTGG | DUP29_stage6_F | CATTGTGGAAGTCAGTGTGGCAATTCC |
| GYP_del_A_seq_F2 | AAGCCCTGTGGGCTCACTGGG | DUP29_stage6_R1 | TCTGCATCTCCACTCCCTCCATCCATT |
| GYP_del_A_seq_F4 | GGAGAAGGAACCAGAAAGCCA | DUP29_stage6_R2 | CTCCTCCCTCTCTGGGATGCTCTG |
| GYP_del_A_seq_R1 | ACTGGATCCCTTCCTTACAC | DUP29_stage7_F | AGGGGAATATTACACTCTGGGGACTGC |
| GYP_del_A_seq_R3 | GAGATATCATCTCACACCA | DUP29_stage7_R | GTCCAGTTCTTTGTCCAGTTGGGTGGC |
| GYP_del_B_seq_F1 | ATATGTCTGCCTGCTACGCT | DUP29_stage8_F | TGAAGGACTAGAACCAACATTTCTCC |
| GYP_del_B_seq_F2 | GGTCAGAATGTGGAATGAGT | DUP29_stage8_F2 | TCAAATGATGGATCGATAAGCAAAATG |
| GYP_del_B_seq_F3 | AGGGGAAGCAAACACATCCT | DUP29_stage8_R | TGTAAACTACCCTATTTTAGTTCTATC |
| GYP_del_B_seq_F5 | CTTCCTAGATACAATGAGGGT | DUP29_stage9_F | CCTTCTGGAGTGATGAAAATGTTCTGG |
| GYP_del_B_seq_F6 | TAGCAGCACCCCACCTCTAGTA | DUP29_stage9_F2 | TCTGTAGAGGTAGAAAACAAATTAG |
| GYP_del_B_seq_R1 | ACCTGGACAACTCTGTTGGGT | DUP29_stage9_R | GCTGTAATCCAAGTGAGAGCAAGTATC |
| GYP_del_B_seq_R3 | GAACAAAATGCAAAGTAAC | DUP29_stage10_F | GTATAGTCATAGTCTATGACCAACG |
| Dup_B_stage2_Forward | AATGCCATCCCCATCAAG | DUP29_stage10_F2 | TAGGTACTTACTAAAGATGTGGATATG |
| Dup_B_stage2_Revese | GTGTGCATGTGTCTTTAAAGCAG | DUP29_stage10_R | GACTAGGGAACAAACCTACGTATGTG |
| Dup_B_stage3_Forward | GGATATGAACAGACACTTCTC | DUP29_stage11_F | CTACTCGGTGCTTGATTGAATAGACAC |
| Dup_B_stage3_Reverse | CTTCAGGTCTAACATTTAAGTC | DUP29_stage11_F2 | GTCATTTTGAGTAGTTGATATTGACTG |
| Dup_B_stage3_Forward | GGATATGAACAGACACTTCTC | DUP29_stage11_R | CAAATTTGGAAAGGGTTCCTGGCAAAG |
| Dup_B_stage3_Reverse | CTTCAGGTCTAACATTTAAGTC | DUP29_stage11_R2 | CAAAATTAGACACTGCGTACTCTAC |
| Del_C_stage2_Forward | CAGCACAGAACACAAACTCTG | DUP29_stage12_F | TAGCAAGGCTCTCTTCTTGGATCCCTG |
| Del_C_stage2_Revese | CAAAGACATCAAAGTGTTGCTAC | DUP29_stage12_R | CTGATTAGCCGAAGACTAAACATGGTG |
| Del_C_stage3_Forward | CTAAATCAAACTCAGACTTACTCTG | DUP29_stage12_F2 | CAGGTTTGGCTTTTCAAGAATTTATAG |
| Del_C_stage2_R | AAAACATTCATGTCTTGTGATGGCCTG | DUP29_stage12_R2 | CCAGTACCCAAGCTGTTATCATCTTCG |
| Del_C_stage4_R | GAGACACTCCTTTATATGCTAGCAGC | DUP29_stage13_F | GATGAATAAAGTAGGCATGATGATATC |
| Del_C_stage5_R1 | CTATGCAGTGTATTTCAGTGTAAACAG | DUP29_Stage 13_R | TGGAGAACCCATTTTTTCTAATATGAC |
| Del_C_stage5_R2 | TATGGAATCATACACATTGTTGAGATC | DUP29_stage14_R1 | TTTGCATTCTAGGTCATAACATAC |
| Del_C_stage5_R1 | CTATGCAGTGTATTTCAGTGTAAACAG | DUP29_stage14_R2 | ACTGGTATGATATCATCATGCCTAC |
| Del_C_stage5_R2 | TATGGAATCATACACATTGTTGAGATC | DUP29_stage14_R1 | TTTGCATTCTAGGTCATAACATAC |
| Del_C_stage6_R1 | CCTTTCAAGAAAATAATCTTAAAAGTG | DUP29_stage14_R2 | ACTGGTATGATATCATCATGCCTAC |
| Del_C_stage6_R2 | CCCATGAGTCTCACAGATTACTACACC | DUP29_stage14_F1 | GGGAATACATGGATGAATAAAGTAGGC |
| Del_C_stage7_R1 | CAGTTTTCTGAGCAATTAAATTCTGGC | DUP29_stage14_F2 | GCTACTATAAATACCAATCGGGAACTG |
| Del_C_stage7_R2 | GTGCATCAAAGGACACTATCAAAAG | Dup5_stage2_F | CAACTGCAAATGTAATATCTTTTGCTC |
| Del_C_stage8_R | TCTTCTGCTAACAATTTATAATACTTC | Dup5_stage2_F2 | TATTATCATATCCACATCTTTAGTAAG |
| Del_C_stage9_R | CATATTTACTATATCACTGTTCCAC | Dup5_stage8_F | TAGAGACTAATCCATGTATAATGATGG |
| Del_C_stage9_R2 | ATTATCTGCCAATATTCTAAACAGAAG | Dup5_stage8_F2 | CTTCCTCACGTGGTGCAGCTAAGGAAC |
| Del6_stage2_Forward | CTCATTACCCTAAAACCTATATTATAATG | Dup5_stage8_F1 | TGCTAGATTCAGACCTCAGAAACATTG |
| Del6_stage2_Reverse | GAATTGTTATGCTAGTTTCACTAATGAGG | Dup5_stage8_F2 | AGCTATTTGAGGCAACTGAGAATAGAG |
| Del_6_stage3_R1 | CATTCATCAACTCTTTACTATTCTG | Dup5_stage9_F | GCCTCCCGGGTTCAAGCAATTCTCCTG |
| Del_6_stage3_F1 | CTGGCAGCAGAGGAATACTTG | Dup5_stage9_F2 | GATTACAGGCATGAGCCATTGCACCAG |
| Del_6_stage3_R2 | ATGACTTTATTCTCTTTTATATATTTC | Dup5_stage10_F1 | GACAAGATTTTGCTCTGTGGCCTAAGC |
| Del_6_stage3_F2 | TATTTTTCCAAGCAATATTCTTCATG | Dup5_stage10_F2 | TCATAGCTTACTTCAGTCTTGAACTGC |
| Del6_stage4_F | CTTTGAAAAATGCAAAAATGCTCACCG | Dup5_stage10_F3 | CTCCAGCTTCTGCCTCTCAAGTAGCTG |
| Del6_stage4_R1 | GTTCTCAGCAAGTAAATAAATAAAAAC | Dup5_stage11_F1 | GAGGCAGGAGAATGGAGTGAACCCGGG |
| Del6_stage4_R2 | ACATACAGTGTGATTGTTCTCAGCAAG | Dup5_stage11_F1 | GAGGCAGGAGAATGGAGTGAACCCGGG |
| Del6_stage5_F | TGTTGTAAACTCACGCATGGGCTTTTG | Dup5_stage11_F2 | GATCACGAGGTCAGGAGATCGAGATCG |
| Del6_stage5_R | TTAGGGCTTCAATATACACATTCTG | Dup5_stage11_F1 | AGAATGGAGTGAACCCGGGAGGCGGAG |
| Del6_stage6_F | GCTTTCAAAGCTCCTCTTCCTTTTGTC | Dup5_stage11_F2 | AGCCGAGATCGGGCCACTGCACTCCAG |
| Del6_stage6_R | GCTTAAGTCTCAAATTTCTTATCTG | Dup5_stage11_F3 | ATTATAGCCCAATAATATAGCAACGAG |
| Del6_stage7_F1 | CTTTGAGGAACATATTATTCCTATTTTTC | Dup5_stage11_F4 | TGATACAATGTTTCTGAGGTCTGAATC |
| Del6_stage7_R1 | TTCTAAGACGTGAATCACTCTCTTTG | Dup5_stage13_F1 | CAGTTAGAGTTCCTTAGCTGCACCACG |
| Del6_stage8_F | TGTCATTATAAGAAGAACACCAGAGAG | Dup5_stage13_F2 | TCTCTACCTTATAGAATTATTGAACAG |
| Del6_stage8_R | GGCACTTTCATTATCCCCATTCTTAAG | Dup5_stage13_F3 | TAGACATATTCTTCGGATAGAATTGAG |
| Del6_stage9_F | GTGATGTCTTTGTGAGAATGCTATATG | GC_stage3_F2 | CAGATTGCATCTGAATGGGAACATAAT |
| Del6_stage9_R | CTTTGGCCCCTTTGCTCCTGATCGGTG | GC_stage4_F | TATAAATTAATAATTAATAATAACATG |
| Del6_stage10_F | TTATTCTGAAGATGACTATTATAATCG | GC_stage4_R | GCTCAGCCACTGGAAGTTTTGGCATAC |
| Del6_stage10_R | CTATCCTATTGTATCCTATTATATCTG | GC_stage5_R | CTTCACCTAAAACTCTGCCCATCCTTC |
| DUP29_stage2_F1 | GATAGGTTAGGACTGGTGCAATGGCTC | GC_stage5_R2 | CTTGCCTGCTACGTGCCCTAAGCCTCC |
| DUP29_stage2_F2 | GCCTGGGCAACACGGTGAAACCCCGTG | GC_stage6_R | GAGACCAAATGTCTATCTGTAAAGTTG |
| DUP29_stage2_R1 | CATAGGTAGAATTTTATGATAATTATG | GC_stage7_R | CTTGGTGATTTTAAGGGATCATATTG |
| DUP29_stage2_R2 | GTGCAGATACGCCACTAGAGAGCACTG | GC_stage8_R1 | GGAATAACAAATTATTTTAAAG |
| DUP29_stage2_R2 | GTGCAGATACGCCACTAGAGAGCACTG | GC_stage8_R2 | CAAGTATTTCTAAAAACATAAGTGGC |
| DUP29_stage3_R1 | GACTTACTTTAATAGGGCCAGGTACTC | GC_stage8_R3 | GGAAGGAAAATAAACTATGAAAAAGG |
| DUP29_stage3_F1 | AGGCTGAGGCAGGAGAATTGCTTG | GC_stage9_R1 | CAGAACAAAATTTCCACCAAGTGCTC |
| DUP29_stage3_F2 | AGTTTTCTCCAATTGTGATAGATCAAG | GC_stage9_R2 | CTGTAATTTTCTGCCCCCTTTAAAGCC |
| DUP29_stage3_F3 | ACAATGTTTTTGAAGTCTGAATCTAGC | GC_stage9_R3 | GTTTGCAAAGGGACTGCTTATCTTG |
| DUP29_stage2_R2 | GTGCAGATACGCCACTAGAGAGCACTG | GC_stage10_R2 | GTTAAGAAGTGATAAGGACCTTCTGTG |
| DUP29_stage3_R1 | GACTTACTTTAATAGGGCCAGGTACTC | GC_stage10_R1 | ATGGGACCCTCTACTGCAAAGCTCTGG |
| DUP29_stage3_R2 | AGCTAGGATATGACATAGGTTACCCGC | GC_stage10_R2 | GTTAAGAAGTGATAAGGACCTTCTGTG |
| DUP29_stage3_F1 | AGGCTGAGGCAGGAGAATTGCTTG | GC_stage3_F | GACTGAGTACAAATATACATATTCCTCC |
| DUP29_stage3_F2 | TTCTTAAGAAATTATGGACCATTATAG | GC_stage10_R3 | GTTATCCTCCATACTACTGACAGTG |
| DUP29_stage3_F3 | GATCTGAAATATATCTGCTTGCAATC | GC_stage4_F1 | AGGCAGGCTGTGTGGTACCAGCAATG |
| DUP29_stage3_F2 | AGTTTTCTCCAATTGTGATAGATCAAG | GC_stage4_F2 | AGCGGTCAACTCTACTTGGAGAGGTAG |
| DUP29_stage3_F3 | ACAATGTTTTTGAAGTCTGAATCTAGC | GC_stage4_F1 | AGGCAGGCTGTGTGGTACCAGCAATG |
| DUP29_stage4_R1 | CTCCATATCTTTATTTTTGAAAACATG | GC_stage4_F2 | AGCGGTCAACTCTACTTGGAGAGGTAG |
| DUP29_stage4_F2 | GGAAAACCTAATCCTAAAATTTCTGTG | GC_stage5_F1 | AGGCACCAAATGCTGTGGGCCATGTTC |
| DUP29_stage5_F | GTGCATCAAAGGACACTATCAAAAG | GC_stage6_F1 | TAGGGAAAATCACTGTCAGTAGTATGG |
| DUP29_stage5_R | ACTTGGTTGACTCTTGAGCACAATTCC | GC_stage6_F2 | CCCATCTTGTGTTCAAGAGTGGCCAAG |
| DUP29_stage5_R2 | TGATACTTATCTCCATCTTTTATCCAG | | |

## Appendix 3: Alignments software settings

The European Molecular Biology Open Software Suite (EMBOSS-Needle) pairwise alignment options that have been used (STEP 2):




Multiple Alignment using Fast Fourier Transform (MAFFT) pairwise alignment options that have been used (STEP 2):



# Appendix 4: All alignments that show the breakpoints

## Multiple sequence alignment for deletion DEL1 breakpoints

```
Del1 GGAAGTATATTGATAAAAGAAAAGGAACAAATATGTAGTCATTTCTAAGATTTATTTTTA
GYPB GGAAGTATACTGATAAAAGAAAAGGAACAAATATGTAGTCATTTCTAAGATTTATTTTTA
GYPE GGAAGTATATTGATAAAAGAAAAGGAACAAATATGTAGTCATTTCTAAGATTTATTTTTA
     *********.**************************************************

Del1 TTAAAATGTAACACTCACAAACTGTAAGTGTGTAACTTATATATATATACACACATAT--
GYPB TTAAAATGTAACACTCACAAAC--TAAGTGTATAACTCATATGTATATATACACACATAC
GYPE TTAAAATGTAACACTCACAAACTGTAAGTGTGTAACTCATATATATATACACACAT--
     **********************  *******.******.****.******.*****.**

Del1 ----ACACACACATACATGAGTATATGTGTGAAACTGTCACTCAAATCAAGACATAGA
GYPB ACACACACACACACATACATGAGTATATGTGTGAAACTGTCACTCAAATCAAGACATAGA
GYPE ----ATACACACACATACATGAGTATATGTGTGAAACTGTCACTCAAATCAAGACATAGA
         ******************************************************

Del1 ACATTTCAGATGTATTCCAGAGATCATGGTGGATGGGAGGCAGGACTGGATTGCAGCTCC
GYPB ACATTTCAGATGTATTCCAGAGATCATGGTGGATGGGAGGCAGGACTGGATTGCAGCTCC
GYPE ACATTTCAGATGTATTCCAGAGATCATGGTGGATGGGAGGCAGGACTGGATTGCAGCTCC
     ************************************************************

Del1 CACTTGAACAGACAAAGCAGCGTGTGGAGGCTTGTATCATGAACTTTGACTGCAGGAATA
GYPB CACTTGAACAGACAAAGCAGCGTGTGGAGGCTTGTATCATGAACTTTGACTGCAGGAATA
GYPE CACTTGGACAGACAAAGCAGCGTGTGGAGGCTTGTATCATGAACTTTGACTGCAGGAATA
     ******.******************************************************

Del1 AATCAGGAAAGCTGAGAGAACCCACAGACCCTCTGAAGGAAGTGGATTGCTCCTGCAGGT
GYPB AATCAGGAAAGCTGAGAGAACCCACAGACCCTCTGAAGGAAGTGGATTGTTCCTGCAGGT
GYPE AATCAGGAACACTGAGAGAACCCACAGACCCTCAGAAAGAAGCGGATTGCTCCTGCAGGT
     ********* .********************** ***.****.******.**********

Del1 CTCAGGAGACACCCCAAATGCTGTGGGAGCCCAAACTGCAAACTGTGGAAGTGGGAAAGG
GYPB CTCAGGAGACACCCCAAATGCTGTGGGAGCCCAAACTGCAAACTGTGGAAGTGGGAAAGG
GYPE CTCAGGAGACACCCCAAATGCTGTGGGAGCCCAAACTGCAAACTGTGGAAGTGGGAAAGG
     ************************************************************

Del1 GGAATAGTCAGCTCCTGAACACACATCCTCACTGGGGAACCTAAAGGTCTAGATCACAGG
GYPB GGAATAGTCAGCTCCTGAACACACATCCTCACTGGGGAACCTAAAGGTCTAGATCACAGG
GYPE GGAATAGTCAGCTCCTGAACACACATCCTCACTGGGGAACCTAAAGGTCTAGATCACAGG
     ************************************************************

Del1 AGAAGATTTTGACCTTACTTGGAGCTGAGTCAATTTANAGAGCCAAGTGACATACACTGC
GYPB AGAAGATTTTGACCTTACTTGGAGCTGAGTCAATTTAGAGAGCCAAGTGACATACACTGC
GYPE AGAAGATTTTGACCTTACTTGGAGCTGAGTCAATTTAGAGAGCCAAGTGACATACACTGC
     ********************************** **********************

Del1 TAGAGAAAGCAGCNCNAAAAGCCCTGTGGGCTCACTGGGTTCCCTAGCCATCNNTTTCTG
GYPB TAGAGAAAGCAGCGCGAAAAGCCCTGTGGGCTCACTGGGTTCCCTAGCCATCCATTTCTG
GYPE TAGAGAAAGCAGGGCGTAAGGCCCTGTGGGCTCACTGGGTTCCCTAGCCATCCATTTCTG
     ************  *  **.*********************************  ******

Del1 NCTTGNCTNACANGGGTCCTTGAGGAGGGCTACCAGAGGCACTGGGAAATGGTCACAAAG
GYPB CCTTGCCTCACAGGGGTCCTTGAGGAGGGCTACCAGAGGCACTGGGAAATGGTCACAAAG
GYPE CCTTGCCTCACGGGGGTCCTTGAGGAGGGCTACCAGAGGCACTGGGAAATGGTCACAAAG
      **** ** **. ********************************************

Del1 AGAAGGAAACTTCCAGCTGAACTTTGTAACAATTTGAACAGATTGAGAAATCTCCTGGCC
GYPB AGAAGGAAACTTCCAGCTGAACTTTGTAACAATTTGAACAGATTGAGAAATCTCCTGGCC
GYPE AGAAGGAAACTTCCAGCTGAACTTTAAAACAATTTGAACAGATTGAGAAGTCTCCTAGCC
     *************************. **********************.******.***
```

**Multiple sequence alignment for deletion DEL2 breakpoints**

```
Del2 GGCCCATGCAAGTCTGANNTNNNNNTGGGGCAGTAATTAAATCTTAAAGCACCTTAATAA
GYPB GGCCCATGCAAGTCTGAAAT-CCAGTGGGGCAGTAATTAAATCTTAAAGCACCTTAATAA
GYPA GGCCCATGCAAGTCTGAAAT-CCAGTGGGGCAGTAATTAAATCTTAAAGCACCTTAATAA
     ****************  *     ************************************

Del2 TCTCTTTGGACTCCATGTCTCACATCCAGTTAATGCTGATGCAAGAGGTGGGCTCCCACA
GYPB TCTCTTTGGACTCCATGTCTCACATCCAGTTAATGCTGATGCAAGAGGTGGGCTCCCACA
GYPA TCTCTTTTAACTCCATGTCTCACATCCATGTAATGCTGATGCAAGAGGTGGGTTCCCACA
     ******* .******************  **********************.*******

Del2 GTCTTGGGAAGCTCCGCTCCTGTGGCTTTGCATAGTACAACCCCCCTCTCGGCTGCTTTC
GYPB GTCTTGGGAAGCTCCGCTCCTGTGGCTTTGCATAGTACAACCCCGCTCTCGGCTGCTTTC
GYPA GTCTTGGGAAGCTCTGCTCCTGTGGCTTTGCATGGTACAACCCCCCTCTCGGCTGCTTTC
     **************.******************.********** ***************

Del2 ACAGGCTGTCTGTTATCCAGTTCCAAAGTCACTTCTGCATTTTTAGGTATCCTTATAGCA
GYPB ACAGGCTGTCTGTTATCCAGTTCCAAAGTCACTTCTGCATTTTTAGGTATCCTTATAGCA
GYPA ACGGG----CTGTTATCCAATTCCAAAGTCACTTCTGCATTTTTAGGTATCCTTATAGCA
     **.**    **********.****************************************

Del2 GCACCCCACCTCTAGTACCAACTTACTGTATTAGTCTGTTCTC-TGCTGCTATAAAAAAC
GYPB GCACCCCACCTCTAGTACCAACTTACTGTATTAGTCTGTTCTCATGCTGCTATAAAAAAC
GYPA GCACCCCACCTCTAGTACCAACTTACTGTATTAGTCTGT-CTCATGCTGCTATAAAAAAC
     ****************************************.*** ***************

Del2 TGCGCAAGACTGTGTAATTTATAAAGGAAAGAGGTTTAGTTGATCTACAGTTTTGCATGG
GYPB TGCCCAAGACTGTGTAATTTATAAAGGAAAGAGGTTTAGTTGATCTACAGTTTTGCATGG
GYPA TTCCCAAGACTGTGTAATTTAT-----AAAGAGGTTTAATTGATCTACAGTTTTGCATGG
     * * ******************     ***********.*********************

Del2 CTTGGAAGGTCTTAGAATACTTACAATCATGACCAAAGGGGAAACAAACACATCTTTCTT
GYPB CTTGGAAGGTCTTAGGATACTTACAATCATGACCAAAGGGGAAACAAACACATCTTTCTT
GYPA CTGGGAAGGTCTCAGGATACTTACAATCATGACCAAAGGGGAAACAAACACATCTTTCTT
     ** **********.**.***********************************************

Del2 ACATAGTGGCAGGAAGGAGAAGAATGAGAGCTGAGTGAAGGGGGAAGCTCCTTTATAAAA
GYPB ACATAGTGGCAGGAAGGAGAAGAATGAGAGCTGAGTGAAGGGGGAAGCTCCTTTATAAAA
GYPA ACATAGTGGCAGGAAGGAGAAGAATGAGAGCTGAGTGAAGGGGGAAGCTCCTTTATAAAA
     ************************************************************

Del2 CTATCAGATTATGTGAGAATTTATTCACTCTCATGAGAATAGCATAGGGGAAACCACCGC
GYPB CTATCAGATTATGTGAGAATTTACTCACTATCATGAGAATAGCACAGGGGAAACCACCGC
GYPA CTATCAGATTATGTGAGAATTTATTCACTCTCATGAGAATAGCATAGGGGAAACCACCGC
     ************************.***** **************.***************

Del2 AATGATTCAAGTACCTCCCACTGGGTTCCTCCCATGACACGTGGGGATTATTGGAACTAC
GYPB AATGATTCAAGTACCTCCCACTGGGTTCCTCCCATGACATGTGGGGATTATTGGAACTAC
GYPA AATGATTCAAGTACCTCCCACTGGGTTCCTCCCATGACACGTGGGGATTATTGGAACTAC
     ************************************.***********************

Del2 AATTCAAGATGAGATTTGAGTGGGAACACAGCCAAACCATATCAGTCATTCCACATATTG
GYPB AATTCAAGATGAGATTTGAGTGGGAACACAGTCAAACCATATTAGTCATTCCACATATTG
GYPA AATTCAAGATGAGATTTGAGTGGGAACACAGCCAAACCATATCAGTCATTCCACATATTG
     *******************************.***********.*****************

Del2 AGTGATTTCTCTCTTCTATTCACTATTCTTCCACAGAGAGGGACCTGTAGTCATTACCTT
GYPB AGTGATTTCTCTCTTCTATTCACTATTCTTCCACAGAGAGGGACCCATAGTCATTACCTT
GYPA AGTGATTTCTCTCTTCTATTCACTATTCTTCCACAGAGAGGGACCTGTAGTCATTACCTT
     ********************************************..*************

Del2 CAAGGAACCTAAATCCTGGTGTCTTTATGGTTATGGGTGGCATATAATATACAGATAAGC
GYPB CAAGGAACCTAAATCCTGGCATCTTTATTGTTATGGGTGGTGTATAATATACAGGTAAGC
GYPA CAAGGAACCTAAATCCTGGTGTCTTTATGGTTATGGGTGGCATATAATATACAGGTAAGC
     ******************..******* ***********..************.*****
```

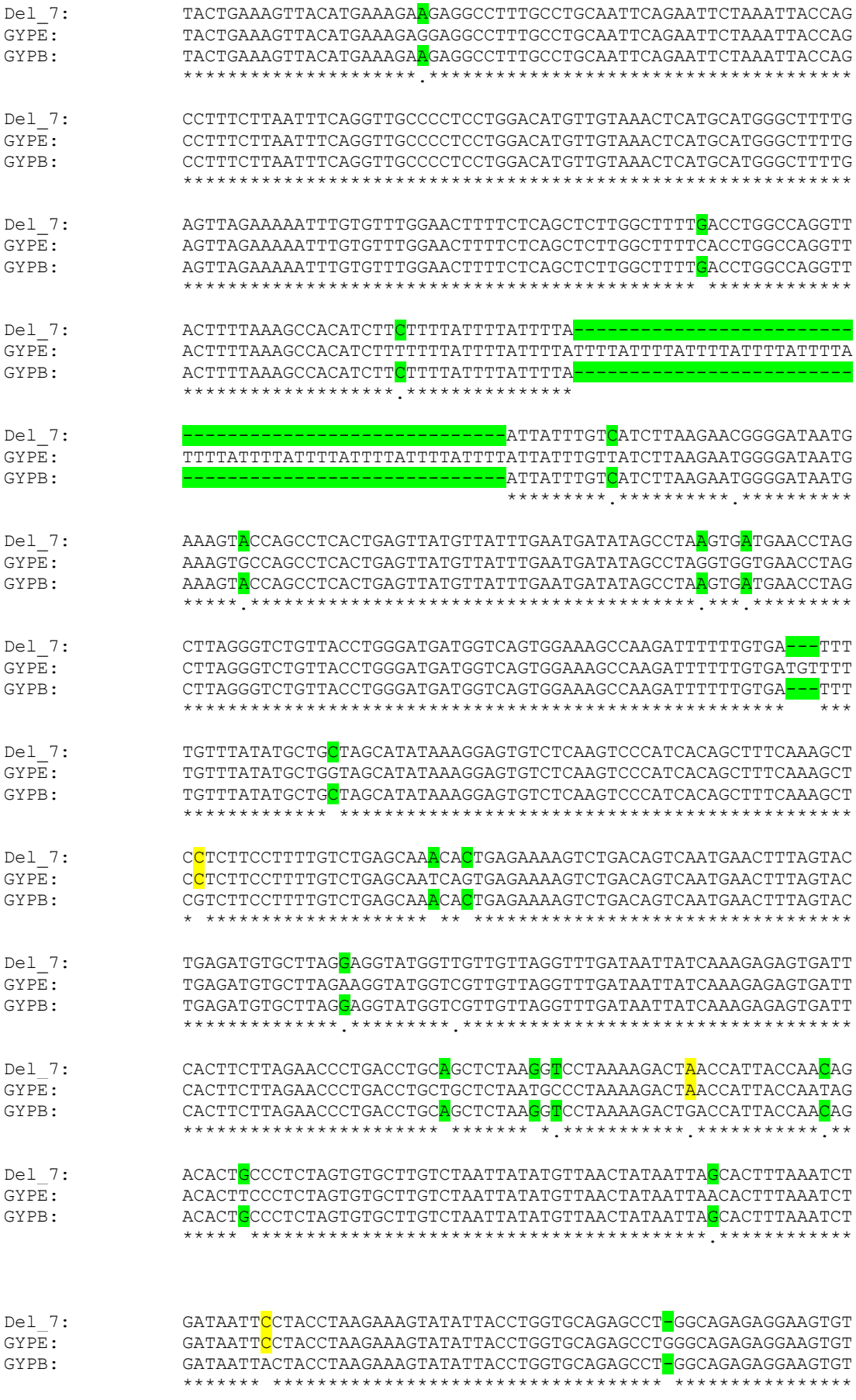
**Multiple sequence alignment for deletion DEL6 breakpoints**

```
Del6:           TTGGTGTAATGTACTCATTACCCTAAAACCTATATTATAATGGACCCGAATACCATAAGG
GYPA:           TTGGTGTAATGTACTCATTACCCTAAAACCTATATTTTAATGGACCTGAAGACCATAAGG
GYPE:           TTGGTGTAATGTACTCATTACCCTAAAACCTATATTATAATGGACCCGAATACCATAAGG
                ********************************** ********.*** *********

Del6:           TAGAAGATCAGAAGAAAAAGAAAATGCATACATTTTTATCTTTAAAACAATTCACTTTTA
GYPA:           TAGAAGATCAGAAGAAAAAGAAAATGCATACATTTTTATCTTTAAAACAATTCACTTTTA
GYPE:           TAGAAGATCAGAAGAAAAAGAAAATGCATACATTTTTATCTTTAAAACAATTCACTTTTA
                ************************************************************

Del6:           AGATTATTTTCTTGAAAGGTATACGTTTGATGGAGCTGAATCTAGTTCCTTCAGGCAGAT
GYPA:           AGATTATTTTCTTGAAAGGTATACCTTTGATGGAGCCGAATCTGGTTCCTTCAGGCAGAT
GYPE:           AGATTATTTTCTTGAAAGGTATACGTTTGATGGAGCTGAATCTAGTTCCTTCAGGCAGAT
                ********************** **********.******.****************

Del6:           ATAATAGGATACAATAGGATAGGACTTCTTGGGTAAAAAAGGTGTAGTAATCTGTGAGAC
GYPA:           ACAATAGGATACAATAGGATAGGACTTCTTGGGTAAAAAAGGTGTAGTAATCTGTGAGTC
GYPE:           ATAATAGGATACAATAGGATAGGACTTCTTGGGTAAAAAAGGTGTAGTAATCTGTGAGAC
                *.********************************************************* *

Del6:           TCATGGGAAGGAGAAAGGAAATTATTAATTAATTTGGCTGAGGCCAGGGTAGTAAGCAGA
GYPA:           TCACCGGAAGGAGAAAGGAAATTATTCATTAATTTGGCTGAGGCCAGGGTAGTAAGCAGA
GYPE:           TCATGGGAAGGAGAAAGGAAATTATTAATTAATTTGGCTGAGGCCAGGGTAGTAAGCAGA
                ***. ********************* **********************************

Del6:           TTCATTATAAATTCTGGCAGCAGAGGAATACTTGAGAAATGCTTATAAAGTTAGATCATC
GYPA:           TTTATTATAAATATTGGCAGCAGAGGAGTACTTGAAAAGTGCTTATAAAGTTAGATCATC
GYPE:           TTCATTATAAATTCTGGCAGCAGAGGAATACTTGAGAAATGCTTATAAAGTTAGATCATC
                **.********* .*************.*******.**.**********************

Del6:           TATTTTTCCAAGCAATATTCTTCATGTCATTCCACCTTGTAAAAACTTTGTGGAAAAGAC
GYPA:           TATTTTTCCAAGCAATATTCTTCATGTCATTCCACCTTGCAAAAACTTTGTGGAAAGGAA
GYPE:           TATTTTTCCAAGCAA-ATTCTTCATGTCATTCCACCTTGTAAAAACTTTGTGGAAAAGAA
                ************** ************************.***************.**

Del6:           TAAGCACATGAATGAGCAATGAATAGACTTTGCATTAGATTGGTAATCTTGCTACTAGTA
GYPA:           TAGGCATATGAATGAGCAATGAATAGACTTTGCATTAGATTGGTAATCTTGCTACTAGTA
GYPE:           TAAGCACATGAATGAGCAATGAATAGACTTTGCATTAGATTGGTAATCTTGCTACTAGTA
                **.***.************************************************************

Del6:           CATTAATTTACTTGTATTACTGATATTTTAAGATAGGATTTTTACAGATGTTTATATATT
GYPA:           CATTAATTTACTTGTATTACTGATATTTTAAGATAGGATTTTTACAGATGTTTATATATT
GYPE:           CATTAATTTACTTGTATTACTGATATTTTAAGATAGGATTTTTACAGATGTTTATATATT
                ************************************************************

Del6:           TTTTAAAAGCCTATTTTTTTCTTTACCCATTTTCTCAATTTTTTTAACCTCGAAAACTTT
GYPA:           TTTAAAAAGCCTATTTTTTTCTTTACCCATTTTCTCAATTTTTTTAACCTCGAAAACTTT
GYPE:           TTTTAAAAGCCTATTTTTTTCTTTACCCATTTTCTCAATTTTTTTAACCTCGAAAACTTT
                *** ************************************************************

Del6:           TTCCTCCCTGAGAGAAATACATGGGCTATCTTCAGCACAGAACACAAACTCTGCCTTCTT
GYPA:           TTCCTCCCTGAGAGAAATACATGGGCTATCTTCAGCACAGAACACAAACTCTGCCTTCTT
GYPE:           TTCCTCCCTGAGAGAAATACATGGGCTATCTTCAGCATAGAACACAAACTCTGCCTTCTT
                ************************************.***********************

Del6:           TTCTCAAGTCCAGTGATTCAAGTTTGACACACCACACCACGGTGGATGTGCCTT-TCATG
GYPA:           TTCTCAAGTCCAGTGATTCAAGTTTGACACACCACACCACGGTGGATGTGCCTT-TCATG
GYPE:           TTCTCAAGTCCAGTGATTCAAGTTTGACACACCACACCACGGTGGATGTGCCTTGTCATG
                ************************************************** *****

Del6:           ACGCATCTTCTCAGCAGCACTAGAAGATACATCCAAAGTCCTGTGAGCAAGGTAAGCCTT
GYPA:           ACGCATCTTCTCAGCAGCACTAGAAGATACATCCAAAGTCCTGTGAGCAAGGTAAGCCTT
GYPE:           ACACA---TCTCAGCAGCACTAGAAGATACATCCAAAGTCCTGTGAGCAAGGTAAGCCTT
                **.**   ************************************************
```

```
Del6:           GTAGCAACCCTTTGATGTCTTTGAAAAATGCAAAAATGCTCACCGATCAGGAGCAAAGGG
GYPA:           GTAGCAACCCTTTGATGTCTTTGAAAAATGCAAAAATGCTCACCGATCAGGAGCAAAGGG
GYPE:           GTAGCAACCCTTTGATGTCTTTGAAAAATGCAAAAATGCTCACCGATCAGGAGCAAAGGG
                ************************************************************

Del6:           GCCAAAGTTGAGGTATTTCCCCCTCACATGTAGTTTTTATTTTTTTAGCCTATTTTTTGT
GYPA:           GCCAAAGTTGAGGTATTTCCCCCTCACATGTAGTTTTTATTTTTTTAGCCTATTTTTTGT
GYPE:           GCCAAAAGAGAAGTATTTCCCCCTCACATGTAGTTTTTATTTTTTTAGCGTATTTTTTGT
                ******.  **.*********************************** **********

Del6:           TTTTCTCTAAATCAAACTCAGACTTACTCTGTTTACACTGAAATACACTGCACAGTGCCG
GYPA:           TTTTCTCTAAATCAAACTCAGACTTACTCTGTTTACACTGAAATACACTGCACAGTGCCG
GYPE:           TTTTCTCTAAATCAAACTCAGACTTACTCTGTTTACACTGAAATACACTGCATAGTGCCA
                ************************************************.******.

Del6:           TAGGTGTAATCAGATACAGATCTCAACAATGTGTATGATTCCATACCCAGTATAAATAAG
GYPA:           TAGGTGTAATCAGATACAGATCTCAACAATGTGTATGATTCCATACCCAGTATAAATAAG
GYPE:           TAGGTGTAATCAGACCCAGATCTCAACAATGTGTATGATTGCATACCCAGTATAAATAAG
                **************. ************************ ******************
```

**Multiple sequence alignment for deletion DEL7 breakpoints**

```
Del_7:          CTTTGTGGAAAAGAATAAGCACATGAATGAGCAATGAATAGACTTTGCATTAGATTGGTA
GYPE:           CTTTGTGGAAAAGAATAAGCACATGAATGAGCAATGAATAGACTTTGCATTAGATTGGTA
GYPB:           CTTTGTGGAATAGAATAGGCATATGAA-GAGCAATGAATAGACTTTGCATTAGATTGGTA
                ********** ******.***.***** ********************************

Del_7:          ATCTTGCTACTAGTACATTAATTTACTTGTATTACTGATATTTTAAGATAGGATTTTTAC
GYPE:           ATCTTGCTACTAGTACATTAATTTACTTGTATTACTGATATTTTAAGATAGGATTTTTAC
GYPB:           ATCTTGCTACTAGTACATTAATTTACTTGTATTACTGATATTTTAAGATAGGATTTTTAC
                ************************************************************

Del_7:          AGATGTTTATATATTTTTTAAAAGCCTATTTTTTTCTTTACCCATTTTCTCAATTTTTTT
GYPE:           AGATGTTTATATATTTTTTAAAAGCCTATTTTTTTCTTTACCCATTTTCTCAATTTTTTT
GYPB:           AGATGTTTATATATTTTTTAAAAGCCTATTTTTTTCTTTACCCACTTTCTCAATTTTTTT
                ******************************************.***************

Del_7:          AACCTCGAAAACTTTTTCCTCCCTGAGAGAAATACATGGGCTATCTTCAGCACAGAACAC
GYPE:           AACCTCGAAAACTTTTTCCTCCCTGAGAGAAATACATGGGCTATCTTCAGCATAGAACAC
GYPB:           AACCTCGAAAACTTTTTCCTCCCTAAGAGAAATACATGGGCTATCTTCAGCACAGAACAC
                ************************.**************************.********

Del_7:          AAACTCTGCCTTCTTTTCTCAAGTCCAGTGATTCAAGTTTGACACACCACACCACGGTGG
GYPE:           AAACTCTGCCTTCTTTTCTCAAGTCCAGTGATTCAAGTTTGACACACCACACCACGGTGG
GYPB:           AAACTCTGCCTTCTTTTCTCAAGTCCAGTGATTCAAGTTTGACACACCACACCACGGTGG
                ************************************************************

Del_7:          ATGTGCCTTGTCATGACACA---TCTCAGCAGCACTAGAAGATACATCCAAAGTCCTGTG
GYPE:           ATGTGCCTTGTCATGACACA---TCTCAGCAGCACTAGAAGATACATCCAAAGTCCTGTG
GYPB:           ATGTGCCTTGTCATGACACATCTTCTCAGCAGCACTAGAAGATACATCCAAAGTCCTGTG
                ********************   ************************************

Del_7:          AGCAAGGTAAGCCTTGTAGCAACACTTTGATGTCTTTGAAAAATGCAAAAATGCTCACCG
GYPE:           AGCAAGGTAAGCCTTGTAGCAACCCTTTGATGTCTTTGAAAAATGCAAAAATGCTCACCG
GYPB:           AGCAAGGTAAGCCTTGTAGCAACCCTTTGATGTCTTTGAAAAATGCAAAAATGCTCACCG
                ********************** ************************************

Del_7:          ATCAGGAGGAAAGGGGCCAAAGGAGAGGTATTTCCCCCTCACATGTAGTTTTTATTTTTT
GYPE:           ATCAGGAGCAAAGGGGCCAAAAGAGAAGTATTTCCCCCTCACATGTAGTTTTTATTTTTT
GYPB:           ATCAGGAGCAAAGGGGCCAAAGGAGAGGTATTTCCCCCTCACATGTAGTTTTTATTTTTT
                ******** ************.****.**********************************

Del_7:          TAGCGTATTTTTTGTTTTTCTCTAAATCAAACTCAGACTTACTCTGTTTACACTGAAATA
GYPE:           TAGCGTATTTTTTGTTTTTCTCTAAATCAAACTCAGACTTACTCTGTTTACACTGAAATA
GYPB:           TAGCGTATTTTTTGTTTTTCTCTAAATCAAACTCAGACTTACTCTGTTTACACTGAAATA
                ************************************************************

Del_7:          CACTGCATAGTGCCGTAGGTGTAATCAGACCCAGATCTCAACAATGTGTATGATTCCATA
GYPE:           CACTGCATAGTGCCATAGGTGTAATCAGACCCAGATCTCAACAATGTGTATGATTGCATA
GYPB:           TACTGCACAGTGCCATAGGTGTAATCAGACCCAGATCTCAACAATGTGTATGATTCCATA
                .******.******.************************************** ****

Del_7:          CCCAGTATAAATAAGGGTTTAATATCAACTCCATCTCTGAGTGAGAATGAATAGCTCTGA
GYPE:           CCCAGTATAAATAAGGGTTTAATATCAACTCCATCTCTGAGTGAAAATAAATAGCTCTGA
GYPB:           CCCAGTATAAATAAGGGTTTAATATCAACTCCATCTCTGAGTGAGAATGAATAGCTCTGA
                ******************************************.***.************
Del_7:          AAAGTCAATTCTTCCTATTTTTCTTTTACATTTGGCCTTGTAATAGCTGTTTTATGTCAT
GYPE:           AAAGTCAATTCTTCCTA-TTTTCTTTTACATTTTGCCTTCTAATAGCTGTTTTATGTCAT
GYPB:           AAAGTCAATTCTTCGTATTTTTCTTTTACATTTGGCCTTGTAATAGCTGTTTTATGTCAT
                ************** ** *************** ***** ********************

Del_7:          TTCCTTTGAGAGCTCTGATACCAAGTTGGAATAGTCCATTATATCCTAGCTCTCAGTCTT
GYPE:           TTCCTTTGAGAGCTCTGATACCAAGTTGGAATAGTCCATTATATCCTAGCTCTCAGTCTT
GYPB:           TTCCTTTGAGAGCTCTGATACCAAGTTGGAATAGTCCATTATATCCTAGCTCTCAGTCTT
                ************************************************************
```


Del_7: TACTGAAAGTTACATGAAAGAAGAGGCCTTTGCCTGCAATTCAGAATTCTAAATTACCAG
GYPE: TACTGAAAGTTACATGAAAGAGGAGGCCTTTGCCTGCAATTCAGAATTCTAAATTACCAG
GYPB: TACTGAAAGTTACATGAAAGAAGAGGCCTTTGCCTGCAATTCAGAATTCTAAATTACCAG
Del_7: CCTTTCTTAATTTCAGGTTGCCCCTCCTGGACATGTTGTAAACTCATGCATGGGCTTTTG
GYPE: CCTTTCTTAATTTCAGGTTGCCCCTCCTGGACATGTTGTAAACTCATGCATGGGCTTTTG
GYPB: CCTTTCTTAATTTCAGGTTGCCCCTCCTGGACATGTTGTAAACTCATGCATGGGCTTTTG
Del_7: AGTTAGAAAAATTTGTGTTTGGAACTTTTCTCAGCTCTTGGCTTTTGACCTGGCCAGGTT
GYPE: AGTTAGAAAAATTTGTGTTTGGAACTTTTCTCAGCTCTTGGCTTTTCACCTGGCCAGGTT
GYPB: AGTTAGAAAAATTTGTGTTTGGAACTTTTCTCAGCTCTTGGCTTTTGACCTGGCCAGGTT
Del_7: ACTTTTAAAGCCACATCTTCTTTTATTTTATTTTA----------
GYPE: ACTTTTAAAGCCACATCTTTTTTTATTTTATTTTATTTTATTTTATTTTATTTTA
GYPB: ACTTTTAAAGCCACATCTTCTTTTATTTTATTTTA----------
Del_7: ----------ATTATTTGTCATCTTAAGAACGGGGATAATG
GYPE: TTTTATTTTATTTTATTTTATTTTATTTTATTATTTGTTATCTTAAGAATGGGGATAATG
GYPB: ----------ATTATTTGTCATCTTAAGAATGGGGATAATG
Del_7: AAAGTACCAGCCTCACTGAGTTATGTTATTTGAATGATATAGCCTAAGTGATGAACCTAG
GYPE: AAAGTGCCAGCCTCACTGAGTTATGTTATTTGAATGATATAGCCTAGGTGGTGAACCTAG
GYPB: AAAGTACCAGCCTCACTGAGTTATGTTATTTGAATGATATAGCCTAAGTGATGAACCTAG
Del_7: CTTAGGGTCTGTTACCTGGGATGATGGTCAGTGGAAAGCCAAGATTTTTTGTGA---TTT
GYPE: CTTAGGGTCTGTTACCTGGGATGATGGTCAGTGGAAAGCCAAGATTTTTTGTGATGTTTT
GYPB: CTTAGGGTCTGTTACCTGGGATGATGGTCAGTGGAAAGCCAAGATTTTTTGTGA---TTT
Del_7: TGTTTATATGCTGCTAGCATATAAAGGAGTGTCTCAAGTCCCATCACAGCTTTCAAAGCT
GYPE: TGTTTATATGCTGGTAGCATATAAAGGAGTGTCTCAAGTCCCATCACAGCTTTCAAAGCT
GYPB: TGTTTATATGCTGCTAGCATATAAAGGAGTGTCTCAAGTCCCATCACAGCTTTCAAAGCT
Del_7: CCTCTTCCTTTTGTCTGAGCAAACACTGAGAAAAGTCTGACAGTCAATGAACTTTAGTAC
GYPE: CCTCTTCCTTTTGTCTGAGCAATCAGTGAGAAAAGTCTGACAGTCAATGAACTTTAGTAC
GYPB: CGTCTTCCTTTTGTCTGAGCAAACACTGAGAAAAGTCTGACAGTCAATGAACTTTAGTAC
Del_7: TGAGATGTGCTTAGGAGGTATGGTTGTTGTTAGGTTTGATAATTATCAAAGAGAGTGATT
GYPE: TGAGATGTGCTTAGAAGGTATGGTCGTTGTTAGGTTTGATAATTATCAAAGAGAGTGATT
GYPB: TGAGATGTGCTTAGGAGGTATGGTCGTTGTTAGGTTTGATAATTATCAAAGAGAGTGATT
Del_7: CACTTCTTAGAACCCTGACCTGCAGCTCTAAGGTCCTAAAAGACTAACCATTACCAACAG
GYPE: CACTTCTTAGAACCCTGACCTGCTGCTCTAATGCCCTAAAAGACTAACCATTACCAATAG
GYPB: CACTTCTTAGAACCCTGACCTGCAGCTCTAAGGTCCTAAAAGACTGACCATTACCAACAG
Del_7: ACACTGCCCTCTAGTGTGCTTGTCTAATTATATGTTAACTATAATTAGCACTTTAAATCT
GYPE: ACACTTCCCTCTAGTGTGCTTGTCTAATTATATGTTAACTATAATTAACACTTTAAATCT
GYPB: ACACTGCCCTCTAGTGTGCTTGTCTAATTATATGTTAACTATAATTAGCACTTTAAATCT
Del_7: GATAATTCCTACCTAAGAAAGTATATTACCTGGTGCAGAGCCT-GGCAGAGAGGAAGTGT
GYPE: GATAATTCCTACCTAAGAAAGTATATTACCTGGTGCAGAGCCTGGGCAGAGAGGAAGTGT
GYPB: GATAATTACTACCTAAGAAAGTATATTACCTGGTGCAGAGCCT-GGCAGAGAGGAAGTGT

## Multiple sequence alignment for duplication DUP5 breakpoints

- **First breakpoints**

```
Dup5              201 TCCAGTCTCCTTTATCATGTTCTGGCTAGCAGAATGAAGGTGGCTATGTT    250
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
non_Coding        201 TCCAGTCTCCTTTATCATGTTCTGGCTAGCAGAATGAAGGTGGCTATGTT    250

Dup5              251 ACATGCCTCCTTCAGAACCACCCTACAAATATATCCTGTGTT    292
                      |||||||||||||||||||||||||||||||||||||||||||
non_Coding        251 ACATGCCTCCTTCAGAACCACCCTACAAATATATCCTGTGTT    291


Dup5:             ----------------CACCCACTGTGAACTCCTTTGCAGAAAAGGTCTGAATAGAGGGGA
GYPB:             ----------------CACCCACTGTGAACTCCTTTGCAGAAAAGGTCTGAATAGAGGGGA
GYPE:             ----------------CACCCACTGTGAACTCCTTTGCAGAAAAGGTCTGAATAGAGGGGA
                                  *****************************************

Dup5:             AAGTAGGGATGGTATCTCAAACTTACTTCGTAGTGATTTTAAATTAGGAAATTTAGCTTC
GYPB:             AAGTAGGGATGGTATCTCAAACTTACTTCGTAGTGATTTTAAATTAGGAAATTTAGCTTC
GYPE:             AAGTAGGGATGGTATCTCAAACTTACTTCGTAGTGATTTTAAATTAGGAAATTTAGCTTC
                  ************************************************************

Dup5:             ACATTCTTGTGATAAATTTCTTTTCACCTTGGTTTCTAGAAGATTATTCAAAACATCTAT
GYPB:             ACATTCTTGTGATAAATTTCTTTTCACCTTGGTTTCTAGAAGATTATTCAAAACATCTAT
GYPE:             ACATTCTTGTGATAAATTTCTTTTCACCTTGGTTTCTAGAAGATTATTCAAAACATCTAT
                  ************************************************************

Dup5:             GAGACTATTTGAGAAGTATACTTTTGGGGAATTTCCCCCAAGTTATCTTTATAGATTATA
GYPB:             GAGACTATTTGAGAAGTATACTTTTGGGGAATTTCCCCCAAGTTATCTTTATAGATTATA
GYPE:             GAGACTATTTGAGAAGTATACTTTTGAGGAATTTCCCCCAAGTTATCTTTACAGATTATA
                  *************************.************************.*********
```

- **Second breakpoints**

```
Dup5:             TATGTTTTCTTCTAAGAGTTTTGTATTTCTTCTCTTATATTTAGATCTTTGGTTTATTAT
GYPE:             TATGATTTCTTCTAAGAGTTTTGTAGTTCTTCTCTTATATTTAGATCTTTGGTTTATTAT
GYPB:             TATGTTTTCTTCTAAGAGTTTTGTATTTCTTCTCTTATATTTAGATCTTTGGTTTATTAT
                  **** ******************** **********************************

Dup5:             CAGTTAATTTTTCTATATGATGTATGATAAGAGTCCACCTTTATTATTTTGCAGCTGTCC
GYPE:             CAGTTAATTTTTCTGTATGATGTATGATAAGAGTCCACCTTTATTATTTTGCAGCTGTCC
GYPB:             CAGTTAATTTTTCTATATGATGTATGACAAGAGTCCACCTTTATTATTTTGCAGTTGTCC
                  **************.************.************************.*****

Dup5:             CAGCACCATTTGTTGAAGAGACTATCCTTTGCCCATTGAATGGTCTTGACACCCTTCTTG
GYPE:             CAGCACCATTTGTTGAAGAGACTATCCTTTGCCCATTGAATGGTCTTGACACCCTTCTTG
GYPB:             CAGCACCATTTGTTGAAGAGACTATCCTTTGCCCATTGAATGGTCTTGACACCCTTCTTG
                  ************************************************************

Dup5:             AAAGTTAATTGGCCATGGATATATGAGTTTATTTCTGGAGTCTCAATTCTATCCGAAGAA
GYPE:             AAAGTTAATTGGCCATGGATATATGAGTTTATTTCTGGAGTCTCAATTCTATCCGAAGAA
GYPB:             AAAGTTAATTGGCCATGGATATATGAGTTTATTTCTGGAGTCTCAATTCTATCCGAAGAA
                  ************************************************************

Dup5:             TATGTCTATTCTTGGGGCAAAATCACACAGATTTTATTGCTGTTACTTGGTTAGAAGTTT
GYPE:             TATGTCTGTTCTTGGGGCAAAATCACACAGATTTTATTGCTGTTACTTGGTTAGAAGTTT
GYPB:             TATGTCTGTTCTTGGGGCAAAATCACACAGATTTTATTGCTGTTACTTGGTTATAAGTTT
                  *******.****************************************** ******

Dup5:             TGATTTCATGAAGTGTGATTCACTGAACTTTGTTCTTCTTCCAGATTGTTTTAGCTATTT
GYPE:             TGAATTCATGAAGTGTGATTCACTGAACTTTGTTCTTCTTCCAGATTGTTTTAGCTATTT
GYPB:             TGCATTCATGAAGTATGATTCACCAAACTGTGTTCTTCTTCAAGATTGTTTTGGCTATTT
                  **  **********.********..**** ********** **********.*******
```

- **Third breakpoints**

```
Dup5:           GCACTGAAATTGCATTTTGCTAGGAAAAAGACCACAAAAGTTCTCCCCTTGCTACCTTTC
GYPB:           GCACTGAAATTGCATTTTGCTAGGAAAAAGACCACAAAAGTTCTCCCCTTGCTACCTTTC
GYPE:           GCACTGAAATTGCATTTTGCTAGGAAAAAGACCACAAAAGTTCTCCCCTTGCTACCTTTC
                ************************************************************

Dup5:           CTGAACTATTCTGCTAGATTCAGACCTCAGAAACATTGTATCAGGAAATACAGAAATGTT
GYPB:           CTGAACTATTCTGCTAGATTCAGACTTCAAAAACATTGTATCAGGAAATACAGAAATGTT
GYPE:           CTGAACTATTCTGCTAGATTCAGACCTCAGAAACATTGTATCAGGAAATACAGAAATGTT
                *************************.***.******************************

Dup5:           CTTTCAAAATGAGTGTATGGGAATGTGGGAATGCCTAATAAAATCTGTCCTCATTGATTC
GYPB:           CTTTCAAAATGAGTGTATGGGAATGTGGGAATGCCTAATAAAATCTGTCCTTATTGATTC
GYPE:           CTTTCAAAATGAGTGTATGGGAATATGGGAATGCCTAATAAAATCTGTCTTCATTGATTC
                ************************.************************.*.********

Dup5:           GTTAGCAAAAAATCATATAAATCAATAGCTTGTGATTGCAAGCAGATATATTTCAGATCC
GYPB:           GTTAGCAAAAAATCATATAAATCAATAGCTTGTGATTGCAAGCAGATATATTTCAGATCC
GYPE:           GTTAGCAAAAAATCATATAAATCCATAGCTTGTGATTGCAAGCAGATATATTTCAGATCC
                ********************** ************************************

Dup5:           TTTCTGTGTTTGTTTTTTGCTTTCTTGATCTATCACAATTGGAGAAAACTTAAAATTTCT
GYPB:           TTTCTGTGTTTGTTTTTTGCTTTCTTGATCTATCACAATTGGAGAAAACTTAAAATTTCT
GYPE:           TTTCTGTGTTTGTTTTTTGCTTTCTTGATCTATCACAATTGGAGAAAACTTAAAATTTCT
                ************************************************************

Dup5:           CAATGGTATTGTATTTTTGCCAATTTCTTATTCTGCTTTATGTTTCTCGTTGCTATATTA
GYPB:           CAATGGTATTGTATTTTTGCCAATTTCTTATTCTGCTTTATGTTTCTCGTTGCTATATTA
GYPE:           CAATGGTATTGTATTTTTGCCAATTTCTTATTCTGCTTTATGTTTCTCGTTGCTATATTA
                ************************************************************

Dup5:           TTGGGCTATAATGGTCCATAATTACTTAAGAATCATTGTGAAATATATTGCTTAATGACA
GYPB:           TTGGGCTATAATGGTCCATAATTACTTAAGAATCATTGTGAAATATATTGCTTAATGACA
GYPE:           TTGGGCTAAAATGGTCCATAATTACTTAAGAATCATTGTGAAATATATTGCTTAATGACA
                ******** ***************************************************

Dup5:           CAAGTAAATCTTTTTCACTGTTTGTAATGTCTTTTCTCTTAATTCTACTTTGCCTAAGAT
GYPB:           CAAGTAAATCTTTTTCACTGTTTGTAATGTCTTTTCTCTTAATTCTACTTTGCCTAAGAT
GYPE:           CAAGTAAATCTTTTTCACTGTTTGTAATGTCTTTGCTCTTAATTCTACTTTGCCTAAGAT
                ********************************* ************************
```

**Multiple sequence alignment for duplication DUP14 breakpoints**

```
Dup_B:          TTTGAAGCAAGAGATGGTGTTAATGTTAAGACAGTCTAGGAAACAGACAGACTCCTAGTT
GYPE:           TTTGAAGCAAGAGATGGTGTTAATGTTAAGACAGTCTGGGAAACAGACAGACTCCTAGTT
GYPB:           TTTGAAGCAAGAGATGGTGTTAATGTTAAGACAGTCTAGGAAACAGACAGACTCCTAGTT
                ************************************.**********************

Dup_B:          AGCATTGTAATTATGAGTTCTTAATATCACCTAAGATGCATAAGATCTCACCATCCCAAA
GYPE:           AGAATGGTAATTATGAATTCTTAATGTCACCTAAGACCCATAAGATCTCACAATCCCAAA
GYPB:           AGCATTGTAATTATGAGTTCTTAATATCACCTAAGATGCATAAGATCTCACCATCCCAAA
                ** ** **********.********.**********. ************* ********

Dup_B:          CAAAAAATAGGTATCGTTTCTAATGTGTATTTTCTTCTCTTTATGCTACTTATCTTTTTA
GYPE:           CAAAAAATAAATATCATTTCTAATGTGTATTTTCTTCTCTTTATGCTAATTATCTTTTTA
GYPB:           CAAAAAATAGGTATCGTTTCTAATGTGTATTTTCTTCTCTTTATGCTACTTATCTTTTTA
                *********..****.********************************* ***********

Dup_B:          CCCAATCAACAGTCTTATTTCATCATATATCAAATATAATTTTCCCATATTGTTGATTGC
GYPE:           CCCAGTCAACAGTCTTACTTCATCATATATCAAATATAATTTTCCCATATTATTGATTGC
GYPB:           CCCAATCAACAGTCTTATTTCATCATATATCAAATATAATTTTCCCATATTGTTGATTGC
                ****.************.*********************************.********

Dup_B:          AGACTATAAGTCAGATATTACTTTGAGGCTTTGGTTTTTACATCTTTGATAGAAGATATA
GYPE:           AGACTACAAGTCAGATATTATTGTGAGGCTTTGGTTTTTACATCTTTGATAGAAGATATA
GYPB:           AGACTATAAGTCAGATATTACTTTGAGGCTTTGGTTTTTACATCTTTGATAGAAGATATA
                ******.*************.* ************************************

Dup_B:          ATTTAACCTCAGTTTTACCTACTGGCTGGTGCCTGTCAACGTATTGCTTAAAGAACATCC
GYPE:           ATTTAACCTCAGTTTTACCTACTGGCTGTTGCCTGTCAACGTATTGCTTAAAGAACATCT
GYPB:           ATTTAACCTCAGTTTTACCTACTGGCTGGTGCCTGTCAACGTATTGCTTAAAGAACATCC
                **************************** ******************************.

Dup_B:          AACTTACAAGGGCTGTGAAGGACCTCTTCAAGGAGACCTACAAACCACTGCTCAATGAAG
GYPE:           AACTTACAAGGGATGTGAAGGACCTCTTCAAGGAGACCTACAAACCACTGCTCAATGAAG
GYPB:           AACTTACAAGGGCTGTGAAGGACCTCTTCAAGGAGACCTACAAACCACTGCTCAATGAAG
                ********** ************************************************

Dup_B:          TAAAAGAGGACACAAACAAATGGAAGAACATTCCATGCTCATGGATAGAAAGAATCAATA
GYPE:           TAAAAGAGGACACAAACAAATGGAAGAACATTCCATGCTCATGGATAGAAAGAATCAATA
GYPB:           TAAAAGAGGACACAAACAAATGGAAGAACTTTCCATGCTCATGGATAGGAAGAATCAATA
                ***************************** ******************.************

Dup_B:          TTGTGAAAATGGCCATACTGCCCAAGGTAATTTATAGATTCAATGCCATCCCCATCAAGC
GYPE:           TTGTGAAAATGGCCATACTGCCCAAGGTAATTTATAGATTCAATGCCATCCCCATCAAGC
GYPB:           TTGTGAAAATGGCCATACTGCCCAAGGCAATTTATAGATTCAATGCCATCCCCATCAAGC
                ****************************.**********************************

Dup_B:          TACCAATGACTTTCTTCACAGAATTTGAAAAAAACTACTTTAAAGTTCATATAGAACCAA
GYPE:           TACCAATGACTTTCTTCACAGAATTTGAAAAA-ACTACTTTAAAGTTCATATAGAACCAA
GYPB:           TACCAATGACTTTCTTCATAGAATTCGAAAAAAACAACTTTAAAGTTCATATAGAACCAA
                ******************.*******.****** ** ************************

Dup_B:          AAAGGAGACCACATTGCCAAGCCAATCCTAAGCCAAAAGAACAAAGCTGGAGGCGTCATG
GYPE:           AAAGGAGACCACATTGCCAAGCCAATCCTAAGCCAAAAGAACAAAGCTGGAGGCGTCATG
GYPB:           AAAAGAGCCCGCATTGCCAAGTCAATCCTAAGCCAAAAGAACAAAGCTGGAGGCATCATG
                ***.*** **.**********.*********************************.*****

Dup_B:          CTACCTGACTTCAAACTATACTACAAGGCTACAGTAACCAAAACAGCATGGTACTGGTAC
GYPE:           CTACCTGACTTCAAACTATACTACAAGGCTACAGTAACCAAAACAGCATGGTACTGGTAC
GYPB:           CTACCTGACTTCAAATTATACTACAAGTCTACAGTAACTGAAACAGCATGGTACTGGTAC
                **************.*********** ********** ..********************

Dup_B:          CAAAACAGAGATATAGACCAATAGAACAAAACAGAGCCATCAGAAATAATACCACACATC
GYPE:           CAAAACAGAGATATAGACCAATAGAACAAAACAGAGCCATCAGAAATAATACCACACATC
GYPB:           CAAAACAGAGATATAGACCAATAGAACAGAACAGGGCCCTCAGAAATAATACCACACATC
                ****************************.*****.*** **********************
```

**Multiple sequence alignment for duplication DUP29 breakpoints**

```
DUP29:          GTAGCTATGTTTGTTTTCTGTTTTCTTTCTGTAATTTTCTTGTCTCTCTGAC-TCTCAGA
GYPE:           GTAGCTATGTTTGTTTTCTGTTTTCTTTCTGTAATTTTCTTTTCTCTCTGACTTTTCAGA
GYPB:           GTAGCTATGTTTGTTTTCTGTTTTCTTTCTGTAATTTTCTTGTCTCTCTGAC-TCTCAGA
                **************************************** ********** *.*****

DUP29:          AATCACCAATCACATACGTAGGTTTGTTCCCTAGTCTATAATTTATTGATAGCTACTAATT
GYPE:           AATCACCAATCACATACGTAGGTTTGTTCCCTAGTCTGTAACTACTGATAGCTACTGCTT
GYPB:           AATCACCAATCACATACGTAGGTTTGTTCCCTAGTCTATAATTTATTGATAGCTACTAATT
                ************************************.***.**.***********. **

DUP29:          ATTTTAATATTACTTATATTATTAATACTTATTAATCATGATAGTTAATTCTAATCATAG
GYPE:           ATTTTAATA----ATATATTATTAATACTTATTAATCATGATAGTTAATTCTAATCATAG
GYPB:           ATTTTAATATTACTTATATTATTAATACTTATTAATCATGATAGTTAATTCTAATCATAG
                *********     *******************************************

DUP29:          AAAATGCGTAGTACGCTTAACATTTTACTTTGTTTCTGGCATCACTAAGTGTATTCCATT
GYPE:           AAAATGCATAGTAAGCTTAACATTTTACTTTGTTTCAGGCATCAGTAAGTGTATTCCATT
GYPB:           AAAATGCGTAGTACGCTTAACATTTTACTTTGTTTCTGGCATCGCTAAGTGTATTCCATT
                *******.***** ********************** ******. ***************

DUP29:          AAATATTTTGTTTAATCTTTATGCAATTATTATGAGGTAAGTACCAACATTGCCTGCATT
GYPE:           AAACATTTTGTTTAATCTTTACGCAATTATTATAAGGCAAGTGCCAACATTGCCTGCATT
GYPB:           AAATATTTTGTTTAATCTTTATGCAATTATTATGAGGTAAGTACCAACATTGCCTGCATT
                ***.*****************.***********.***.****.*****************

DUP29:          TTAAAGACGAGGACACATAAAAATGTGGGCCCTTAATTTGCCTTAATGTTCCCCAGCCTT
GYPE:           TTAAAGACAAGGACACATAAAAATGTGGGCACATAATTTGCCTTAATGTTCCCCAGCCTT
GYPB:           TTAAAGACGAGGACACATAAAAATGTGGGCCCTTAATTTGCCTTAATGTTCCCCAGCCTT
                ********.********************* * ***************************

DUP29:          GAGAAGTAAAACTAGATTTGAACCCAGTTCTGTGTGTTCCTCATGCCTGTGCTCTTAACT
GYPE:           GAGAAGTAAAACTAGATTTGAACCCAGTTCTGTGTGTTCCTCATGCCTGTGCTCTTAACT
GYPB:           GAGAAGTAAAACTAGATTTGAACCCAGTTCTGTCTGTTCCTCATGCCTGTGCTCTTAACT
                ******************************** ************************

DUP29:          ATTGCACTATTTTATCCTCCCCTAGGGTATCTTAGAAAATGCTGATCTCTCTTCCCTGTA
GYPE:           ATTGCACTATTTTATCCTCCCCTAGGGTATCTTAGAAAATGCTGATCTCTCTTCCCTGTA
GYPB:           ATTGCACTATTTTATCCTCCCCTAGGGTATCTTAGAAAATGCTGATCTCTCTTCCCTGTA
                ************************************************************

DUP29:          TGCTACAGAGTATGCTCAGTAGACCCTCTCTTCCCTGTATGCTACTGTATGCTCAGTAGA
GYPE:           TGCTACAGAGTATGCTCAGTAGATCCTCTCTTCCCTGTATGCTACTGTATGCTCAGTAGA
GYPB:           TGCTACAGAG----------------------------------------TGCTCAGTAGA
                **********                                        ***********

DUP29:          TCCTTGATACTTGCTCTCACTTGGATTACAGCAATGTCTAATAGACTCAGAATGTTTCAA
GYPE:           TCCTTGATACTTGCTCTCACTTGGATTACAGCAATGTCTAATAGACTCAGAATGTTTCAA
GYPB:           TCCTTGATACTTGCGCTCACTTGGGTTACAGCAATGTCTAATAGACTCAGAATGTTTCAA
                ************** *********.***********************************

DUP29:          GCAACTGCCTGCCCCCTGTGTGCTGGGCACTAGTATTGCTAACACATAAGCACTGCACTT
GYPE:           GCAACTGCCTGCCCCCTGTGTGCTGGGCACTAGTATTGCTAACACATAAGCACTGCACTT
GYPB:           GCAACTGCCTGCCCCCTGTGTGCTGGGCACTAGTATTGCTAACACATAAGCACTGCACTT
                ************************************************************

DUP29:          GAGAGGATGTTTAACTATAATATTCTCAGAAAGAGGCCAAAATCATGTTTATTTAATACA
GYPE:           GAGAGGATGTTTAACTATAATATTCTCAGAAAGAGGCCAAAATCATGTTTATTTAATACA
GYPB:           GAGAGGATGTTTAACTATAATATTCTCAGAAAGAGGCCAAAATCATGTTAATTTAAGACA
                ********************************************** ****** ***

DUP29:          CTAAATTAATAATTAAATTTAAACTAAGTTAAATTCAAATTATTG----AATACAATTTA
GYPE:           CTAAATTAATAATTAAATTTAAACTAAGTTAAATTCAAATTATTG----AATACAATTTA
GYPB:           CTAAATTAATAATTAAATTTAAACTAAGTTAAATTCAAATTATTGAATTAATACAATTTA
                ********************************************    ***********
```

## Multiple sequence alignment for gene conversion (*GYP*: *E*-*B*-*E*) breakpoints

- First breakpoints part

```
E1              TCATGGGAACAGTCTTCTTCAAACATCTTTTAGCACAGGCAAGATTCCCATTTATACGTT
GYPE:           TCATGGGAACAGTCTTCTTCAAACATCTTTTAGCACAGGCAAGATTCCCATTTATACGTT
GYPB:           TCATGGGAACAGTCTTCTTCAAACATCTTTTAGCACAGGCAAGATTCCCATTTATACGTT
                ************************************************************

E1              AATTCTTTCCAAGACAATGAGATTGGGCAGAAAAGGCATTGAGTTGGAAGTCAATGGATA
GYPE:           AATTCTGTCCAAGACAATGAGATTGGGCAGAAAAGGCATTGAGTTGGAAGTCAATGGATA
GYPB:           AATTCTGTCCAAGACAATGAGATTGAGCAGAAAAGGCATTGAGTTGGAAGTCAATGGATA
                ****** ******************.**********************************

E1              TGAGTTTTTGTCCCAGTTTTACCACAAATTAGCTGAGCATAACTTCCACAGATGCATTTA
GYPE:           TGAGTTTTTGTCCCAGTTTTACCACAAATTAGCTGAGCATAACTTCCACAGATGCATTTA
GYPB:           TGAGTTTTTGTCCCAGTTTTACCACAAATTAGCTGAGCATAACTTCCACAGATGCATTTA
                ************************************************************

E1              TCAAGTAGTTTTCATGGTCATTGCAATGCCAAAAAACTGTAGCATTTAGAAAATTTAGTT
GYPE:           TCAAGTAGTTTTCATGGTCATTGCAATGCCAAAAAACTGTAGCATTTAGAAAATTTAGTT
GYPB:           TCAAGTAGTTTTCATGGTCATTGCAATGTCAAAAAACTGTAGCATTTAGAAAATTTAGTT
                ****************************.*******************************

E1              TTCAGACTTGGAAACTATTTAAGGCATTTCATATGAAGGGTGTGTCCTTGTAAGAGTTTG
GYPE:           TTCAGACTTGGAAACTATTTAAGGCATTTCATATGAAGGGTGTGTCCTTGTAAGAGTTTG
GYPB:           TTCAGACTTGGAAACTATTTAAGGCATTTCATATGAAGGGTGTGTCCTTGTAAGAGTTTG
                ************************************************************

E1              CTTATGCAAGATAAGGCTTCTTTCAGCTGCAAGTCAGGAGCGAACCAAAACTCAAACCAG
GYPE:           CTTATGCAAGATAAGGCTTCTTTCAGCTGCAAGTCAGGAGCGAACCAAAACTCAAAGCAG
GYPB:           CTTATGCAAGATAAGGCTTCTTTCAGCTGCAAGTCAGGAGCGAACCAAAACTCAAACCAG
                ************************************************** ***

E1              CAGC--------------TTATCACATCTTGATGAGAGCTCAGCCACTGGAAGTTTTGGC
GYPE:           CAGCTGCATGAGCTGACTTTATCACATCTTGACAAGAGCTCAGCCACTGGAAGTTTTGGC
GYPB:           CAGC--------------TTATCACATCTTGATGAGAGCTCAGCCACTGGAAGTTTTGGC
                ****              **************..**************************

E1              ATACAGCAAAACTGAAGGGTACTTATACAATATCACATTTTATTTTTATTGTTTCTAATA
GYPE:           ATACAGCGAAACTGAAGCGTACTTATACAATATCACACTTTATTTTTATTGTTTCTAATA
GYPB:           ATACAGCAAAACTGAAGGGTACTTATACAATATCACATTTTATTTTTATTGTTTCTAATA
                *******.********* ******************.**********************

E1              GCATTCCAGGTTAGAAATGTCAATTATTTGGGAAAGCTGAGTGGTCTGGTAGATAAAGCA
GYPE:           GCATTCCAGGTTAGAAATGTCAATTATTTGGGAAAGCTGAGTGGTCTGGTAGATAAAGCA
GYPB:           GCATTCCAGGTTAGAAATGTCAATTATTTGGGAAAGCTGAGTGGTCTGGTAGATAAAGCA
                ************************************************************

E1              TATAGCAGAGAGCTAGGAGGCTGGCTATTTCCAGTTGTTATCCTAACATGTCTTGGGCCC
GYPE:           TGCAGCAGAGAGCTAGGAGGCTGGCTATTTCCAGTCGTTATCCTAACATGTCTTGGGCCC
GYPB:           TATAGCAGAGAGCTAGGAGGCTGGCTATTTCCAGTTGTTATCCTAACATGTCTTGGGCCC
                *..*******************************.**********************

E1              CCAAGTCACCCCACCTCCTTGGTACAATGGGAACAGTGGCAGAAGTCCATGCTCTCTCCC
GYPE:           CCAAGTCACCCCACCTCCATGGTACAATGGGAACTGTGGCAGAAGTCCACGCTCTCTCCC
GYPB:           CCAAGTCACCCCACCTCCTTGGTACAATGGGAACAGTGGCAGAAGTCCATGCTCTCTCCC
                ***************** *************** ***************.**********
```

- Second breakpoints part

```
E1              AAACTCATTTTTATGGGAACTATTGTTCCTTCAAGATGACACTGTTTTGTAAACTATAGG
GYPE:           AAACTCATTTTTATGGGAGCTATTGTTCCTTCAAGATGACACTGTTTTGTAAACTATAGA
GYPB:           AAACTCATTTTTATGGGAACTATTGTTCCTTCAAGATGACACTGTTTTGTAAACTATAGA
                ******************.****************************************.

E1              CTTCCAATAACAAGCCTCTGTGCCTTCCTCTTACCACTAAGCAAGCATGGGTATGAATTC
GYPE:           CTTCCAGTAACAAGCCTCTGTGCCTTCTTCTTACCACTAAGCATGCATGGGTATTAATTC
GYPB:           CTTCCAATAGCAAGCCTCTGTGCCTTCCTCTTACCACTAAGCAAGCATGGGTATTAATTC
                ******.**.*****************.*************** ********** *****

E1              CTACTGAAAGGCTTATGCTATCTTTTTTCCAGAAATGGAAGGAAAATAAACTATGAAAAA
GYPE:           CTACTGAAAGACTTATGCTATCTTTTTTCCAGAAATGGAAGAAAAATGAACTATGAAAAA
GYPB:           CTACTGAAAGGCTTATGCTATCTTTTTTCCAGAAATGGAAGAAAAATGAACTATGAAAAA
                **********.*******************************.*****.************

E1              GGTCATTTTATAGGTCAGCTACCATTATGAGATTGTTGAGGAAATGATAT-AAAAACAAT
GYPE:           GGTCATTTTATAGGTCAGCTACCATTATGAGATTGTTGAGGAAATGATATAAAAAACAAT
GYPB:           GATCATTTTATAGGTCAGCTACCATTATGAGATTGTTGAGGAAATGATATAAAAAACAAT
                *.************************************************ *********

E1              TTTTATCAAATTATCTTTAGGGCATTTATATGTTTATTTTCTTACTGTGTTTGACTTAGGT
GYPE:           TTTTATCAAATTATCTTTAGGGAATTTATATGTTTATTTTCTTACTATGTTTGACTTAGGT
GYPB:           TTTTATCAAATTATCTTTAGGGCATTTATATGTTTATTTTCTTACTGTGTTTGATTTAGGT
                ********************* ***********************.******.*******

E1              GACTATAAGAAGTTGTATCAGAGCAACTGATTCTGGAGAATTAAAGCAAGTATTTCTAAG
GYPE:           GACTATAAGAAGTTGTATCAGAGCAACTGATTCTGGTGAATTAAAGCAAGTATTTCTAAG
GYPB:           GACTATAAGAAGTTGTATCAGAGCAACTGATTCTGGTGAATTAAAGCAAGTATTTCTAAG
                ********************************** ************************

E1              AACATAAGTGGCAACTTTCAATCTCAAATCAATTTGGCCACCAATCAGTTTTTGTAAGGG
GYPE:           AACATAAGTGGCAACTTTCAGTCTCAAATCAATTTGGCCACCAATCAGTTTTTGTAAGGG
GYPB:           AACATAAGTGGCAACTTTCAGTCTCAAATCAATTTGGCCACCAACCAGTTTTTGTAAGGG
                ********************.************************.***************

E1              TACAAATAGGACATAACATGCTCAGATGCGACTTGGATGAAGTGTATACAATTTTACATC
GYPE:           TACAAATAGGACATAACATGCTCAGATGGGACTTGGATAAAGTGTATACAATTTTACATC
GYPB:           TACAAATAGGACATAACATGCCCAGATGGGACTTGGATAAAGTGTATACAATTTTACATT
                **********************.****** *********.*********************.

E1              GAGGAAATTGTGTCAATGTGTTACCTTCAATGTTAGAAATTCCCAAGTTCTGACAATAGT
GYPE:           GAGGAAATTGTGTCAATGTGTTACCTTCAATGTTAGAAATTCCCAAGTTCTGACAATAGT
GYPB:           GAGGAAATTGTGTCAATGTGTTACCTTCAATGTTAGAAATTCCCAAGTTCTGACAATAGT
                ************************************************************

E1              TCAGAGCCTTGTTAAAAGCCAGAGTGGAGGCATGTAGATCCAGCTGGAAAGAGAGGCATT
GYPE:           TCAGAGCCTTGTTAAAAGCCAGAGTGGAGGCATGTAGATCCAGCTGGAAAGAGAGGCATT
GYPB:           TCAGAGCCTTGTTAAAAGCCAGAGTGGAGGCATGTAGATCCAGCTGGAAAGAGAGGCATT
                ************************************************************

E1              ATGGTCTAAGTTAGGACAAATTTTAAAGCCAGTGTTAGGGTCTGAGTCCAGCTTTGTATA
GYPE:           ATGGTCTAAGTTAGGACAAATTTTAAAGCCAGTGTTAGGGTCTGAGTCCAGCTTTGTATA
GYPB:           ATGGTCTAACTTAGGACACATTTTAAAGCCAGTGTTAGGGTCTGAGTCCAGCTTTGTAAA
                ********* ******** ****************************************** *

E1              CTTGAGTACAGTGTTTGATCTCTGGGGTTTCAGCCTTCACTTCAGAACAAAATTTCCACC
GYPE:           CTTGAGTACAGTGTTTGATCTCTGGGGTTTCAGCCTTCACTTCAGAACAAAATTTCCACC
GYPB:           CTTGAGTACAGTGTTTGATCTCTGGGGTTTCAGCCTTCATTTCGGTACAAAATTTCCACC
                ***************************************.***.* **************

E1              AAGTGCTCATTTACTGTGAGGAGTAGCTGTTGAAGAAGAAAGAAACTGTTGAAGAAGAAA
GYPE:           AAGTGCTCTTTTACTGTGAGCAGTAGCTGTTGAAGAAGAAAGAAACTGTTGAAGAAGAAA
GYPB:           AAGTGCTCTTTTACTGTGAGGAGTAG------------------CTGTTGAAGAACAAA
                ******** *********** ******                  *********** ***
```

## Appendix 5: Sequence and primers for ARMS

**rs186873296 (R= A/G)**

TCCTGACGTCAAGTGATCCGCCCACCTCRGCCTCCCAAAGTGYTGGGATTACAGRTGTGAGCCACCAT
GCCYRGCCTCTACTAAATTTTTAACATTKAAAAATATTTATTTTACAGTCCTTGYCTGTTAATTACAA
CACCTGGGTCRTCTATGGGTTTTCTGCTGTTGGCATTTTTACAATTATTATGAGWCACATTGAGAATA
ATGCATTGTRGAAGTGTTCTCTTMAGTTACAATTTTCTAAATAGTGTYGAATTCGTTCTTGAAGGCAA
TTAAATTACCAAAAYACAAAYGKTACCTTTGATAMAATTATAAAAACAGGGCATATGGCACCATGTTT
TGAGTAGAGGAATTTGCAGATTAGCATTCACCCAGATGTGAAGAAGCTGGGAACCCTGTCRTACAAGA
AATGACAAAKAAAGCTTATTCTGGAAGGAAGGATGCTTACTCTGGAGCAGAGAAACCAGTAGGAGGCA
TGACAAGCTGGGCCYCTCCAATAATTTCAATAATCTAYTCTTYATAGAAATGTCTGAGTACTTTTTAA
AGACAAACTTCCCAGTACTGTTTGCTTGTCTTACAATAAAAARTMTGGCTTAAAATAAATGACAATTT
TCAGTCTTTTCAYACTTGGCATTTAAGCTTAAGCTCCAAAATAGTTCATTAAATAATGARTCTRCTTG
GCAGRGTCATCTTGTTCTTATTTATTGACATCTTTGCWGAAGGTYGGCTCAGAGAAGATGCAATAAAA
TGATTTACTATAYGCTRTATTTTGCTGTTCATATGTCTGCATCACATCTAACM

rs186873296_ARMS_Forward 5' AGAAGCTGGGAACCCTGTC[G] 3'
rs186873296_ARMS_Reverse 5' ATTGTAAGACAAGCAAACAGTACTGG 3'

>chr4:144702455+144702654 200bp

AGAAGCTGGGAACCCTGTCGTACAAGAAATGACAAAGAAAGCTTATTCTGGAAGGAAGGATGCTTACT
CTGGAGCAGAGAAACCAGTAGGAGGCATGACAAGCTGGGCCTCTCCAATAATTTCAATAATCTACTCT
TCATAGAAATGTCTGAGTACTTTTTAAAGACAAACTTCCCAGTACTGTTTGCTTGTCTTACAAT

Seq_SNP_WA_Forward 5' GAATTCGTTCTTGAAGGCAATT 3'
Seq_SNP_WA_Reverse 5' GGAGCTTAAGCTTAAATGCCAAGT 3'

>chr4:144702326+144702722 397bp

GAATTCGTTCTTGAAGGCAATTAAATTACCAAAACACAAATGTTACCTTTGATAAAATTATAAAAACA
GGGCATATGGCACCATGTTTTGAGTAGAGGAATTTGCAGATTAGCATTCACCCAGATGTGAAGAAGCT
GGGAACCCTGTCGTACAAGAAATGACAAAGAAAGCTTATTCTGGAAGGAAGGATGCTTACTCTGGAGC
AGAGAAACCAGTAGGAGGCATGACAAGCTGGGCCTCTCCAATAATTTCAATAATCTACTCTTCATAGA
AATGTCTGAGTACTTTTTAAAGACAAACTTCCCAGTACTGTTTGCTTGTCTTACAATAAAAAATCTGG
CTTAAAATAAATGACAATTTTCAGTCTTTTCACACTTGGCATTTAAGCTTAAGCTCC

## Appendix 6: R script for scatterplot/histograms

```
rm(list=ls(all.names=TRUE))
compare<-read.table("NGS_vs_cis_prt1.txt",header=T)
png(filename = "NGS_vs_cis_prt1.png",res=100, width = 800,
height = 600 )
layout(matrix(c(2,0,1,3),2,2,byrow=TRUE),
widths=c(3,1),heights=c(1,3),TRUE)
par(mar=c(5.1,4.1,0.1,0))
plot(compare$NGS_cn,compare$cis_prt1_cn,pch=compare$cn,cex=0.5
,col=compare$cn,xlab="NGS_cn",ylab="cis_prt1_cn")
legend("topright",legend=c("1","2","3"),ncol=1,cex=1,bty="n",p
ch=c(1,2,3), col=1:3)
par(mar=c(0,4.1,3,0))
hist(compare$NGS_cn,breaks=50,ann=FALSE,axes=FALSE,col="gray60
",border="black")
Yhist <- hist(compare$cis_prt1_cn,breaks=50, plot = FALSE)
par(mar=c(5.1,0,0.1,1))
barplot(Yhist$density,horiz=TRUE,space=0,axes=FALSE,col="gray6
0",border="black")
dev.off()
```

## Appendix 7: Samtools script to generate the normalized 5 Kb window reads file (Thanks to Dr Daniel Zadik)

```
# samtools program on Spectre after download the BAM file of any sample
Module load samtools
samtools
#use perl reads_per_region.pl file and glycophorin.bed to create the 5 Kb
window reads file
perl reads_per_region.pl glycophorin.bed
NA11994.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.bam NA1199.txt
```

**The glycophorin.bed file contains the 5 Kb windows for a certain region, while the reads_per_region file is a Pl file contains commands in order read each 5 kb window in the glycophorin.bed file automatically**

```
#!/usr/bin/perl
use strict;
use warnings;
# usage is
# perl reads_per_region.pl XXXXXX.bed XXXXXXXXX.bam XXXXXOUT.txt
# where bed file indicates the regions, bam file is the input alignment and
txt file is the output
my ($bed_file, $bam_file, $out_file) = @ARGV;
open (BED, "<$bed_file") or die "cant open $bed_file - $!";
open (OUT, ">$out_file") or die "cant write to $out_file - $!";

while( <BED> ){
      my $row = $_;
      chomp ($row);
      my ($csome,$start,$stop) = split (/\s+/,$row);
      my $result = `samtools view -c -F 4 $bam_file "$csome:$start-$stop"`;
      print OUT "$row $result";}
close BED;
close OUT;
tables
```

## Appendix 8: R script for generating the 5 Kb windows plot (using a set of normal samples as an example)

```
data<-read.table("glycophorin_normal_hg19.txt",header=T)
par(xpd=NA)
plot(data$end,data$NA19257,pch=3,cex=0.5,col="#a6cee3",ylim=c(0.4,1.8),xlim
=c(144750000,145100000),ylab="normalised sequence read
depth",xlab="chromosome 4 (hg19)",bty="n")
points(data$end,data$NA19256,pch=3,cex=0.5,col="#1f78b4")
points(data$end,data$NA19247,pch=3,cex=0.5,col="#b2df8a")
points(data$end,data$NA19240,pch=3,cex=0.5,col="#33a02c")
lines(lowess(data$end, data$mean,f=0.1))
rect(144792019,1.8,144826716,1.85,col='dark blue')
rect(144917257,1.8,144940496,1.85,col='dark blue')
rect(145030456,1.8,145061904,1.85,col='dark blue')
text(144810019,1.75,"GYPE",col='dark blue')
text(144929257,1.75,"GYPB",col='dark blue')
text(145045456,1.75,"GYPA",col='dark blue')
text(145100000,1.75,"Key",cex=1,col="black")
text(145100000,1.65,"NA19257",cex=0.6,col="#a6cee3")
text(145100000,1.6,"NA19256",cex=0.6,col="#1f78b4")
text(145100000,1.55,"NA19247",cex=0.6,col="#b2df8a")
text(145100000,1.5,"NA19240",cex=0.6,col="#33a02c")
text(145046716,1.9,"c1A",cex=0.6,col="dark red")
text(144924531,1.9,"c1B",cex=0.6,col="dark red")
text(144806225,1.9,"c1E",cex=0.6,col="dark red")
text(144997338,1.9,"c2A",cex=0.6,col="dark red")
text(144894139,1.9,"c2B",cex=0.6,col="dark red")
text(144773141,1.9,"c2E",cex=0.6,col="dark red")
text(145066345,1.9,"c4A",cex=0.6,col="dark red")
text(144945004,1.9,"c4B",cex=0.6,col="dark red")
text(144834767,1.9,"c4E",cex=0.6,col="dark red")
text(144960732,1.9,"t2B",cex=0.6,col="dark red")
text(145015246,1.9,"t3A",cex=0.6,col="dark red")
text(144911981,1.9,"t3B",cex=0.6,col="dark red")
text(144791186,1.9,"t3E",cex=0.6,col="dark red")
```

## Appendix 9: Cluster analysis script for analysis of Benin deletions

```
# cis PRT1 v cis_PRT2
# load data from the file
data<-read.table("Benin_samples_Heterozygote_deletions.txt",header=T)
# number of clusters
nclust=2
# subset data
data_to_cluster<-data[c(2,3)]
# do the cluster analysis
clusters<- kmeans(data_to_cluster,nclust)
# append cluster assignment to the data
data<-data.frame(data,clusters$cluster)
# plot the clusters
plot(data$cis_prt1_cn,data$cis_prt2_cn,col=data$clusters,pch=1) data<-
read.table("Benin_samples_Homozygote_deletions.txt",header=T)
# number of clusters nclust=3
# subset data data_to_cluster<-data[c(2,3)]
# do the cluster analysis clusters<- kmeans(data_to_cluster,nclust)
# append cluster assignment to the data data<-
data.frame(data,clusters$cluster)
# plot the clusters
plot(data$cis_prt1_cn,data$cis_prt2_cn,col=data$clusters,pch=1)
```

## Appendix 10: CNVtools script for analysis of Benin *GYP* CNs and histograms

```
Library (CNVtools)
# load the data from the file into a dataframe called CNVdata
CNVdata<-read.table("cis_prt1_cn.txt",header=T)
# create a matrix with just the raw data
raw.signal <- as.matrix(CNVdata[, -c(1, 2)])
dimnames(raw.signal)[[1]] <- CNVdata$Sample
# Perform principal component analysis
# pca.signal <- apply.pca(raw.signal)
# Export first principal component to a text file which can be
# processed in R or Excel
# write.table(pca.signal, "CNVdata_PCA.txt", sep="\t")
# Draw histogram
hist(raw.signal, breaks=50,main='First PC signal',col="light blue")
batches<-factor(CNVdata$Cohort)
sample<-factor(CNVdata$Sample)
ncomp=4
fit.pca<-CNVtest.binary(signal=raw.signal,sample=sample,
batch=batches,ncomp=ncomp,n.H0=10, model.mean="~cn",model.var="~1")
print(fit.pca$status.H0) # this should be "C"
# plots fitted data
cnv.plot(fit.pca$posterior.H0, batch = 'Benin', main = 'Benin cis PRT1',
breaks = 50, col = 'red')
# write data to file
write.table(fit.pca$posterior.H0,file="fitted_cisPRT1.txt",sep="\t")
```

## Appendix 11: QTDT methodology

### Software installation

The QTDT software can be downloaded for windows from the Rockefeller website (http://csg.sph.umich.edu/abecasis/QTDT/) as a zipped folder. This needs to be extracted into an accessible format into a folder that is known to the windows command prompt, usually the C drive. The Software Simwalk2 also needs to be downloaded from this website and unzipped into the same folder as all the QTDT files.

### File preparation

The software requires 2 matched files as input files, a pedigree file and a data file. There are examples in the downloaded QTDT files. Initial files can be generated in an excel format and then once finalised these files can be saved in a text format, in the same folder as the QTDT executable files.

The data file is just a small file that basically tells the software what information is in which column of the pedigree file. It is important that it is matched to the pedigree file, but the order of the columns is not important. The required columns are markers (M), phenotypes (T) and any covariates (C), such as age or sex.

The pedigree file is a large file containing all the data, with each individual's data on a single line. The first 5 columns of the file contains the pedigree information and includes; family identifier; individual identified; paternal and maternal identifies; and sex (1=males, 2=females). After this is the information that is described in the data file.

Marker genotypes are presented as two consecutive integers, and missing values are encoded by a 0 or x. When preparing the initial files in an excel format then use 2 columns for each genotype, with one integer in each column. (NB, for your data as 0 0 is a valid genotype we will need to change the integers prior to analysis).

Phenotypes and covariates are encoded as numbers in a single column with missing values as an x.

**Analysis**

Initial analysis can be performed using the pedstats command. This gives a summary of the contents of the files and also checks that's the files have been formatted correctly. This command also gives information if there are any segregation errors.

```
>pedstats  -p PEDFILE.txt -d DATAFILE.txt
```

Once the files are formatted correctly then you need to generate an IBD file using the Simwalk2 program and the commands prelude and then the command finale to generate a single file

```
>prelude -p PEDFILE.txt -d DATAFILE.txt

>simwalk2

>finale IBD-01.*
```

This will generate an ibd file called qtdt.ibd which can be used in further analysis. The file qtdt.ibd is a default name, and this will the overwritten by subsequent analysis, so it is a good idea to change the name.

Now there are the required files we are able to perform the analysis.

First we can estimate the heritability of the phenotypes using the flags -a- -c- -we -veg and using the qtdt command. (details of the flags and what they all stand for can be found on the QTDT software page, in the on-line tutorial)

```
>qtdt -p PEDFILE.txt -d DATAFILE.txt  -i QTDT.ibd -a- -c- -we
-veg
```

To model within family association use the flag –ao, and to include environmental, polygenic and additive components of variance use the flag –wega

```
   >qtdt -p PEDFILE.txt -d DATAFILE.txt  -i QTDT.ibd -ao- -wega
```

## Appendix 12: Comparison between both Leffler *et al.* (2018) and our Sanger sequencing results of DEL1

### Leffler et al. (2018) Sanger sequence result across the DEL1 breakpoint.


### Our project Sanger sequence result across the DEL1 breakpoint.

## Appendix 13: Benin samples that are positive for the novel SNP (rs186873296)

| DNA sample | rs186873296 alleles |
|---|---|
| A023 | A/G |
| A073 | A/G |
| A195 | A/G |
| C385 | A/G |
| C389 | A/G |
| C489 | A/G |
| G685 | A/G |