COPY NUMBER VARIATION AND RELEVANCE TO DISEASE OF THE COMPLEMENT C3b/C4b RECEPTOR 1 (*CR1*) GENE

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Ezgi Kucukkilic MSc

Department of Genetics

University of Leicester

2017

ABSTRACT

Copy Number Variation and Relevance to Disease of the Complement C3b/C4b Receptor 1 (*CR1*) Gene

Ezgi Kucukkilic

The complement 3b/4b receptor 1 (*CR1*) gene is located at chromosome 1q32.2 in a cluster of complement-related genes. CR1 regulates both classical and alternative pathways of the complement system. CR1 is a major receptor for *Plasmodium falciparum*, and variation within the gene has been associated with different malarial clinical phenotypes. *CR1* shows intragenic copy number variation (CNV) resulting in variation in protein length and number of C3b/C4b binding domains. Previously, CR1 was related to Alzheimer's disease (AD) via complement system regulation. Furthermore, *CR1* variation is responsible for the alleles of the Knops blood group, including McCoy and Swain-Langley.

In this thesis, Novel paralogue ratio test (PRT) assays were developed to robustly type CNV of the low-copy repeat (LCR) regions (which defines the common *CR1*-A and *CR1*-B alleles, but also rarer alleles) within the gene in large cohorts, and an allele-specific hybridisation assay to genotype alleles of the Knops blood group system.

Variation was analysed across global populations, and in the Tori-Bossito cohort (563 infants) from Benin, followed since birth to observe malaria acquisition and treatment. This showed that the Swain-Langley Sl2 polymorphism is not in strong linkage disequilibrium (LD) with the CNV, nor with other Knops blood group alleles. It appears to provide protection against early acquisition of malaria and subsequent number of malarial infections in the Tori-Bossito cohort but these results were not confirmed in an independent cohort (n=276).

The association between the *CR1*-B allele and AD (early-onset (EOAD) and late-onset (LOAD)) was explored, showing that the *CR1* risk loci (rs3818361, rs6656401 (only for EOAD) and rs6701713) were in moderate LD with *CR1*-B, but revealing no association between *CR1*-B (p=0.755) and EOAD (n=633). However, the *CR1*-B allele (risk) appears to be associated with LOAD (n=2185) with (p=0.015) and without (p=0.048) use of a junction fragment PCR assay.

ACKNOWLEDGMENTS

First and foremost I would like to express my sincerest and deepest gratitude and thanks to my supervisor Dr. Edward Hollox who has supported me throughout my thesis and my PhD study with his patience, motivation and immense knowledge. His guidance helped me during my research and writing of this thesis. One simply could not wish for a better supervisor.

I wish to thank all collaborators: Prof. Kevin Morgan; the University of Nottingham, David Courtin and André Garcia; PRES Sorbonne Paris Cité, Université Paris Descartes, Paris, France for providing DNA samples and helping in data analysis.

I would like to thank Rita Rasteiro for help with malaria prevalence estimation, Ihthisham Ali for help with ASO analysis, Etienne Patin for allele frequency data, and Paul Kelly for samples.

Thanks to all the members in lab G3; Shamik Polley, Razan Abujaber, Hasret Ozturk, Walid Algady, Faisal Almalki, Ade Adewoye and Lee Machado for the support, useful advice and their friendship. I would also like to thank all past and present members of Hollox group and Jobling group at University of Leicester for all the help, encouragement and friendship.

I would also like to thank my AP thesis committee Prof. Raymond Dalgleish and Dr. Christopher Talbot for going carefully through my first, second and third year reports, and giving useful suggestions regarding my project. I would also like to thank Prof. Flaviano Giorgini for his advice and help regarding to my research and this thesis.

Finally, from the bottom of my heart special thanks to my husband Ben Stephens and my parents; Ali Kucukkilic and Hale Kucukkilic who made all this possible with their great continuous emotional and financial support and believing in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACGH	array-Comparative Genomic Hybridisation
AD	Alzheimer's disease
ANR	l'Agence Nationale de la Recherche
APOE	Apolipoprotein E
bp	Base pairs
BSA	Bovine Serum Albumin
CEPH	Centre de'Etude du Polymorphisme Humain
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry
CHB	Han Chinese in Beijing, China
CMT1A	Charcot-Marie-Tooth disease type 1A
CN	Copy Number
CNP	Copy Number Polymorphism
CNV	Copy Number Variation
CoNVEM	Copy number variation frequency estimation by expectation maximisation
CR1	Complement C3b/C4b receptor 1
CR1L	Complement component C3b/C4b receptor like
CR2	Complement C3d receptor 2
ddNTP	Dideoxy Nucleotide Triphosphate

DGV	Database of Genomic Variants
DNA	Deoxyribonucleic acid
dNTPs	Deoxy Nucleotide Triphosphate
EOAD	Early-Onset Alzheimer's Disease
ESR	Erythrocyte Sedimentation Rate
FAM	6-Carboxyfluorescein
FoSTeS	Fork-stalling and template switching
gDNA	Genomic DNA
GWAS	Genome-Wide Association Study
HEX	Hexachloro-fluorescein
HGDP-CEPH	The Human Genome Diversity Project-Centre d'Etude du Polymorphisme Humain
HRC	Human Random Control
IgG	Immunoglobulin G
JPT	Japanese in Tokyo, Japan
kb	Kilobase
LCR	Low-copy repeat
LD	Linkage Disequilibrium
LHR	Long homologous region
LNA	Locked nucleic acid
LOAD	Late-Onset Alzheimer's Disease
MAQ	Multiplex Amplicon Quantification

NAHR	Non-allelic homologous recombination
NaOH	Sodium Hydroxide
ng	nanogram
NHEJ	Non-homologous end joining
OR	Odds Ratio
PBS	Phosphate buffered saline
PCR	Polymerase Chain Reaction
PRT	Paralogue Ratio Test
RFLP	Restriction fragment length polymorphism
PfRh4	*P. falciparum* reticulocyte-binding homolog
SCR	Short consensus repeat
SD	Segmental duplication
SLE	Systemic Lupus Erythematosus
SNP	Single Nucleotide Polymorphism
Tolimmunpal	Projet Tolérance Immunitaire Associée au Paludisme
UCSC	University of California Santa Cruz
WHO	World Health Organization
YRI	Yoruba in Ibadan, Nigeria
µg	Microgram

1 INTRODUCTION

1.1 Copy Number Variation

The human genome shows genetic variation between different individuals (Girirajan et al., 2011). A variant is defined as any form of DNA variation irrespective of frequency and phenotypic effect (Jobling et al., 2014 p.690). These include single nucleotide polymorphisms (SNPs), deletions, duplications and even changes in copy number of whole chromosome (Girirajan et al., 2011). Polymorphism is usually defined as a genetic variation which has a minor allelic frequency of greater than 1% in a given population. Single nucleotide polymorphisms are variants due to a base substitution or the insertion or deletion of a single base. Many variants described as SNPs can have minor allele frequencies lower than 1% (despite the definition of polymorphism) (Jobling et al., 2014 p.637). SNPs are widely distributed and frequent (Jobling et al., 2014 p.49). Two fundamental processes cause base substitution (SNPs): misincorporation of nucleotides during replication and mutagenesis caused by the chemical modification of bases or physical damage due for example UV or ionizing radiation. The human genomes sequenced to date revealed that each genome contains ~3-4 million SNPs (one SNP about every 1006 bp and 1250 bp for Africans and Europeans, respectively) (Jobling et al., 2014 p.62). According to the 1000 Genomes Project, a human genome contains on average ~10,000 nonsynonymous (missense) and ~30 nonsense SNPs (Jobling et al., 2014 p.58).

The human genome contains long duplicated sequences which comprises almost 5% of the genome and have arisen within the last 40 million years called segmental duplications. Segmental duplications (also known as low-copy repeats) are continuous segments of DNA that show high levels of sequence identity (>90%) and their size ranges from 1 to 400kb (this limit is somewhat arbitrary) (Eichler, 2001). Stankiewicz and Lupski (2002) suggested that combination of segmental duplications and non-allelic homologous recombination (NAHR) can results in chromosome rearrangements including deletion, duplication, inversions, translocations and complex chromosome rearrangements because presence of highly homologous flanking repeats prone these regions to repetitive rearrangement by NAHR (Figure 1.1). Therefore, these regions are

probable hotspots of genomic instability and because of this they are prone to copy number variation (Sharp et al., 2005).

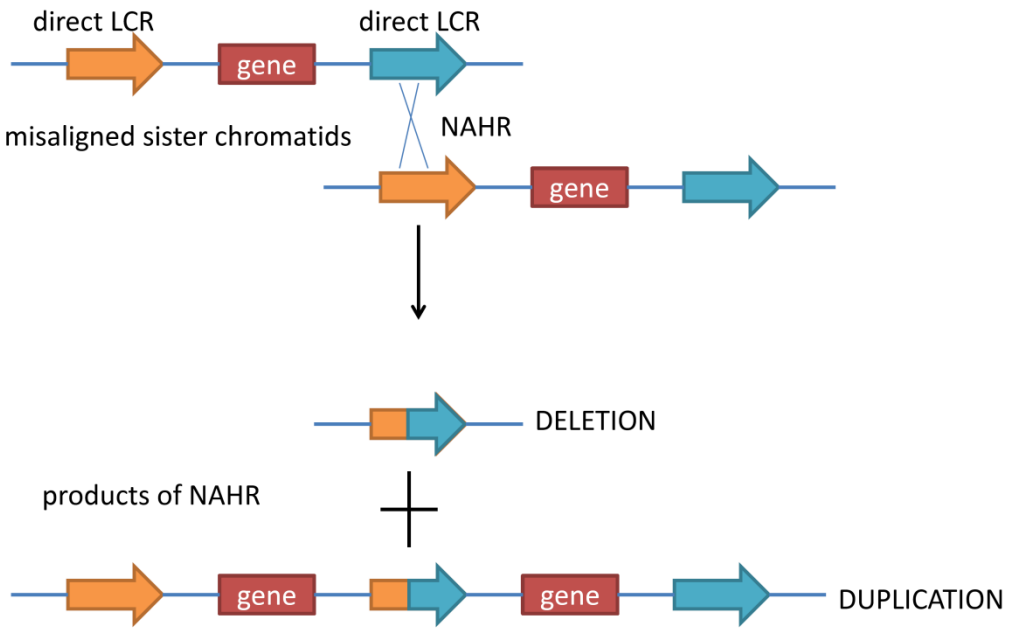


Figure 1.1: Products of NAHR between direct repeats. NAHR between direct repeats generates deletions or duplications. Adapted and reproduced from Jobling et al. (2014) p.76.

Bailey and Eichler (2006) showed that segmental duplications created novel primate genes and also triggered current human genic and phenotypic variation. In addition, natural structural variation of these regions is linked to susceptibility to common diseases. For example, CNV proximal of the *PMP22* gene causes Charcot-Marie-Tooth disease (Marian et al., 2010). Also, it has been found that copy number variation of *CCL3L1* is associated with susceptibility to HIV-1 infection. A lower copy number of *CCL3L1* is associated with an increased risk of HIV-1 infection whereas a higher copy number of *CCL3L1* is associated with reduced risk (Liu et al., 2010).

Structural variants that occur in chromosomes can be balanced, involving no alteration of copy number, or can be copy number variable involving differences in the number of particular sequences between alleles. The impact of structural variation on the genome can be assayed effectively by technologies such as microarrays, comparative genomic hybridization (CGH) and genome sequencing.

Structural variation is generally defined as a region of DNA at least 1 kb in size or larger referred to CNVs excludes small indels, including *Alu* insertions. Copy number

polymorphism (CNP) describes a CNV that exists at >1% frequency in a population (Jobling et al., 2014 p. 77).The mechanisms of mutation of CNVs include NAHR but other mechanisms are also involved. Some simple duplication/deletion CNVs is generated by nonhomologous end-joining (NHEJ) which is a mechanism for repairing double-strand breaks (Gu et al., 2008).

Most CNV is not likely to have phenotypic effect (Jobling et al., 2014 p. 77) because only 2% of human genome is coding (Jobling et al., 2014 p. 26); therefore most random CNVs will not contain genes. However, some variation causes genetic disease and others contribute to the susceptibility to complex disease (Table 1.1). For example, higher copy number of β-defensin genes increases the risk of common inflammatory skin disease psoriasis (Hollox et al., 2008).

Table 1.1: Some of the complex diseases related to CNV at specific loci. Reproduced from Zhang et al. (2009).

DISEASE	LOCUS	RELATION	REFERENCES
Alzheimer's disease	*APP*	Duplications are responsible for rare Mendelian forms of disease	(Helbig et al., 2009)
Parkinson's disease	*SNCA*	Duplication/triplication is responsible for rare Mendelian forms of disease	(Farrer et al., 2004; Zhang et al., 2009)
Systemic Lupus Erythematosus (SLE)	*FCGR3B*	Common variation (copy number loss) is weakly associated with SLE	(Aitman et al.,2006; Zhang et al., 2009)
SLE	*C4*	Common variation (copy number loss) is weakly associated with SLE	(Lalic et al., 2005)
Psoriasis	*DEFB4*	High copy number increases risk of disease	(Hollox et al., 2008; Stuart et al., 2012)
HIV	*DEFB4*	High copy number increases HIV load prior to highly active antiretroviral therapy (HAART) and poor immune reconstitution following initiation of HAART	(Hardwick et al., 2012)

Some CNV may have a role in natural selection. An interesting example of this is the salivary amylase gene, where copy number varies between populations that have different dietary starch intakes (Perry et al., 2007). Many phenotypes associated with CNVs are regulated by gene-dosage effects, disruption of a coding sequence or alteration of transcriptional regulation (Jobling et al., 2014). For example, partial

duplication of *SRGAP2* contributed to human brain evolution by producing an incomplete protein that antagonizes the original (Tyler-Smith and Xue, 2012).

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. Rate of CNV mutations when LCR involved lie in the range of 10^{-6} to 10^{-4} per generation (1 in 800 bp on average in, 270 individuals, HapMap samples). Therefore, CNV mutation rates are ~3 orders of magnitude than those SNPs that this disrupts the LD relationship (Jobling et al., 2014 p.78).

Indeed, Campbell et al. (2011) has shown that a significant proportion of copy number polymorphisms (CNPs) involving segmental duplications (SDs) are not in linkage disequilibrium with nearby single nucleotide polymorphisms. This research has shown that 40% of 192 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).

1.2 Methods to Detect and Measure CNV

The susceptibility to common diseases can be influenced by genomic copy number. Therefore, high-throughput measurement of gene copy number is critical but it can be challenging especially on large sample sizes. The confirmation of obtained copy number measurement with aCGH and sequencing is critical. There are few methods to detect and measure CNVs and they have advantages as well as disadvantages (Veal et al., 2013).

1.2.1 The Paralogue Ratio Test (PRT)

The Paralogue ratio test (PRT) is a method to obtain copy number by accurately quantifying the amplification ratio between a target and reference amplicon. This method is cost effective and has been used effectively to several studies testing common CNV (Veal et al., 2013). PRT is an inexpensive method meaning that it can be used in large scale association studies with low quantity genomic DNA (10-20 ng) to determine copy number (Cantsilieris et al., 2013). Due to the assay design challenges it has not been widely used (Veal et al., 2013). In order to overcome the PRT assay design challenge PRTPrimer (www.prtprimer.org) software can be used (Veal et al., 2013).

PRT is particularly designed to measure complex regions of the genome such as segmental duplications. This technique has been used to analyse the copy number variations of Beta-defensin, *FCGR3B*, *CC3L1* and complement component 4 (*C4*) genes. This proved that PRT is an accurate methodology to detect copy number variation. The strength of the PRT relies on the first pass assignment of integer copy number which is very high at 85% for *CCL3L1*, 93% for *FCGR3B*, 93% for the beta-defensin locus and 91% for *C4*. Previously, PRT has been shown to be better than quantitative PCR (qPCR) in terms of determining integer copy number and accuracy in copy number measurement (Cantsilieris et al., 2013). qPCR monitors the accumulation of specific PCR product from a candidate CNV region in real time which measured against an independent control sequence being invariant in copy number (Jobling et al., 2014 p. 116). Haridan et al. (2015) showed that qPCR assay is not suitable for the use in large scale case-control studies for multi allelic genes such as *FCGR3* whereas PRT is providing more reliable results. PRT's workflow procedure is very rapid and does not require the lengthy overnight hybridization steps as with multiplex ligation-dependent probe amplification (MLPA) and multiplex amplifiable probe hybridization (MAPH). Several PRT studies showed accuracy with well validated methods such as MLPA and MAPH. There are few limitations of PRT method. This methodology assumes that the paralogous reference loci do not itself vary in copy number, that sites of primer binding sites are free from polymorphisms which can affect primer binding efficiency and the dependence of identifying paralogous sequence outside the target region. A

technical limitation of PRT is regions of high sequence identity such as *FCGR3A*/*FCGR3B* and *C4A*/*C4B*, do not contain the sequence differences required for primer specificity. Lower multiplexing capacity compared to MLPA and MAPH is another limitation of PRT (Cantsilieris et al., 2013).

1.2.2 Southern Blotting and Pulsed Field Gel Electrophoresis (PFGE)

Southern blotting which is a strong method to type structural rearrangements relies on fragmentation of DNA with a restriction endonuclease enzyme and then followed by agarose gel electrophoresis and transfer to a nylon membrane. A labelled DNA probe is used for hybridization step and exposure to film reveals visualization of target regions. This technique includes both hybridization and electrophoresis steps therefore it requires a normal control with unknown samples or altered fragment sizes in order to compare hybridization intensities between them to detect structural rearrangements (Cantsilieris et al., 2013).

Southern blot hybridization can be used to resolve copy number polymorphisms. For example, Aitman et al. (2006) showed that Southern blotting can be used to determine the copy number at the *FCGR3B* locus (commonly 1-4 copies per diploid genome) through differential band intensities after normalization on a reference locus.

There are disadvantages of this technique such as large amount of high quality of genomic DNA and labour-intensive workflow. In addition to those disadvantages, the errors through the analysis can affect the results negatively. For example, uneven transfer of DNA to the nylon membrane or unfinished washing of probes can result in misinterpretation of band intensities. Therefore, careful probe design and choosing a restriction enzyme that creates significant length distances between fragments has a critical role in minimizing the effect of these limitations (Cantsilieris et al., 2013).

In order to avoid scoring copy numbers incorrectly by conventional agarose gel electrophoresis and Southern blotting, Pulsed field gel electrophoresis (PFGE) can be used, also in combination with Southern blotting. PFGE can be used for both restriction mapping and characterization of large genomic rearrangements. PFGE uses periodic alteration of electric field which generates continuous reorientation of DNA molecules to obtain high resolution and sizing of large DNA fragments. The advantage of this

technique is the direct detection of chromosomal rearrangements (deletions, duplications, translocations, insertions and also inversions) which are more challenging to detect by other methods apart from FISH. All types of rearrangements are in principle detectable by fragment size measurement in PFGE, and therefore comparison of hybridization intensities is not required. Southern blotting with PFGE is a useful and powerful method that provides high resolution (resolving DNA fragments >12Mb) and accurate copy number measurement, including the analysis of altered junction fragments (Cantsilieris et al., 2013).

PFGE has been shown to be useful for showing absolute copy number of multi-allelic copy number polymorphism. For example, Hollox et al. (2009) showed that Fcγ receptor copy numbers of 3-5 estimated by PRT agreed with PFGE fragment sizes increasing by 84 kb intervals, matching the length of the duplicon containing the Fcγ region.

However, there are some limitations to this technique such as the labour-intensive workflow, the requirement of high molecular weight DNA (not always available from archived samples) and dependence on suitable restriction enzyme sites to characterize rearrangements (Cantsilieris et al., 2013).

1.2.3 ArrayCGH (aCGH)

aCGH is a method which relays on dual hybridization of test and reference DNA to either immobilized short oligonucleotides or long DNA sequences (e.g. bacterial artificial chromosomes (BACs)). In order to obtain copy number, the signal ratio between test and reference sample is normalized. Previous approaches using BAC clones (resolution 100-200 kb) provided deep understanding of the landscape of structure variation in the human genome. Poor breakpoint resolution can cause the overestimation of copy number variation size therefore more recent studies prefers to use long nucleotide arrays which provides more accurate picture of the copy number variation landscape (resolution between 0.5-2 kb) (Cantsilieris et al., 2013 This method provides clear indication of the presence of a CNV region on particular location; however it cannot give clear information about the precise copy number of a sequence. For example, while it is easy to detect the difference between one and two

copies of a sequence; it is harder to differentiate between four and five copies. Alternatively, qPCR and PRT can be used to give information about the copy number of a test locus; or PFGE can be used to provide information about *per-chromosome* copy number of a variant (Jobling et al., 2014 p. 116).

1.2.4 Fluorescence in-situ hybridization (FISH)

Fluorescence in-situ hybridization (FISH) is a visual technique which is commonly used to identify chromosomal abnormalities from metaphase to interphase spreads using fluorescent probes. Although, the direct visualization of DNA copy number at the single cell level is simple and standard when using FISH, analyzing multi-allelic copy number polymorphisms by FISH can be challenging especially for tandem duplications. Alternatively, Fibre FISH can be used to accurately resolve simple and complex structural rearrangements and also repetitive sequences. It is the most accurate methodology to type multi-allelic copy number polymorphisms (Cantsilieris et al., 2013). For example, the salivary amylase (*AMY1*) gene was studied by Fibre FISH as complex multi-copy gene rearrangements. Individuals with more than 10 copies of the tandem duplication can be easily resolved (Perry et al., 2007). The core of Fibre FISH underlies in hybridization of DNA fibres with fluorochrome-labelled DNA probes. Multiple DNA targets can be analyzed by using multi-coloured probes. The advantage of using Fibre FISH is ability to identify copy number per allele which is critical for inheritance and disease studies. There are some limitations such as labour-intensive workflow, low throughput and a requirement for high quality sample. Besides, highly variable regions can be challenging for Fibre FISH if there are overlapping signals (Cantsilieris et al., 2013).

1.2.5 Sequence read-depth

Read depth is a methodology to measure the absolute copy number of the target location. The number of the sequencing reads that are mapped to a specific region is proportional to the number of copies of that region in the genome. This method relies on a Poisson distribution of sequencing reads, and thereby can indicate if the region is deleted or duplicated depending on the region having fewer or more mapped reads than expected. For smaller CNV events which also contain high genomic copy number require sequence read depth to achieve accurate copy number measurements. Hollox

(2012) suggested that there is a reciprocal relationship between the size of the CNV and the required sequence read depth in order to discriminate four from five copies accurately. Hollox (2012) found that increasing sequence read depth improves the accuracy of CNV measurement (up to 50X coverage) and is needed to gain higher resolution. The limitation of using massively parallel short-read sequencing is the inability to uniquely map short reads to regions such as segmental duplications. The advantage of using sequence read depth is being able to transform relative copy number obtained from aCGH data into accurate copy number. The limitation of this is inaccuracy with high copy number counts (Cantsilieris et al., 2013).

1.2.6 Multiplex Amplicon Quantification (MAQ)

Multiplex Amplicon Quantification (MAQ) is a method for the measurement and analysis of CNVs. It requires PCR amplification of several fluorescently labelled target and reference amplicons. The method relies of firstly fragment analysis and then secondly comparison of the relative intensities of the target amplicons in the test individual and a control individual which result in the copy number of those target amplicons (Multiplicom, 2015). This method was efficiently used in few studies (Brouwers et al., 2012; Kumps et al., 2009). Brouwers et al. (2012) used MAQ for the *CR1* LCR dosage analysis in a late-onset Alzheimer's disease cohort. Copy number status of the 18-kb-long LCRs in *CR1* and *CR1L* was defined using MAQ in a French late-onset Alzheimer's Disease (LOAD) cohort (n=2003) (Brouwers et al., 2012). Kumps et al. (2009) showed that MAQ is a low-cost, time-effective, closed-tube and high-throughput PCR-based technique for detection of copy number variations in regions with prognostic relevance for neuroblastoma, and comparison with aCGH revealed that it can reliably detect genomic aberrations. According to Kumps et al. (2009), compared to MLPA, MAQ uses less DNA, involves reduced handling, and has limited contamination problems. However, although no false negatives were detected for MLPA, MAQ gave one false negative result at chromosome 1p (false negative rate 1.1%). Therefore, the regions where the reference amplicons bind should be chosen carefully because even low-changes in these regions (reference) can change copy number results dramatically.

1.3 CNV Mechanisms

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 1.2).

1.3.1 Nonallelic Homologous 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 (Lupski, 1988; Stankiewicz and Lupski, 2002). NAHR requires substrates with extended homology for genomic rearrangements. The substrates for NAHR are low copy repeats (LCRs) (also known as segmental duplications (SDs)) which have length >10 kb and similarities of 95-97% (Shaw and Lupski, 2004; Stankiewicz and Lupski, 2002). NAHR can occur in both meiosis and mitosis, leading to different consequences. For instance, in meiosis NAHR can cause unequal crossing over as revealed by segregation of marker genotypes and lead to constitutional genomic rearrangements which can be benign polymorphisms or manifest sporadic or or inherited genomic disorders (if de novo) (Lupski, 2007; Lupski and Stankiewicz, 2005; Turner et al., 2008). NAHR can also occur in mitosis causing mosaic populations of somatic cells carrying copy number or structural variations (Flores et al., 2007; Lam and Jeffreys, 2006; Lam and Jeffreys, 2007).

1.3.2 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 (Zhang et al., 2009).

1.3.3 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 (Lee et al., 2007).

1.3.4 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).

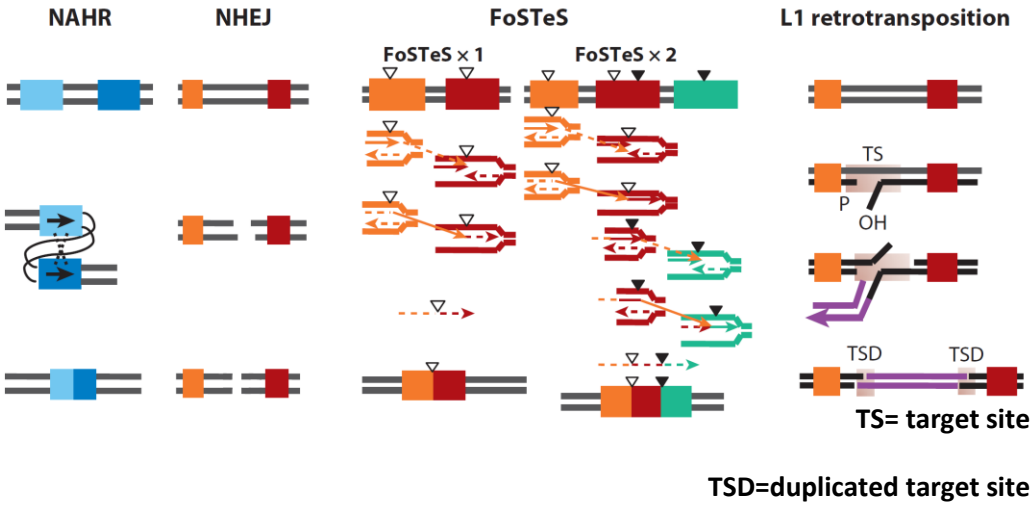


Figure 1.2: Mechanisms causing CNV formation: NAHR between repeat sequences (LCRs/SDs, Alu, or L1 elements); NHEJ recombination repair of double strand break; FoSTeS multiple FoSTeS events (×2 or more) resulting in complex rearrangement and single FoSTeS event (×1) causing simple rearrangement; and retrotransposition. Thick bars of different colours indicate different genomic fragments; completely different colours (as orange and red or orange/red/green in FoSTeS×2) symbolise that no homology between the two fragments is required. The two bars in two similar shades of blue indicate that the two fragments involved in NAHR should have extensive homology with each other. The triangles symbolise short sequences sharing micro-homologies. Each group of triangles (either filled or empty) indicates one group of sequences sharing the same micro-homology with each other. Taken from Zhang et al. (2009).

1.4 The Complement System

The complement system was discovered by Bordet in 1896 and is a network of plasma- and membrane-associated serum proteins that are involved in the defence against

infectious organisms such as bacteria and viruses. The complement system is older than the adaptive immune system (Walport, 2001). The proof of this idea is the identification of a functional component of the complement system (C3) in horseshoe crab *(Carcinoscorpius rotundicauda) and* cnidarian anthrozoans *(Nematostella vectensis)* (Miller et al., 2007; Zhu et al., 2005) whereas the adaptive system is younger and is generally considered to be restricted to jawed vertebrates (Du Pasquier and Litman, 2000). The complement system is a proteolytic cascade of more than 30 proteins, comprising 15% of globular proteins in plasma. This system can be triggered by the identification of surfaces of pathogens (innate immune response) and leads to formation of potent proinflammatory mediators, opsonization (antigen binds to antibody or complement proteins) and pathogen targeting by membrane-penetrating pores (MAC) for the lysis process. Deficiencies in the complement system can cause some diseases such as systematic lupus erythematosus which can result from the deficiency of C1q, C1r, C1s, C2 and C4 complement proteins (Figure 1.3) (Sarma and Ward, 2011). Recent research indicates that the complement system has an important role in the adaptive and innate immune system as it can regulate B- and T-cells and gives a fast response to pathogens (Dunkelberg and Song, 2010). It is also involved in typical haemolytic uraemic syndrome, age related macular degeneration (Wagner and Frank, 2010), tissue regeneration and tumour growth (Qu et al., 2009).

PROTEIN	PATHWAY AFFECTED	IMMUNOLOGIC FUNCTION
C1q	Classical Pathway	Binds to antigen-bound antibody
C1r	Classical Pathway	Activates C1s
C1s	Classical Pathway	Cleaves C4 and C2
C4	Classical Pathway	Component of C3 convertase complex
C2	Classical Pathway	C2 binds to C4 forms C3 convertase
FACTOR D	Alternative Pathway	Activates Factor B
FACTOR B	Alternative Pathway	Binds to C3a
PROPERDIN	Alternative Pathway	Stabilizes C3 convertase
C3	Common Pathway	Key component for generation of C5 convertase activity
C5, C6, C7, C8 and C9	Alternative Pathway	Components of the MAC
Complement Receptor type 1 (CR1 or CD35)	Classical, Alternative and Lectin Pathways	Blocks formation of C3 convertase by binding C4b or C3b; cofactor for Factor I-catalyzed cleavage of C4b or C3b

Figure 1.3: The summary of the protein components and functions of the complement system with the affected pathways. Reproduced from Kindt et al. (2007).

Complement C3b/C4b receptor 1 (CR1 or CD35) is a large type–I transmembrane glycoprotein and is also known as immune adherence receptor or complement (3b/4b) receptor 1 (Brouwers et al., 2012). CR1 is a complement regulatory protein of both classical and alternative pathways of complement system, and is a large glycoprotein with four isoforms and variable membrane expressions which differs from cell to cell (Cooling, 2015). CR1 is expressed in the majority of peripheral blood cells and it can bind with high affinity to C4b and C3b as well as ic3b, C3dg, C1q and mannose-binding protein (Dunkelberger and Song, 2010). As CR1 can bind to immune complexes it is important for the immune system and can be found on erythrocytes, B-cells, some T-cells, monocytes, neutrophils, follicular dendritic cells and glomerular podocytes. The expression level of CR1 can vary according to the cell type. For example, erythrocytes can carry between 200 and 1000 receptors per cell whereas this number is higher for nucleated cells which can have between 10,000 and 20,000 receptors per cell (Funkhouser and Vik, 1999). As CR1 can bind to complement components C3b and C4b, it has a role in the activation of the complement system and inhibition of the

complement cascade. CR1 regulates phagocytosis by neutrophils and monocytes in order to help the clearance of the immune complexes by binding C3b in immune complexes. Failure of this process cause the deposition of complexes in tissues and activation via Fc receptors results in tissue injury (Sarma and Ward, 2011). CR1 can inhibit classical and alternative pathways of the complement system by showing decay-accelerating activity for both C3 and C5 convertases, or by exhibiting collapse cofactor activity for Factor-I mediated cleavage of C3b and C4b (Brouwers et al., 2012). On erythrocytes, CR1 functions as an immune adherence molecule which transports complement-opsonized particles to liver and spleen for clearance from the blood, whereas in other cells CR1 binds to complement components and induces phagocytosis of the opsonized complexes. CR1 loss from erythrocytes causes receptor deposition and this may lead to chronic inflammatory damage. While CR1 loss from these cells can cause their accumulation in kidney and joints, it leads to the activation of cytotoxic immune cells and lysis of cells via membrane attack complex (MAC) (Funkhouser and Vik, 1999).

The complement system can be activated by three pathways which are the classical, lectin and alternative pathways (Dunkelberger and Song et al., 2010).

The classical pathway of complement system is initiated when C1, in complex with C1r and C1s serine protease (the C1 complex), binds to the Fc region of complement-fixing antibodies (IgG or IgM) attached to pathogenic surfaces as shown in Figure 1.4 (Ricklin et al., 2010). Activated C1r and C1s cleave C4 and C2 into C4b, C4b, C2a and C2b. Larger fragments make a complex (C4bC2a) on the pathogen surface and become C3 convertase. This complex has an ability to cut C3 into anaphylatoxin C3a and opsonin C3b. The cleavage of C3 into C3b reveals an internal thioester bond which enables C3b binding to hydroxyl groups on nearby proteins and carbohydrates. Tagging organisms by complement system leads to further complement activation on and around the opsonized surface and finally causes the production of anaphylatoxins and membrane attack complex (MAC). C5 convertase, which is the second molecule of the alternative and classical pathway, is formed by the association of C3b with C4bC2a (C3 convertase) (Walport, 2001).

The lectin pathway is similar to the classical pathway but is activated without the need of immunoglobulin. Instead of antibody-antigen immune complexes, the lectin pathway uses pattern-recognition receptors (PRRs) such as mannose-binding lectin (MBL) and ficolins in order to regulate non-self recognition. PRRs mostly detect highly conserved structures on microorganisms such as pathogen-associated molecular patterns (PAMPs) as shown in Figure 1.4 (e.g., endotoxin or lipopolysaccharide of Gram-negative bacteria) whereas antigen-recognition receptors (antibody or T-cell receptors) of the adaptive immune system can detect and bind a variety of antigens by means of somatic diversity. MBL makes a complex with MBL-associated serine proteases (MASPs) which are structurally similar to C1s and C1r. Binding of MBL to pathogenic surfaces cause the activation of associated MASPs and cleavage of C2 and C4. MAPS1 cleaves both C4 and C2 whereas MAPS 2 cleaves C2 but not C4 (Chen and Wallis, 2004). The cleavage of C2 and C4 leads to the generation of C3 convertase (C4bC2a) in both classical and lectin pathways. Then, C3 cleavage leads to formation of C3 and C5 convertases (Dunkelberger and Song, 2010).

The alternative pathway of the complement system is different from the classical and lectin pathways in many ways (Figure 1.4). Low-level sudden hydrolysis of C3 to C3b analog, C3 (H_2O), which binds to Factor B, allows the cleavage of Factor B into Bb and Ba by Factor D and formation of C3 convertase (C3 (H_2O)). C3 convertase, as in classical and lectin pathways, turns C3 into C3b and C3a. C3b is activated by Factor D to form C3Bb (C3 convertase) and is stabilized by properdin (Factor B) which amplifies activation of the alternative pathway. Properdin promotes alternative pathway activation by stabilizing C3 and C5 convertases (Ricklin et al., 2010).

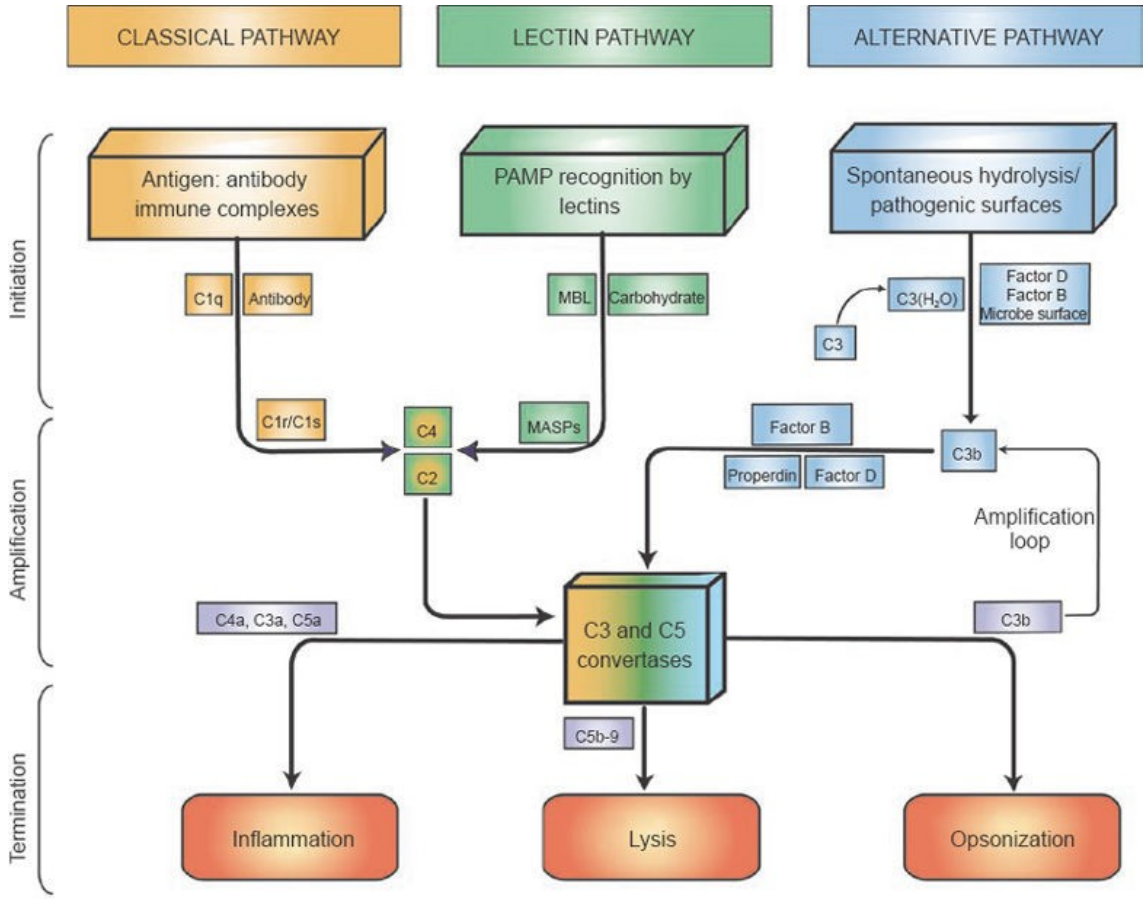


Figure 1.4*:* Three pathways of the complement system: the classical pathway, lectin pathway and alternative pathway. Taken from Dunkelberger and Song (2010).

Finally, the formation of the complement system's convertases leads to the generation of anaphylatoxins (C4a/C3a/C5a), the membrane attack complex (MAC) and opsonins (C3b). Anaphylatoxins, which are effective proinflammatory molecules, arise from the cleavage of C4, C3 and C5. MAC, which is formed by the assembly of C5b and C9 complement components leads to lysis of targeted surfaces. C3b causes phagocytosis of opsonized targets and amplifies complement activation (Dunkelberger and Song, 2010; Ricklin et al., 2010).

1.5 The Structure of the CR1 Protein

The CR1 extra cellular domain is composed of 30 short consensus repeats (SCRs), each having 60 to 70 amino acids. Twenty-eight of the consensus repeats come together to make higher-order repeat units composed of seven SCRs termed long homologous consensus repeats (LHRs) (Figure 1.5).

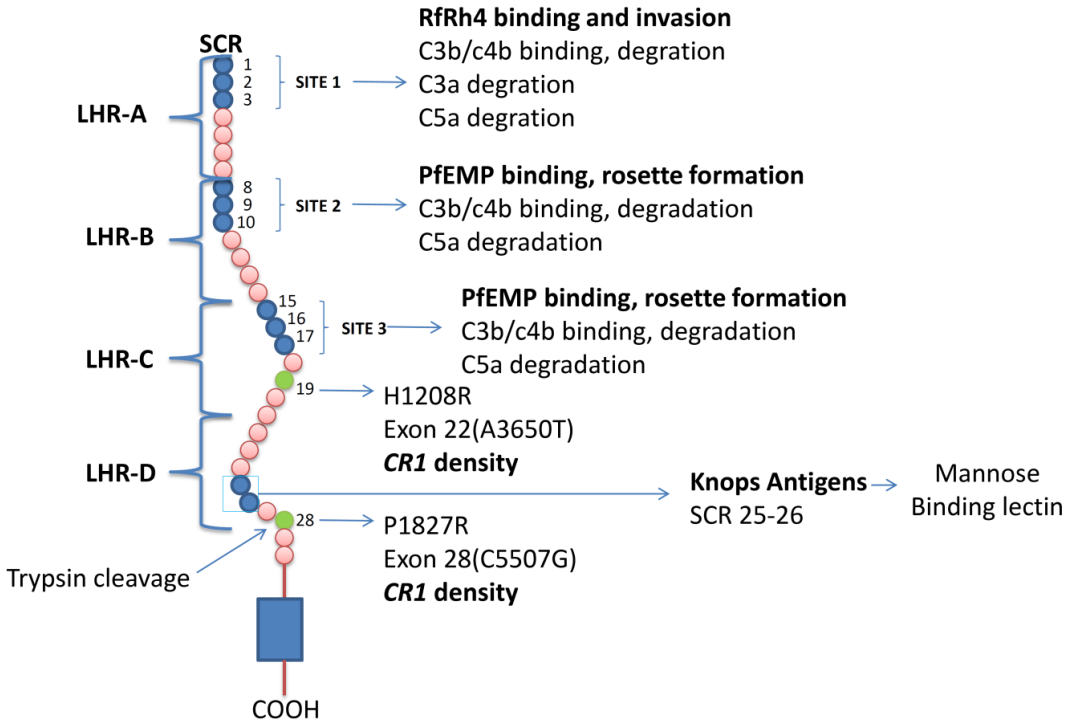


Figure 1.5*:* SCRs of CR1 protein. The CR1 variant on erythrocytes is composed of 30 SCRs which are arranged as 4 LHRs (LHR-A to-D). The blue regions which are marked are the regions related to complement binding and also binding sites for malaria proteins (PfRh4 and PfEMP). The mannose binding site is also highlighted. The Knops antigens are located in SCR25 and SCR26 (LHR-D) (blue). SCR 19 and SCR 28 (Tyrpsin cleavage site on CR1) are sites of two mutations associated with weak CR1/Knops antigen expression (green). Reproduced from Cooling (2015).

The larger alleles (*CR1*-B and *CR1*-D) have an additional C3b/C4b binding site as shown in Figure 1.6 (Brouwers et al., 2012).

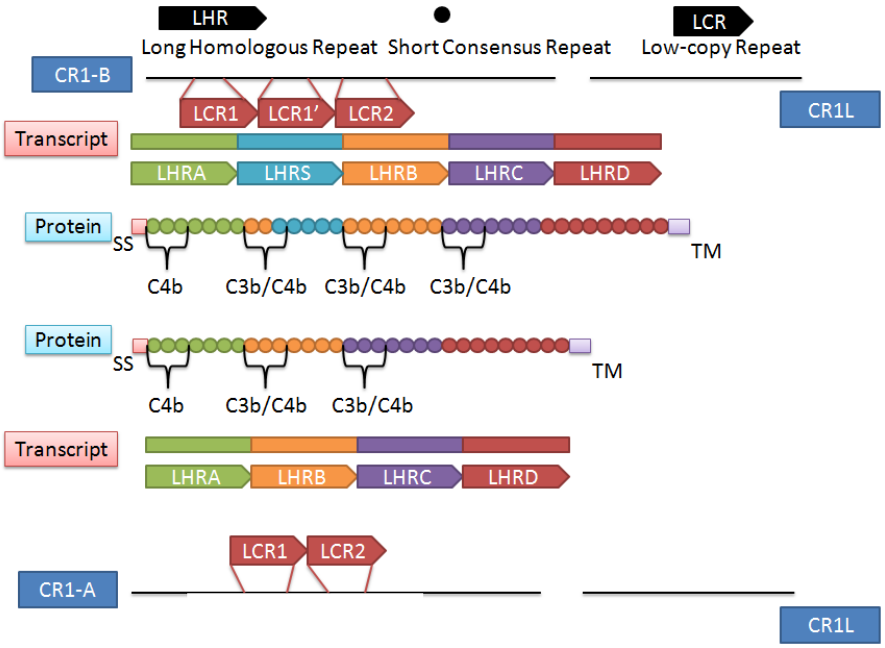


Figure 1.6: Two most frequent alleles of *CR1*: *CR1*-A and *CR1*-B. The *CR1*-B allele has an additional C3b/C4b binding site (LHRS) that shows increased inhibition of complement activity. Therefore, the *CR1*-A has less C3b/C4b binding site compared to the CR1-B allele because of an absence of an LCR region (LCR1'). Reproduced from Brouwers et al. (2012).

The smaller allele (*CR1*-C) has only one binding site for C3b/C4b that shows a loss of function related with this allele (Wong 1989; Liu and Niu 2009). It has been reported that there is an association between the *CR1*-B allele and Alzheimer's disease risk. The idea supporting this study is that *CR1*-B has an additional binding site for C3b/C4b therefore it has increased inhibition of complement activity, increasing AD risk in these allele carriers (Brouwers et al., 2012).

1.6 The Complement C3b/C4b Receptor 1 (*CR1*) gene

The complement receptor 1 (*CR1*) gene is located at chromosome 1q32.2 in a gene cluster encoding complement-related proteins (plasma complement regulatory proteins, C4 binding protein and factor H) (Weis et al., 1987). The *CR1* gene is expressed at high levels in myeloid cell lines, whole blood, breast, ovary and spleen and on the surface of erythrocytes, whereas it is expressed at low levels in pancreatic islets, brain cortex, hypothalamus and salivary glands (Kent et al., 2002). It has three types of polymorphisms which are sequence (Knops Blood groups), size and quantity (Stoute, 2011).

1.6.1 Knops Blood Groups

Sequence polymorphisms of the *CR1* gene are responsible for Knops blood groups. The Knops blood group consists of single antigen York (Yk^a) and the following allelic pairs: Knops a and b (Kn^a and Kn^b), McCoy a and b (McC^a and McC^b), Swain-Langley/Vil (Sl1 (Sl^a) and Sl2 (Vil)) and Sl3 (Sl3+ and Sl3-) and KAM (KAM+ and KAM-) as shown in Table 1.2 (Helgeson et al., 1970; Covas et al., 2007; Veldhuisen et al., 2011). It has been discovered that CR1 protein carries Kn^a, McC^a and Sl^a (McC^c) and also Yk^a (Petty et al., 1997) blood group antigens (Rao et al., 1991). A SNP at position 4646A > G (Asn1540Ser) in the Knops blood group system was recently identified (Covas et al., 2007) and 12 haplotypes were defined by this SNP in combination with SNPs determining Kn, McC, Sl, Sl4, Sl5 and KCAM. The haplotypes 1 (4646A, Kna, McCa, Sl1, Sl4, KAM+), haplotype 2 (4646A, Kna, McCa, Sl1, and KAM–) and haplotype 3 (4646A, Kna, McCa, Sl2, Sl4, KAM–) are the most frequent haplotypes in all populations (Covas et al., 2007). Because of the differences in haplotype frequencies, Covas et al. (2007) suggested that haplotype 2 (similar frequency in all populations) to be ancestral whereas haplotype 3 (higher frequency in the African group compared to other populations) was suggested to have increased in Africa because of positive selection. It has been found that the Sl:2 (Swain-Langley) phenotype (higher frequency in African-derived populations) displayed less binding to the parasite rosetting ligand Plasmodium falciparum erythrocyte membrane protein 1 (Rowe et al., 1995). This phenotype is found in less than 1% of Europeans, whereas in African-Americans and West Africans, it is found at ~ 40% and 70%, respectively. Therefore, it has been claimed that this polymorphism might provide protection against severe malaria as it is selected for in malaria-related regions (Moulds, 2002). Besides, it has been shown that Sl2 and McC^b are associated with protection against severe forms of malarial disease in Papua New Guinea (Cockburn et al., 2004). In Western Kenya, a protective relationship between Swain-Langley Knops blood group allele (Sl2) and cerebral malaria has been found. It has also been shown that a second blood antigen, McCoy McC^b is associated with severe malaria in Western Kenya (Thathy et al., 2005). McC^b and Sl2 are common in Africans whereas they are very rare in Europeans which might indicate that they are under selection pressure (Barreiro et al., 2008; Moulds et al., 2000; Moulds et al., 2004; Thathy et al., 2005). However, some studies showed no association between

malaria and Knops blood group alleles. For example, no effect of either the Sl2 (Jallow et al., 2009) or McC^{b} alleles has been observed in severe malaria in the Gambia and no effect of the alleles McC^{b}, Sl2, Kn^{b}, Kam- or combined haplotypes has been observed in southern Ghana (Gandhi et al., 2009; Hansson et al., 2013; Zimmerman et al., 2003).

The Knops blood group system continues to expand and all new findings provide a better understanding for this blood group system. For example, an additional polymorphism which is Q981H (substitution in SCR16) (rs200082366) (Birmingham et al., 2003) outside of the complement binding site has been discovered. Intriguingly, it has been found that the Q981H substitution is uncommon in Africans and in Europeans whereas it is common in Asians (Thomas et al., 2005), causes increased binding of C4b (Birmingham et al., 2003) and might suggest a selection pressure. In addition, Moulds et al. (2005) reported the antigen formerly known as Sl^{a} is subdivided. At amino acid 1601 (CR1 protein) Sl3 requires both R1601 and S1610 and also SL4 and SL5 hypothetical epitopes represented by S1610 and T1610, respectively. Also, it has been found that the absence of the Yk^{a} is caused by a mutation in Exon 26 of the *CR1* gene (Table 1.2). This 4223C>T mutation results in a 1408T>M change at the protein level (Veldhuisen et al., 2011).

Table 1.2: Antigens of the Knops blood group system encoded by the *CR1* gene: Different base changes in specific locations of exon 29 of *CR1* define different allelic pairs. This change determines prevalence of specific amino acids at different positions of antigens. Taken from Covas et al. (2007).

Blood Group Antigens	Antigens	Protein Variation	DNA Variation	Location	Reference
Knops (Kn^a and Kn^b)	Kn^a	NP_000564.2:p.Val1561Met (Val)	rs41274768-G	Exon 29 of *CR1*	Moulds et al., 2004
	Kn^b	NP_000564.2:p.Val1561Met (Met)	rs41274768-A		
McCoy ($McCoy^a$ and $McCoy^b$)	$McCoy^a$	NP_000564.2:p.Lys1590Glu (Lys)	rs17047660-A	Exon 29 of *CR1*	Moulds et al., 2001
	$McCoy^b$	NP_000564.2:p.Lys1590Glu (Glu)	rs17047660-G		
Swan-Langley (Sl^a/Sl1 and Vil/Sl2)	Sl^a/Sl1	NP_000564.2:p.Arg1601Gly (Arg)	rs17047661-A	Exon 29 of *CR1*	Moulds et al., 2001
	Vil/Sl2	NP_000564.2:p.Arg1601Gly (Gly)	rs17047661-G		
KAM (KAM+ and KAM-)	KAM+	NP_000564.2:p.Ile1615Val (Ile)	rs6691117-A	Exon 29 of *CR1*	Moulds et al., 2005
	KAM-	NP_000564.2:p.Ile1615Val (Val)	rs6691117-G		
York (Yka^+/Yka^-)	Yka^+	NP_000564.2:p.Thr1408Met (Thr)	rs3737002-C	Exon 26 of *CR1*	Veldhuisen et al., 2011
	Yka^-	NP_000564.2:p.Thr1408Met (Met)	rs3737002-T		

In a recent study, it has been shown that Kn^b allele and KAM^+ allele combinations are associated with susceptibility to *P.falciparum* infection in a Brazilian Amazon population (Fontes et al., 2011).

1.6.2 Size Polymorphism

There are four co-dominant alleles of the *CR1* gene which are responsible for different length protein products. The different protein isoforms vary in size by units of 30 kDa. CR1-A (190 kDa, also known as CR1-F) and CR1-B (220 kDa, also known as CR1-S) are the most frequent alleles CR1-C (160 kDa, also known as CR1-F') and CR1-D (250 kDa) are rarer. They show different frequencies in different populations. For instance, in Europeans, the CR1-A and CR1-B have frequency of 0.87 and 0.11, respectively. These

frequencies shows differences as 0.82 and 0.11 in African-Americans, 0.89 and 0.11 in Mexicans (Moulds et al. 1996), and 0.91 and 0.08 in Asian Indians (Katyal et al. 2003). The rarer alleles, such as CR1-C and CR1-D are infrequent in all populations (Eikelenboom and Stam, 1982; Moulds et al., 1996; Wong et al., 1983). This size variation is due to the deletion and insertion during the imperfect chromosome crossing-over process (Crehan et al., 2012b). This is thought to involve the variable number of 18-kb long low-copy repeats (LCRs: LCR1, LCR2 and LCR1') in the *CR1* coding region. Therefore, *CR1* alleles are created by the structural variation in low-copy repeat (LCR) (18-kb) region of *CR1* gene (Figure 1.7). The deletion of LCR1 or duplication or triplication of LCR1 (LCR1') causes different alleles of *CR1*.

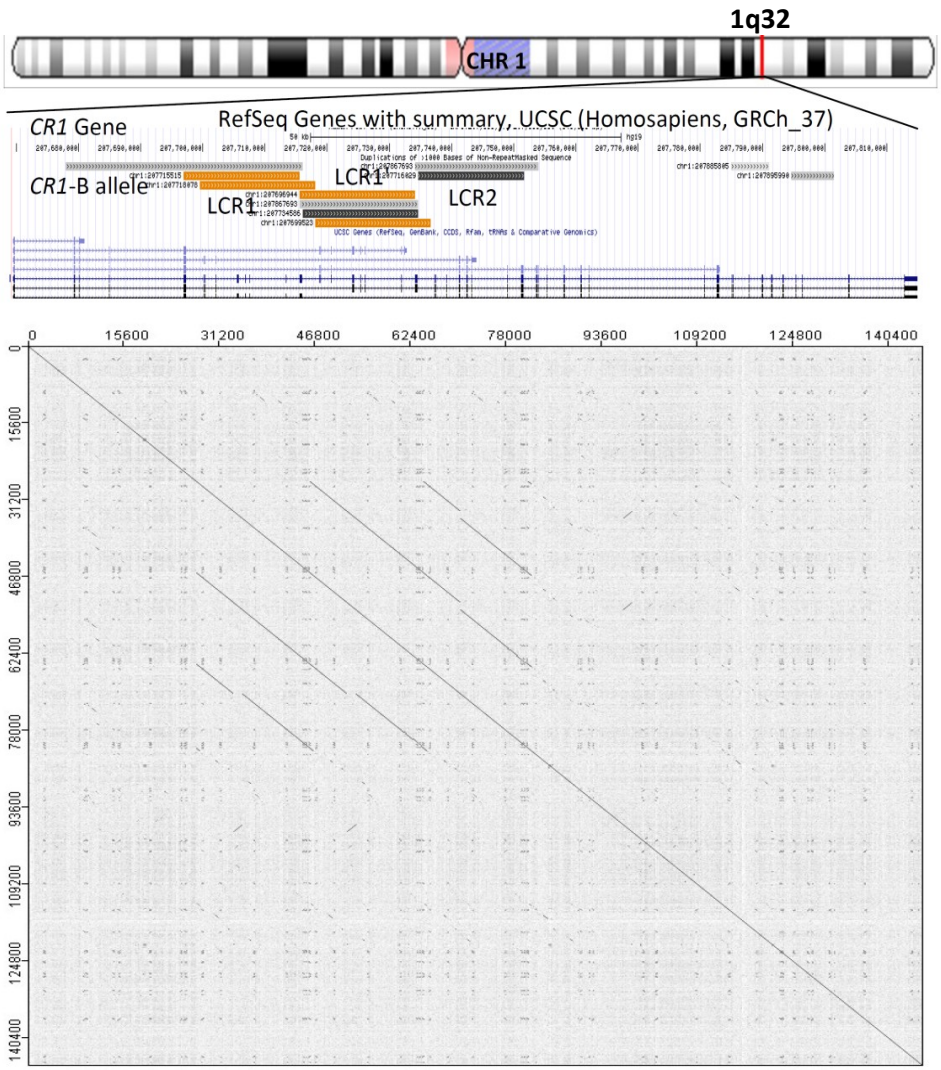


Figure 1.7: Dot plot analysis of *CR1* gene (exons and introns) again itself. Dot plot analysis of entire sequenced region against itself was provided by Dr Edward J Hollox. The upper panel shows chromosomal location of *CR1*, second panel shows seven *CR1* gene annotations from different transcripts and lower panel shows segmentally duplicated regions on *CR1*; the lines shows homologous sequences. Dot plot analysis of; the entire sequenced region against itself was performed with Jdotter (http://athena.bioc.uvic.ca/virology-ca-tools/jdotter/) to identify regions of local similarity between sequences after masking repetitive sequences. The orange and grey bars are used to distinguish levels of similarity: Light to dark grey displays 90-98% similarity and light to dark orange displays greater than 99% similarity.

These regions (LCR1, LCR2 and LCR1') are located in the *CR1*-coding region. The extra copy of LCR1 provides an extra set of C3b/C4b binding and cofactor activity sites so it has been proposed that these alleles have different functional roles in the immune complement cascade (Brouwers et al., 2012). In addition, loss of an C3b binding site (CR1-F' or CR1-C allotype) has been suggested to alter function, as the CR1-C with only one C3b binding site was at least 10-fold less effective (compared to CR1-A) in the inhibition of the alternative pathway C3 and C5 convertases. Therefore, the capacity

of the CR1-C allotype to bind opsonized immune -complexes to inhibit the formation of the alternative pathway C3 and C5 convertases may be compromised (Wong and Farrell., 1991). In the hg19 assembly of the *CR1* gene, LCR1 and LCR1' are 99% similar whereas LCR1 and LCR2 are 97% similar at the DNA level (Figure 1.8).

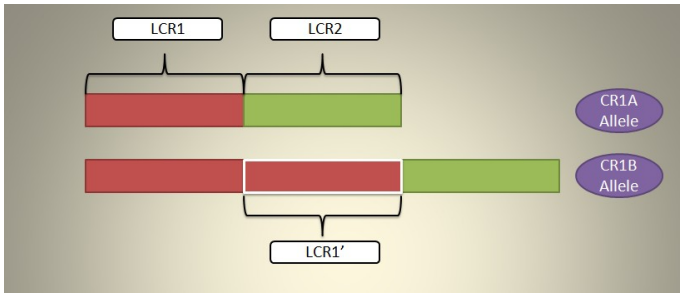


Figure 1.8: CR1 coding region of CR1-A isoform contains low copy repeats, LCR1 and LCR2 whereas CR1-B isoform includes LCR1, LCR2 and LCR1'. This extra LCR is associated with an additional long homologous repeat (LHR) in the *CR1*-B allele. Reproduced from Crehan et al. (2012b).

1.6.3 Quantitative Polymorphism

Another CR1 polymorphism is a quantitative polymorphism. Complement expression level on erythrocytes is genetically determined by the presence of H (High) and L (Low) expression alleles which give rise to low expressors (LL), intermediate expressors (HL) and high expressors (HH) (Rowe et al., 2002). Quantitative expression of CR1 on erythrocytes is linked to the site of *Hin*dIII restriction fragment length polymorphism (RFLP) of the *CR1* gene. The *Hin*dIII RFLP is due to a point mutation in intron 27 (rs2274567: A/T allele) (Wilson et al., 1986) of the *CR1* gene and it has also been linked to several other substitutions in the coding nucleotide sequence.

In Europeans, CR1 levels are genetically determined by three different single nucleotide polymorphisms which are in exon 22, intron 27 and exon 33 of the *CR1* gene (Xiang et al., 1999). They only control the erythrocyte CR1 level (high (H) or low (L) level CR1 haplotypes) as they have no effect on other cell types (Wilson et al., 1986).

In addition, it has been suggested that the genetically determined low CR1 density on erythrocytes may be a risk factor for developing a more severe form of malaria in Thailand (Nagayasu et al., 2001).

1.7 CR1 and Malaria

Repetitive cell cycle growth of *Plasmodium falciparum* (a unicellular organism) is transmitted by Anopheles mosquitoes into erythrocytes and causes malaria, which is responsible for almost 1 million deaths per year. Severe malarial anaemia and cerebral malaria are the reasons for greatest numbers of child deaths in sub-Saharan Africa. During the life cycle of the parasite, sporozoites are injected infected by female Anopheles mosquitoes and they are carried to the liver. In the liver, sporozoites infect hepatocytes and divide into thousands of merozoites. As a consequence of the bursting process 1-2 weeks after infection, they join the blood circulation where they can infect blood cells and multiply asexually. These infected blood cells become gametocytes which are ingested by mosquitoes and undergo sexual reproduction in the mosquito's midgut which causes more sporozoite production before migrating to the salivary gland where the parasite finishes its life cycle) (Figure 1.9) (Stoute, 2011).

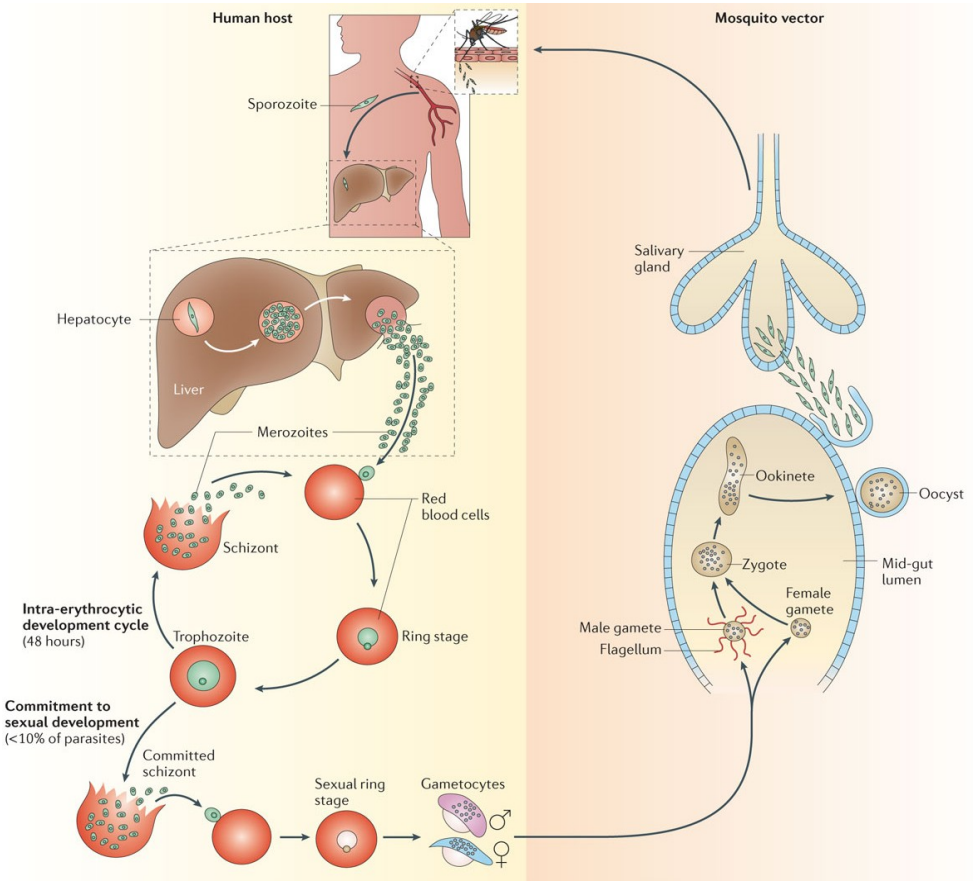


Figure 1.9: Repetitive cell cycle growth of *Plasmodium falciparum*. After the bite of an infected *Anopheles* spp. mosquito sporozoites are transmitted to the human host. Sporozoites travel and develop in the liver. Tens of thousands of merozoites are released into the blood and invade red blood cells. The parasites show a repeating asexual multiplication (the intra-erythrocytic developmental cycle) and progress a ring of trophozoite and schizont stages in turn. During each cycle, a small proportion (<10%) of parasites begin to develop into the sexual form of the parasite (gametocyte), which is required for productive transmission to the mosquito host. Sexual development is believed to occur in schizonts, which form sexual rings and then gametocytes following reinvasion of red blood cells. Between 10 and 12 days of development, mature male and female gametocytes are taken up by the mosquito, in which they undergo the sexual phase of the life cycle. Gametocytes differentiate into gametes and, after fertilization, the resulting zygote (which develops into a motile form known as the ookinete) forms an oocyst. When the oocyst ruptures, haploid sporozoites are released and migrate to the salivary glands. Then, they can be transmitted to humans. This figure is taken from Josling and Llinás (2015).

In addition, during blood stage infection of the erytrocytes, *P.falciparum* merozoites invade erythrocytes by releasing invasion ligands from apical organelles that interact with receptors on the erythrocyte surface. *P.falciparum* has two invasion ligands which are *P.falciparum* reticulocyte-binding homologues (PfRh) and erythrocyte binding antigens (EBAs). These proteins (PfRh and EBA) help facilitate the use of alternate invasion pathways by *P.falciparum* which can cause repeated and chronic invasions. There are two invasion pathways which are sialic acid-dependent invasion and sialic

acid-independent invasion. The PfRH ligands (PfRh1 to 5) are located in the rhoptries of merozoites. Specifically, PfRh4 binds to CR1 and is important for acid-independent invasion (Reiling et al., 2012) (Figure 1.5).

CR1 has also been shown to be involved in rosetting of uninfected erythrocytes to *Plasmodium falciparum*-infected cells through its interactions with *P.falciparum* erythrocyte membrane protein 1 (PfEMP-1) (Moulds et al., 1997) and rosetting formation is associated with severe malaria (Carlson et al., 1990; Kun et al., 1998; Rowe et al., 1995; Treutiger et al., 1992). In addition, Knops blood group antigens (*CR1* polymorphism) show marked frequency differences between Europeans and Africans. Therefore, they may have crucial importance in the mechanism of *P.falciparum* malaria infection such as reducing susceptibility to malaria pathogenesis. Knops blood group variation has been suggested to affect malaria as discussed in section 1.6.1 Knops Blood Groups.

1.7.1 Severe Malaria

In Africa, severe malarial anaemia usually affects children under 5 years old (Snow et al., 1997; Marsh and Snow, 1999) and is the largest cause of the childhood death in the world (Stoute, 2011). Severe malaria has alternative mechanism for the destruction of red cells as even the patients treated for malaria still have red cells reduction (Camacho et al., 1998).

C3b and IgG deposition which leads to erythrophagocytosis, is important for red cell destruction during normal cell senescence. These molecules role in red cell destruction during malaria infection has been investigated and C3 and IgG deposition has been found on red cells of children with severe malarial anaemia (Stoute, 2011).

In a few studies, it has been shown that CR1 and CD55 levels are lower in the erythrocytes of children with severe malaria than in those of children with uncomplicated malaria or uninfected (Stoute et al., 2003; Waitumbi et al., 2000; 2004). On the other hand, Cockburn and colleagues have suggested that CR1-deficient individuals should be protected against malaria as they show reduced *Plasmodium falciparum* rosetting (Cockburn et al., 2004). Al yaman et al. (1995) has found that there is no association between rosetting and severe malaria.

Rosetting is an adhesion property of parasitized erythrocytes that enables them to bind to unparasitized erythrocytes and makes clumps of cells (David et al., 1988). It is believed that rosetting is contributing to malaria by causing obstruction and impaired impaired tissue perfusion (Kaul et al., 1991; Nash et al., 1992). Rosetting association with severe malaria is quite varied geographically. For instance, rosetting is associated with severe malaria in many of studies in Africa (Carlson et al., 1990; Ringwald et al., 1993) whereas no association with rosetting has been found in Southeast Asia (Ho et al., 1991) or Papua New Guinea (al-Yaman et al., 1995).

Uninfected erythrocytes are involved in the rosetting process via the parasite ligand PfEMP1 (Figure 1.5) on the surface of *P. falciparum* infected erythrocytes which can attach to a variety of receptors on uninfected erythrocytes such as complement receptor 1 (CR1) and cause rosetting (Rowe et al., 1997). The number of complement receptor 1 molecules on erythrocytes can change between individuals in the range of 50-1,200 molecules per cell (Rowe et al., 2002; Wilson et al., 1986). Erythrocyte CR1 level depends on both *CR1* exon 22 genotype and α-globin (α-thalassaemia) genotype (Cockburn et al., 2004). Cockburn and colleagues also showed that HL (intermediate expressor) individuals for the exon 22 polymorphism were highly protected against severe malaria (odd ratio (OR) =0.33 and P=0.01) whereas homozygotes for the LL genotype (low expressor) have a reduced odds ratio (though not significantly). In a related study, Nagayasu et al. (2001) has reported that LL genotype is a risk factor for severe malaria in Thai adults.

1.7.2 Cerebral Malaria

In Kenya, Thathy et al. (2005) reported an association between the S/2/2 genotype of the Swain-Langley and protection against cerebral malaria. In 2008, Teeranaipong et al. carried out a study in Thailand and identified 17 polymorphisms in the promoter region of *CR1* which were associated with protection against cerebral malaria and increased cell CR1. Therefore, rosetting may not have a role in the protection against severe malaria because lower CR1 is related to lower risk of cerebral malaria in rosetting. In the Indian study of Rout et al. (2011), the results are completely different from the Thailand results. The increased level CR1 which was determined by

genotyping of the intron 27 *Hin*dIII allele, the exon 22A3650G and the exon 33G5507G was associated with cerebral malaria.

1.7.3 GWAS Results implicating Polymorphisms Surrounding CR1 and Erythrocyte Sedimentation Rate (ESR)

Recent GWAS studies have shown the variants within *CR1* or surrounding 250 kb region to be related to several diseases and ESR (Table 1.3). ESR is a blood test to measure how long it takes for erythrocytes to fall to the bottom of a test tube. The quicker erythrocytes fall, the more likely it is there are high levels of inflammation. This test help diagnose conditions associated with inflammation such as arthritis, endocarditis, Crohn's disease, giant cell arteritis. It can also be used to confirm whether you have an infection (NHS, 2017). Lambert et al. (2009) used 2,032 AD cases and 5,328 controls in first stage of their study and 3,978 cases and 3,297 controls and with the combined data, showing that two SNPs at the *CR1* locus (rs6656401 and rs3818361) reached a genome-wide significance. The SNP rs6656401 had an OR of 1.21 (95% CI 1.14-1.29) and a p value $3.5x10^{-9}$ whereas the SNP rs3818361 had an OR of 1.19 (95% CI 1.11-1.26) and a p value $8.9x10^{-8}$. Another GWAS study of Hollingworth et al. (2011) used 6,688 AD cases and 13,685 controls in first stage of their study and 4,896 cases and 4,903 controls and with the combined data, one SNP at the *CR1* locus (rs3818361) reached genome-wide significance. The SNP rs3818361 had an OR of 1.18 (95% CI 1.13-1.24) and p value $3.7x10^{-14}$. Another GWAS study of Naj et al. (2011) used 8,309 AD cases and 7,366 controls in first stage whereas 3,531 AD cases and 3,565 controls in second stage and by combining these stages made another analysis named as joint analysis. They used stage 3 samples for meta- analysis. One SNP at *CR1* locus (rs6701713) has reached a genome-wide significance. The SNP rs6701713 had an OR of 1.16 (95% CI 1.11-1.22) and a p value $4.6x10^{-10}$ for meta-analysis while for joint analysis, rs6701713 had an OR of 1.17 (95% CI 1.12-1.23) and the p value $5.2x10^{-11}$. Many other studies have shown the relationship of CR1 to AD. For example, Antunez et al. (2011) group showed that *CR1* polymorphism rs3818361 increases the risk of Alzheimer's disease although its effect size can be smaller than previously estimated. 2440 samples from individuals from Spain showed an association between rs3818361 and late-onset AD (OR=1.114, 95% confidence interval: 0.958-1.296, P= 0.16) whereas

samples from genome-wide association studies supported the effect of rs3818361 (odds ratio=1.180, 95% confidence interval: 1.113-1.252, P<0.05) on Alzheimer's disease (Antunez et al., 2011).

In addition, Kullo et al. (2011) has found one SNPs at *CR1* locus, rs12034383 (n=5,628 and $p=2x10^{-18}$ also n=7,607 and $p=5x10^{-28}$), and one surrounding *CR1* (*CR1L*), rs7527798 (n=7,607 and $p=2x10^{-9}$), to be associated with ESR.

Table 1.3: SNPs surrounding 250 kb region of *CR1* gene and their correlation with some diseases which obtained from GWAS. (-) indicates no information is available.

SNP	Nearest genes	GWAS Disease Association	Allele	Odds Ratio or beta	References
rs6656401	*CR1*	Alzheimer's Disease	A	1.21	**Lambert et al., 2009**
rs3818361	*CR1*	Alzheimer's Disease	A	1.18/1.19	**Hollingworth et al., 2011** **Lambert et al., 2009**
rs6701713	*CR1*	Alzheimer's Disease	A	1.16	**Naj et al., 2011**
rs12034383	*CR1*	Erythrocyte Sedimentation Rate	G	0.13	**Kullo et al., 2011**
rs12034598	*CR1*	Erythrocyte Sedimentation Rate	A	-	**Naitza et al., 2012**
rs17045328	*CR2*	Type 2 Diabetes mellitus (T2D)	G	1.38	**Kooner et al., 2011**
rs4844614	*CR1L*	Metabolic Trait (low-density lipoprotein)	A	0.10	**Sabatti et al., 2009**
rs7527798	*CR1L*	Erythrocyte Sedimentation Rate	-	0.10	**Kullo et al., 2011**

1.8 CR1 and Alzheimer's Disease

Alzheimer's disease is neurological disease which affects mostly old people (Small and Duff, 2008) and cause death of the patient between 3 to 9 years after diagnosis (Querfurth and LaFerla, 2010). Alzheimer's Disease can be divided into two classes as early-onset AD (EOAD) and late-onset AD (LOAD) according to the first disease symptoms' existence before and after age 65 (Kowalska, 2004). It shows the symptoms of memory disorders, cognitive decline and also loss of autonomy. Neurofibrillary degeneration and amyloid deposits are the most commonly observed neuropathological lesions in Alzheimer's disease. Neurofibrillary degeneration is caused by intraneuronal accumulation of hyperphosphorylated Tau proteins whereas amyloid plaques which are primarily composed of Aβ peptides. Aβ peptides are generated from amyloid precursor protein (APP) and tau-associated neurofibrillary tangles (Small and Duff, 2008; Zhu et al., 2015). Mutations in three genes are related with rare hereditary early-onset of Alzheimer's disease. These genes are *APP*, encoding amyloid precursor protein, on chromosome 21; *PS1*, encoding presenilin 1, on chromosome 14; and *PS2*, encoding presenilin 2, on chromosome 1 (Hardy and Selkoe, 2002). These mutations do not fully explain all cases of Alzheimer's disease (Campion, 1999). Some genetic studies have showed that the Ɛ4 allele of *APOE* is a susceptibility locus for late-onset Alzheimer's disease (Farrer, 1997). According to twin studies, 60% of disease susceptibility results from the genes (Gatz, 2006) and it has been found that 50% of this is related to *APOE* (Ashford and Mortimer, 2002). 550 genes are still under study as possible genes for susceptibility to the disease (Bertram et al., 2007). Recently, in genome-wide association studies it has been found that common genetic variants in *CLU, CR1, PICALM, ABCA7, BIN1, EPHA1, CD33, CD2AP* and *MSA4* contribute to increased risk in the progression of LOAD (Harold et al. 2009; Lambert et al., 2009; Hollingworth et al., 2011; Naj et al., 2011). There is recent evidence that shows the complement system is involved in Alzheimer's disease. Veerhuis (2011) reported that the expression of the complement factors of classical complement pathway are upregulated in AD-affected brain. Although complement proteins exist in high levels in the blood, they cannot pass through the blood-brain barrier in normal brain but it has been shown that they can enter the brain in many degenerative diseases including AD. Rogers et al. (1992) showed that Aβ plaques can activate the classical complement

pathway in the absence of the antibody in AD brain. Aβ plaques are accumulated in AD brains and bind to specific parts within the collagen-like domain of C1q, activating the classical complement pathway. In addition, fibrillar Aβ can bind to C1q and can activate classical and alternative complement pathways. It has been found that neurofibrillary tangles and neuritic plaques can activate the complement system. For instance, Tau which is the major part of the neurofibrillary tangles, can activate the classical complement pathway (Crehan et al., 2012a).

The SNPs (in *CR1* or surrounding) which are found to be risk loci for AD obtained from GWAS studies has been shown previously (Table 1.3).

1.8.1 The Complement System's Role in the Brain

The brain is protected by the blood-brain-barrier (BBB). This shield has three barrier sites, which are endothelial cells in the cerebral capillaries, the arachnoid barrier formed by the arachnoid multi-layered epithelium, and the blood cerebrospinal fluid (CSF) barrier formed by the CSF-secreting choroid plexus epithelium. BBB-related cells (including pericytes and astrocytes) make a major contribution to this integrity. This barrier blocks the access of immune cells (B and T lymphocytes) to the central nervous system and also diminishes influx of plasma proteins (Abbott et al., 2010).

The complement system, which is composed of 30 proteins, is a regulator of inflammation and very important for the host defence system (Morgan and Gasque, 1996) (see section 1.4 The Complement System)). The main source of complement proteins is the liver. Most of the complement proteins are unable to pass the brain parenchyma (the functional tissue in the brain) unless the blood brain barrier (BBB) integrity is disrupted. In addition, some macromolecules can escape the barriers using extracellular routes (Broadwell and Sofroniew, 1993). Because of these reasons the local synthesis of the complement proteins by brain cells has a crucial role in local defence of the brain. It is known that neurons, astrocytes, migroglia and endothelial cells are able to synthesise most of the complement proteins (Mcgeer et al., 2003). Neuronal and glial cells can locally express complement receptors and also almost all complement components in response to injury or developmental signals (Tenner et al., 2011). Complement has critical roles in central nervous system. It has been found that

complement proteins can regulate proliferation and regeneration in various tissues that they may show similar functions in brain as neuronal stem cells differentiate and migrate in response to complement (Ricklin et al., 2010). In addition, complement activation products can regulate synapse formation during brain development. Also, neurons which were isolated from the developing eye showed high levels of C1q mRNA expression (Chu et al., 2010; Stevens et al., 2007). C1q alone, or together with C3, promotes migroglial clearance of misfolded proteins, apoptotic neurons, and damaged cells and promotes cytokines profiles to suppress probable neurotoxic inflammatory gene expression (Fraser et al. 2010; Trouw et al., 2008). Therefore, the complement cascade promotes neural development as well as stimulating chronic inflammatory response contributing to neurodegeneration. In addition, proteins related to C1q, cerebellins and C1q, expressed in cerebellum and other regions of developing and mature brain may serve as regulators of synapse development and maintenance (Bolliger et al., 2011). C1q is neuroprotective to certain toxic agents such as fibrillar amyloid and serum amyloid P and may also improve neuronal survival (Pisalyaput and Tenner et al., 2008).

Besides, complement can have a critical role during infection by bacteria, fungi and viruses in immune privileged brain, as complement fights and kills of invading pathogens. As excessive complement activation can be detrimental to bystander cells, it needs to be balanced critically. Some of the central nervous system invading microorganisms have developed mechanisms to avoid destruction by complement. One of these mechanisms is mimicry of human complement regulatory proteins, factor H and factor I (Cregs), that enables control and down-regulation of complement against invading microorganism (Cooper and Nemerow, 1989). The other mechanism invading microorganisms use is the membrane-bound complement receptors and Cregs to enter the host cell (Speth et al., 2002). For example, the meningitis-causing bacterium *Neisseria meningitides* invades the central nervous system by binding membrane-bound Creg CD46 which interacts with bacterium pili to cross the brain barrier (Johasson et al., 2003) and the gram-negative bacteria *Escherichia coli K1* avoids destruction by complement by binding to C4bp and regulating degradation of C3b and C4b complement components (Wooster et al., 2006). Also, HIV-1 is detected

in the brain of 85% of patiens who died with AIDS (Carel et al., 1989) and HIV-1 uses Cregs CD46, CD55 and CD59 upon budding from host cells (Frank et al., 1996) and binds to soluble Creg fH to escape from complement-mediated lysis of virion particles (Stoiber et al., 1995).

1.8.2 The Potential Role of CR1 in AD

Characteristics of the pathology of AD brain include accumulation of neurotoxic amyloid-β (Aβ) peptides generated from the amyloid precursor protein (APP), tau-associated neurofibrillary tangles (Eikelenboom et al., 2006) and inflammation, including complement system activation (Jiang et al., 2012). Aβ peptides are 40 to 42 amino acid long peptides and derived from a large Aβ precursor protein which can be found in both brain and peripheral tissues such as the heart, liver, lymph nodes and salivary glands (Joachim et al., 1989). Aβ plaques are capable of activating the classical complement pathway in the absence of an antibody in the AD brain (Roger et al., 1992) as Aβ peptides accumulate in the collagen-like domain of C1q (Jiang et al., 1994). There are several ways that *CR1* may relate to AD. It has been suggested that CR1 contributes to AD by regulating clearance of Aβ in two ways: peripherally by erythrocytes and directly in brain. In serum, complement can be activated by Aβ.

On erythrocytes, CR1's function is to transport opsonised immune complexes from the circulatory system. CR1 also can remove Aβ if it binds to complement component C3b complex with same the mechanism. In AD brain, the increased level of Aβ peptides can activate C3 convertase which is cleaved into C3bi. It surrounds Aβ and leads to capture of Aβ by the phagocyte expressing CR3 (the receptor for C3bi). As CR1 cannot suppress the cleavage of C3b in the presence of Aβ, this suggests CR1 is functioning in microglial phagocytosis. Therefore, Crehan et al. (2013) showed that microglial-CR1 can contribute to neuronal death during AD. They found that microglia showing an activated phenotype displayed an increase in CR1 expression which is related to an increase in microglial intracellular superoxide generation, tumour necrosis factor and interleukin-1β secretion.

Most of the complement proteins can be produced by glial cells and neurons in the central nervous system and their production is increased in AD. Also, complement

proteins and membrane attack complex (active form) are related to AD pathology such as dying neurons and amyloid plaques. Therefore, these indicate that reducing complement activation in brain could be protective and CR1 could have a critical role in the regulation of the complement system during AD, as it could prevent brain damage from occurring (Akiyama et al., 2000). However, this has been contradicted by another study. In a mouse model (human amyloid precursor protein (hAPP) transgenic mice) Wyss-Coray et al. (2002) showed that inhibiting complement activation increases plaque deposition and neurodegeneration, as complement regulator C3 reduces plaque load. This leads to the conclusion that complement activation may be protective. In these circumstances, it should be expected that larger forms of CR1 are more effective in preventing the protective complement cascade. This research is supported by the findings of Brouwers et al. (2012). They found that the presence of *CR1*-B increases Alzheimer's disease (AD) risk with 30% in *CR1*-B carriers in a French late-onset Alzheimer's disease (LOAD) cohort. They suggested that this is caused by the presence of *CR1*-B as it increases the number of C3b/C4b binding and cofactor activity sites.

Finally, it has been found that Tau, which is the major part of the neurofibrillary tangles observed in AD, is able to activate the classical complement pathway (Shen et al., 2001). Therefore, it has been suggested by Zotova et al. (2008) that the chronic and low level of inflammation during AD is caused by these plaques' and tangles' ability to activate the complement pathway. It has been suggested that CR1's linkage with AD may act through unknown pathways to cause the supersensitivity of neurons to Aβ-induced toxicity or to cause hyperaccumulation of tauopathy. Unpublished observations of Haroutunian's laboratory showed that *CR1* mRNA levels correlate better with neurofibrillary tangle density and phosphorylated tau abundance than with neuritic plaque density (Haroutunian et al. 2013). More studies are needed to explore the linkage between tauopathy and *CR1*.

1.9 Aims of the Project

- The development of novel robust PRT assays to investigate and characterize the pattern of CNV at the *CR1* gene in worldwide populations.
- Comparison and contrast of the different methods used for calling LCR *CR1* CNVs.
- Using robust PRT assays to genotype LCR *CR1* CNV in large case-control studies to seek associations between CNV *CR1* and diseases such as malaria and AD.
- Development of an allele-specific hybridisation assay (ASO) to genotype alleles of the Knops blood group system in order to understand their relationship to malaria.
- Development of a junction fragment PCR method (LNA) to detect the existence of the *CR1*-B allele, which helps to make more accurate and distinct copy number calling for the samples giving the PRT copy number between 4 and 5.

2 MATERIALS AND METHODS

2.1 DNA samples used

2.1.1 HapMap Samples

The International HapMap project started in October 2002. The main aim of this project was to study the genetic diversity of four different populations which are the Yoruba from Ibadan in Nigeria (YRI-90 individuals), the Japanese from Tokyo (JPT-45 individuals), the Han Chinese from Beijing (CHB-45 individuals) and the northern and western Europeans from Utah (CEU-90 individuals) (The 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 called a trio) whereas the total of 90 HapMap Asian samples consist of 45 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/). The blood samples were converted into lymphoblastoid cell lines by Epstein Barr Virus transformation of peripheral blood lymphocytes by the Coriell Institute (http://ccr.coriell.org). DNA and cell lines which were from the samples for research projects that have been approved by the appropriate ethics committees were provided by the Coriell Institute. These samples were typed for estimating *CR1* CN using the Paralogue ratio test (PRT) method.

2.1.2 ECACC Human Random Control (HRC) samples

The HRC DNA samples represent a population of 480 UK native blood donors (https://www.phe-culturecollections.org.uk/products/dna/hrcdna/hrcdna.jsp). The DNA samples were extracted from lymphoblastoid cell lines derived by Epstein Barr Virus (EBV) transformation of peripheral blood lymphocytes from unrelated single-donor blood samples. The genomic DNA was provided by Dr Edward Hollox as a solution at a standard concentration of 100 ng/µl in 10mM Tris-HCl buffer (pH 8.0) with 1mM EDTA. Mostly 5-10 ng/µl was used as working concentration in this study.

2.1.3 HGDP-CEPH Panel

The HGDP (Human Genome Diversity Project) collection is the most complete worldwide human DNA collection that is available to not-for-profit researchers (http://www.cephb.fr/en/hgdp_panel.php). A resource of 1063 cultured lymphoblastoid cell lines from 1050 individuals in 52 world populations is banked at the Fondation Jean Dausset-CEPH in Paris. The information for each of the cell lines is limited to sex of the individual and population and geographic origin. Previous studies identified atypical and duplicated samples and pairs of close relatives within the HGDP-CEPH panel, and three standardized subsets (H1048, H971 and H952) were recommended for most population-genetic studies (Rosenberg, 2006). The subset H971 was defined by avoiding first-degree relative pairs and all analysis was performed based on subset H971 in the present study.

2.1.4 Tori-Bossito Cohort

Tori Bossito (southern Benin) is a cohort of 656 infants with a parasitological (symptomatic and asymptomatic parasitaemia) and nutritional follow-up from birth to 18 months. Ecological, entomological and behavioural data were collected along the duration of the study (Port et al. 2011). David Courtin and André Garcia (Université Paris Descartes, Paris, France) provided us with 583 DNA samples which were collected from a rural area in Benin with two seasonal peaks in malaria transmission.

2.1.5 Tolimmunpal Cohort

Tolimmunpal is a cohort of 400 children which were followed from their birth to 24 months of age (as were their mothers during pregnancy). The Tolimmunpal (Immune Tolerance and Malaria) project provides data on 1179 women as they were followed up during their pregnancy which made it possible to obtain data based on malaria infection during pregnancy (contact of the foetus with soluble *P. falciparum* antigens) of the mother. The samples were collected from the Allada region of Benin. David Courtin and André Garcia provided us with 278 DNA samples (children) of this cohort.

2.1.6 EOAD Cohort

Prof Kevin Morgan (University of Nottingham, UK) provided us with 633 DNA samples which were 449 cases (and 184 controls) of early-onset Alzheimer's disease samples

collected from Europeans (collected from Nottingham, Bristol, Manchester, Oxford, Bonn and Southampton).

2.1.7 LOAD Cohort

Prof Kevin Morgan (University of Nottingham, UK) provided us with 2268 DNA samples which were 1263 cases (and 1005 controls) of late-onset Alzheimer's (LOAD) disease samples collected from Europeans (collected from Manchester, Nottingham, Oxford, Bristol, Belfast and Southampton).

2.2 Genomic DNA Extraction from Lymphoblastoid Cell Lines

2.2.1 Cell Line Samples

B-Lymphoblastoid cell lines (Appendix 1) were obtained from the Coriell Cell Repository having been established there by transformation of B-lymphocytes (isolated from peripheral blood mononuclear cells) with Epstein-Barr virus (EBV) using phytohaemagglutinin as a mitogen (Henderson et al., 1977). The lymphoblastoid morphology is small (7-9 micron) round cells that grow as loose aggregates in suspension. The cultures obtained from the Coriell Cell Repository were stored in liquid nitrogen (- 196°C).

2.2.2 Lymphoblastoid Cell Lines

A cryovial of cells was taken from liquid nitrogen and the cells were left on the ice (4°C) for 20 minutes. The cells were then thawed in a water bath (37°C) for half an hour. The entire contents of the cryovial was resuspended in prewarmed (37°C) fresh culture medium (RPMI 1640 with 2mM glutamine and 10% (v/v) foetal calf serum) and centrifuged at 3000g for 5 minutes. The supernatant was discarded and the cell pellet was resuspended in fresh culture medium (RPMI 1640 with 2mM glutamine and 10% (v/v) foetal calf serum) and was cultured in a humidified incubator at 37°C with 5% (v/v) CO_2 and media changed three times in a week.

2.2.3 Growing Lymphoblastoid Cell Lines

The lymphoblastoid cell lines were grown in RPMI 1640 medium. 2mM glutamine and 10% (v/v) foetal calf serum were supplemented in the medium. The cells were grown in the mixed medium inside the humidified incubator at 37°C with 5% (v/v) CO_2.

Lymphoblastoid cell lines were grown in suspension culture. When the medium changes colour from red to yellow that is an indicator of pH change (phenol red (phenolsulfonphthalein)), the cells are subcultured with fresh medium generally after two days. The cell aggregates are dissociated by gentle pipetting before transfer into new medium.

2.2.4 Storage of Lymphoblastoid Cell Lines

A cell suspension (4°C) was centrifuged at 400g for 5 mins at 4°C and the supernatant was discarded. The cells are resuspended in 1 ml of 4°C cryomix (90% [w/v] FCS and 10% [v/v] DMSO). Cell suspension (1 ml) was transferred into labelled cryovials. Cryovials were immediately placed in a cryopreservation chamber (Mr Frosty) and stored overnight at –80°C. Vials were then transferred to liquid nitrogen storage tanks and details were recorded on the cryostore database. B-Lymphoblastoid cell lines were stored in liquid nitrogen (-196°C) for future use.

2.2.5 Genomic DNA Extraction From Lymphoblastoid Cell Lines

The genomic DNA samples used as controls for PRT assays extracted from lymphoblastoid cell lines (Table 2.4).

Table 2.4: The control samples which are used in PRT assays.

Sample ID	Population of Origin	Sex/ age	Source of sample	Source of CN data	Conrad et al., 2010 CN	Optimized PRT1-3 CN
NA18507	YRI	Male/ nk	HapMap	Conrad et al., 2010	2	3
NA18555	CHB	Female/ nk	HapMap	Conrad et al., 2010	3	4
NA18517	YRI	Female/ nk	HapMap	Conrad et al., 2010	3	4
NA19239	YRI	Male/ nk	HapMap	Conrad et al., 2010	4	5
NA18572	CHB	Male/ nk	HapMap	Conrad et al., 2010	2	3
C0140	UK Caucasian	Female/ 40yr	HRC1	Our study (PRT1-3)	Not available	5
C0182	UK Caucasian	Female/ 39yr	HRC1	Our study (PRT1-3)	Not available	5

In order to isolate genomic DNA, a standard phenol/chloroform extraction protocol (Appendix 2) was used (Sambrook and Russell, 2001). All solutions that were used during the extraction process were freshly made. Genomic DNA was isolated from lymphoblastoid cell lines by proteinase-K digestion (20mg/ml) at 55°C for 2 hours, followed by phenol/chloroform extraction and ethanol precipitation (Appendix 2).

2.3 Copy Number Analysis of *CR1*

2.3.1 Paralogue Ratio Test (PRT) for CR1 and CR1L Genes

The PRT1-3 assays are designed to assign the copy number of the *CR1* gene. The DNA sequence of LCRs (copy number variable) and *CR1L* regions (non-copy number variable: checked by UCSC- DGV (Data base of Genomic Variants) (https://genome.ucsc.edu/)) of *CR1*-B allele was obtained from the UCSC genome browser (https://genome.ucsc.edu/). The regions from LCRs and *CR1L* were aligned by Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/). The primers were designed to take advantage of a 7-bp (2) and 23-bp deletion polymorphism (*CR1L*) producing PCR products of different lengths (Table 2.5). The primers that bind to those regions and the product sizes from these primers were checked by UCSC in silico PCR (https://genome.ucsc.edu/cgi-bin/hgPcr). The forward PRT primers; PRT1_F (RP1 purification), PRT2_F and PRT3_F (RP1 purification); are labelled on their 5' end by fluorescent dyes; HEX (hexachloro-fluorescein), NED (Applied Biosystems) and FAM (6-Carboxyfluorescein); respectively.

The PRT4-7 assays are designed to assign the copy number of the *CR1L* gene. The DNA sequence of LCRs (copy number variable) and *CR1L* regions (non-copy number variable: checked by UCSC- DGV (Data base of Genomic Variants) (https://genome.ucsc.edu/)) of the *CR1*-B allele was obtained from the UCSC genome browser (https://genome.ucsc.edu/). The regions from LCRs and *CR1L* were aligned using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/). The primers were designed to take advantage of 4-bp (CR1L_PRT4) and 6-bp (2) (CR1L_PRT6-7) deletion polymorphisms of *CR1* and a 5-bp (CR1L_PRT5) deletion polymorphism of *CR1L* producing PCR products of different lengths (Table 2.5). The primers that bind to those regions and the product sizes from these primers were checked by UCSC *in silico* PCR (https://genome.ucsc.edu/cgi-bin/hgPcr). The forward PRT primers; CR1L_PRT4 (RP1 purification), CR1L_PRT5 (RP1 purification), CR1L_PRT6 (RP1 purification) and CR1L_PRT7 (RP1 purification); are labelled on their 5' ends with fluorescent dyes, HEX or FAM.

These PRT primers were used in PCR (Table 2.6) and then the PCR products were run in ABI3130xl capillary electrophoresis.

Table 2.5: The three different pairs of primers designed for PRT assay. One of three different fluorescent dyes: HEX, NED and FAM) were covalently linked to the 5´ nucleotide of the forward primers of each primer set. The other primers CR1L_PRT4-7 were labelled using two different fluorescent dyes: HEX or FAM. This was to allow detection of PCR products during capillary electrophoresis. M=A/C base, Y=C/T, R=A/G and S=C/G.

Primer Name	**Sequence (5'→3')**	**Product Length (bp)**	**5' modification**	**Ordered From**
PRT1_F **PRT1_R**	GAGGAGACCCATAGTTCTTTACCA CATCACCTATCACACTGGTGC	Reference-108 Test-115	HEX (green)	Sigma-Aldrich
PRT2_F **PRT2_R**	GCTGTTCCAGGGTCAGAGTTA TTGGTCACATGATAGTCCTGC	Reference-164 Test-187	NED (yellow)	Applied Biosystems
PRT3_F **PRT3_R**	CTGTTTGAATAACTAGGTGGGAAGA TTCCCTCCAGATCTATCTAGATCTAGA	Reference-142 Test-149	FAM (blue)	Sigma-Aldrich
CR1L_PRT4F **CR1L_PRT4R**	GGAACAGCTAAGCCAGGTA RRGACTTTGGATTTATTTCYAATT	Test-159 Reference-155	HEX (green)	Sigma-Aldrich
CR1L_PRT5F **CR1L_PRT5R**	AASTAGAAAATGTGAGAMAGCAR AACCTMTTTTCTCAATTTYAGAGTC	Test-102 Reference-107	FAM (blue)	Sigma-Aldrich
CR1L_PRT6F **CR1L_PRT6R**	TTCTGAATCTGTAAGTATCATGTT TGGCYCAGTATAACATCTTT	Test-122 Reference-118	HEX (green)	Sigma-Aldrich
CR1L_PRT7F **CR1L_PRT7R**	AGTTTAGGTGTCAGCCTGGC ACTMATCTCCTGATCCAACAGC	Test-171 Reference-165	FAM (blue)	Sigma-Aldrich

Table 2.6: The volumes of components used to make PCR mix for PRT. Each PCR mix is prepared and run separately (Table 2.7).

PCR mix	PRT 1-3 and CR1L_PRT4-7 1X (µl)
10X low dNTP Buffer (see Appendix 3)	1
Forward Primer (10µM)	0.5
Reverse Primer (10µM)	0.5
Taq DNA Polymerase (5U/µl)	0.1
Water	6.9
DNA (10ng)	1

Table 2.7: PCR Cycles were optimized and run for each PRT and followed for each experiment as shown below.

Step	PRT1	PRT2	PRT3	PRT4	PRT5	PRT6	PRT7
Initial Denaturation	95°C for 2 minutes	95°C for 2 minutes	95°C for 2 minutes	95°C for 2 minutes	95°C for 2 minutes	95°C for 2 minutes	95°C for 2 minutes
Denaturation	95°C for 20 seconds	95°C for 20 seconds	95°C for 20 seconds	95°C for 20 seconds	95°C for 20 seconds	95°C for 20 seconds	95°C for 20 seconds
Annealing	63°C for 20 seconds	63°C for 20 seconds	61°C for 20 seconds	56°C for 30 seconds	57°C for 30 seconds	55°C for 20 seconds	60°C for 20 seconds
Extension	72°C for 20 seconds	72°C for 20 seconds	72°C for 20 seconds	72°C for 20 seconds	72°C for 20 seconds	72°C for 20 seconds	72°C for 20 seconds
Elongation	72°C for 10 minutes	72°C for 10 minutes	72°C for 10 minutes	72°C for 10 minutes	72°C for 10 minutes	72°C for 10 minutes	72°C for 10 minutes
Cycles	24	25	27	29	29	25	24
Final	+4°C	+4°C	+4°C	+4°C	+4°C	+4°C	+4°C

2.3.2 Capillary Electrophoresis

After PCR amplification of DNA, capillary electrophoresis was the second step of PRT in order to distinguish size or nucleotide differences to infer the copy number of the gene. The PCR amplicons were resolved using capillary electrophoresis, performed on an ABI 3130xl Genetic Analyzer (Applied Biosystems, UK).

10 µl of MapMarker© 1000 size standard (BioVentures, Inc., USA) was added to 1 ml of Hi-Di Formamide (Applied Biosystems, UK) and mixed gently. 9.4µl of the mixture was then aliquoted into each well of a 0.2-ml non-skirted 96-well plate (ABI Plate). 0.5-1.0 µl of each sample from each of the three PCR plates (run separately) was added into this single ABI plate. The plate was denatured for 3 minutes at 96°C then immediately placed on ice for a few minutes. The injection time for samples on the ABI3130xl capillary electrophoresis unit was fixed at 30 seconds. Then plate was placed on the ABI Genetic Analyzer and run. In order to analyse the data, run files were loaded to GeneMapper © software v.3.7 (Applied Biosystems, UK) and analysed using the 'CR1' bin setup (according to the predicted PCR product sizes of PRT1, PRT2 and PRT3). A table with all peak sizes and heights was then transferred to Excel where data were normalized by running a linear regression of the control samples (Table 2.4) ratios against expected integer copy number.

The resulting regression equation was used to adjust all other PRT data for each of the three dyes for PRT1, PRT2 and PRT3 assays. The example of the regression for PRT1 is shown below (Figure 2.10).

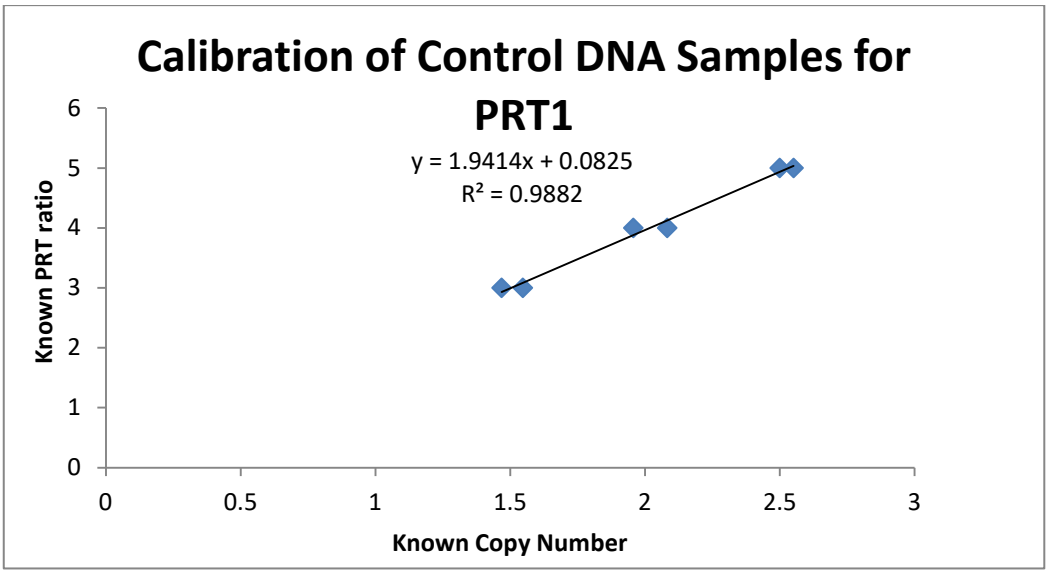


Figure 2.10: The calibration standard on reference DNA samples for PRT1 from a single PCR reaction.

2.3.3 Touchdown PCR for CR1 Breakpoint Analysis

According to Latorra et al. (2003), locked nucleic acid (LNA or LNA™) primers are better than to DNA primers for SNP detection. They found that LNA nucleosides at the 3' end of the primers significantly improve the discriminatory power of the primer. Adding LNA increases primer melting temperature, thereby improving the mismatch ability of the assay. In this research, LNA primers were designed to define the breakpoint for the duplication (*CR1*-B) allele of *CR1*. Breakpoints are locations on a chromosome where DNA might be deleted, translocated or inverted and these breakpoint regions are enriched for structural variations such as segmental duplications, copy number variants and single nucleotide polymorphisms. The LNA bases were positioned at the site of a variant in order to get the highest level of mismatch discrimination (Table 2.8). The forward primers were designed to bind LCR1' region and also to LCR2 region of *CR1* and the reverse primers were designed to bind the LCR1' region and also to LCR1 region of *CR1*. Therefore, PCR products only from LCR1' region (*CR1*-B allele) were expected.

The PCR products were amplified in a total volume of 10 µl with 1 µl of 10 x KAPA Taq Buffer A (Kappa Biosystems), 0.5 µl of 10 µM of each primer, 0.5 µl of 2.5 mM dNTPs (Promega), 0.1 µl of 5 U/µl Taq DNA polymerase (Kappa Biosystems) and 6.4 µl of dH_2O and 5-10ng of DNA as template in every PCR reaction as shown in Table 2.8 below. The

following thermocycler conditions were used for LNA PCR amplification: initial denaturation of 95°C for 2 minutes, followed by 20 cycles of 95°C for 30 seconds, 70°C for 30 seconds followed by incremental drops in temperature of 0.5°C every 30 seconds until 60°C is reached and 70°C for 30 seconds, followed by 15 cycles of 95°C for 30 seconds, 60°C for 30 seconds and 70°C for 30 seconds, followed by a final extension at 70°C for 5 minutes. The PCR products were separated using a 2% (w/v) Agarose gel (containing 0.5 µg/ml Ethidium bromide) in 1x TBE buffer. 4 µl of HyperLadder^{TM} I (BIOLINE) 1000bp was also run alongside the samples to check the size of the amplified PCR bands.

Table 2.8: Primer sequences used to amplify for duplication breakpoint of *CR1*-B and *CR1*-D alleles. Bold indicates the LNA base.

LNA Primer Name	**Primer Sequence (5'-3')**	**PCR product size (bp)**
LNA_ F1	CAAGACATCTTCCTTGCC**C**	**F1 and R1**: 287 **F2 and R1**: 564 **F3 and R1**: 867 **F1 and R3**: 856 **F2 and R3**: 1133 **F3 and R3**: 1436
LNA_F2	GATGGAGGGACCTATTAGAT**G**	
LNA_F3	AATGTGTTTTGATTTCCCAAGATC**A**	
LNA_ R1	CAGCCCCTCTGAGGT**C**	
LNA_R3	CTCAACCTCCCAAAGGTGC**T**	

2.4 SNP genotyping by Allele-Specific Oligonucleotide (ASO) Hybridization

In order to obtain genotyping data for Knops blood group-determining SNPs, an ASO assay was used. The SNPs typed with this assay were rs17047661, rs17047660, rs41274768 and rs6691117. These SNPs are all within a single exon (exon 29 of *CR1*-A allele, exon 39 of *CR1*-B allele) of the *CR1* gene and therefore 420 bp length (Appendix 4) DNA sequence (*CR1*-B allele, exon 29 and the position: 207,782,654-207,783,073) which captures the region containing these SNPs was selected for primer design. The

DNA sequence of the *CR1*-B allele of *CR1* was obtained from UCSC genome browser (Human Feb. 2009 (GRCh37/hg19) Assembly) (https://genome.ucsc.edu/). The primers were designed using Primer3 (http://primer3.sourceforge.net/). The specific parameters which were considered for the primer design were gc% (max 50%) and length (max 22 bp).

The PCR primers were designed and the PCR cycles with optimized temperatures were applied as shown in Table 2.9 and Table 2.10 below, respectively.

Table 2.9: The primers designed for the ASO assay. The PCR products at the end of PCR reaction were shown in below.

Primer Name	**Sequence (5'→3')**	**Product Length (bp)**
ASO_F ASO_R	TCCAATGGAGACTTCTACAGCA CACACCCAGCAAAGTCTTGA	404

Table 2.10: PCR Cycles (were optimized and run) for ASO primers.

Step	**ASO primers**	**Cycles**
Initial Denaturation	95°C for 2 minutes	1
Denaturation	95°C for 30 seconds	35
Annealing	55°C for 30 seconds	
Extension	70°C for 30 seconds	
Elongation	70°C for 5 minutes	1
Final	+4°C	

PCR was performed in a total volume of 20 µl reactions on a Veriti thermal cycler (ABI) with 2 µl of 10 x KAPA Taq Buffer A (Kappa Biosystems, 15mM Mg^{2+}), 0.4 µl of 10 µM of each primer (forward and reverse primer), 0.4 µl of 2.5mM dNTPs (Promega), 0.1 µl of 5 U/µl Taq DNA polymerase (Kappa Biosystems) and 15.7 µl of dH_2O and 1 µl of 5-10ng of DNA as template in every PCR.

Probes were designed according to corresponding SNPs that determine the Knops blood group antigen in *CR1*. They were labelled using γ^{32}P-ATP using oligo-labelling solution (see Appendix 5) before each experiment (Table 2.11). The labelling reaction was stopped using the kinase stop solution (see Appendix 6).

Table 2.11: The probes used for oligo-labelling. Each smaller letter base indicates the variable nucleotide, positioned 8-11 nucleotides from the 5′ end.

Probe Name	SNPs (*CR1*)	Sequence
Probe 1G	rs41274768	TTGAGCTTgTGGGAGAAC
Probe 1A		TTGAGCTTaTGGGAGAAC
Probe 2G	rs17047660	CTACTAATgAATGCACAG
Probe 2A		CTACTAATaAATGCACAG
Probe 3G	rs17047661	ATGCAATTgGAGTACCAG
Probe 3A		ATGCAATTaGAGTACCAG
Probe 4G	rs6691117	CTGAGATCgTCAGATTTA
Probe 4A		CTGAGATCaTCAGATTTA

PCR products were denatured using denaturing mix (included Bromophenol blue) (see Appendix 7). The denatured PCR products were transferred onto nylon membrane as dots for dot blotting. DNA was fixed to the membranes using the UV crosslinker (HoeferTM UVC 500 Ultraviolet Crosslinker was used with 70,000 μJ/cm^2 for 30 seconds). After dot blotting the membranes were placed into numbered bottles in a rotisserie oven with TMAC hybridization solution (see Appendix 8) at 53°C for 10 minutes (pre-hybridization). ^{32}P-labelled probes (incubated at 60°C for 3 minutes) were added into the bottles and incubated at 50°C for 4 hours in the rotisserie oven. The membrane was washed with TMAC wash solution at 50°C (see Appendix 9). After that, the blots were placed into a cassette with X-ray film and exposed at -80° for 6-7 hours (for weaker dots even overnight).

2.5 Sequencing of PCR Products

Each PCR was prepared to give a final volume of 20 µl, of which 1 µl was run on gel electrophoresis with 9 µl of dH_2O (including 1 µl of 5X DNA loading buffer blue, BIOLINE) to confirm the presence of PCR products of expected length.

For gel electrophoresis, a 2% (w/v) Agarose gel (containing 0.5 µg/ml Ethidium bromide) in 0.5x TBE buffer was used. 5 µl of HyperLadder™ IV (BIOLINE) 100bp was also run alongside the samples to check the size of the amplified PCR bands. When a single band was present, the remaining 24 µl of PCR product were again loaded on another 2% (w/v) Agarose gel (containing 0.5 µg/ml Ethidium bromide) in 0.5x TBE buffer.

After gel electrophoresis, the bands of interest were cut from the gel, under a blue light reader. The removed gel can be stored at $+4^0$C. Then, the removed gel containing the PCR product of interest was loaded into another 2% (w/v) Agarose gel (containing 0.5 µg/ml Ethidium bromide) in 1x TAE buffer with the centre of the gel removed and filled with a 0.2 % (w/v) Agarose gel (containing 0.5 µg/ml Ethidium bromide) in 1x TAE buffer. The agarose gel was then run at 100V until the band reached the centred 0.2 % Agarose gel and the band was removed with a 200-µl pipette (Thermo Scientific™). The removed gel (containing 0.2 % (w/v) Agarose gel and DNA) was kept at -20^0C overnight in a 1.5-ml eppendorf tube (Sigma-Aldrich®) (Difazio, 2008).

The removed gel (frozen) was defrosted at room temperature for 10 minutes and then centrifuged for 30 min at 1500g. The DNA was extracted from the solution with a pipette, and its concentration was checked by NanoDrop spectrophotometer (NanoDrop™ 1000 Spectrophotometer, Thermo Scientific); NanoDrop can measure the total absorbance of 1 µl of DNA sample on the assumption that the maximum absorbance for DNA is at 260 nm.

Following these steps, the sequencing reactions were set as follows, in a total volume of 20 µl: 1 µl of Big Dye Terminator Ready Reaction mix (Applied Biosystems and stored in -20^0C) containing four standard deoxynucleotides (dNTPs) and the four dideoxynucleotides (ddNTPs) each labelled with a different fluorescent dye; 3.5 µl of 5X Big Dye Terminator Buffer (50mM Tris-HCl (pH 9.0)) and 2mM $MgCl_2$ and stored at

+4^{0}C); 3.2 µM of left or right primer and 3-10 ng of PCR product template. The sequencing reactions were then run using an Applied Biosystems Veriti 96-well thermal cyclers, firstly 96°C for 1 min followed by 26 cycles protocol: 96 °C for 10 seconds followed by 50 °C for 5 seconds followed by 60 °C for 10 seconds, followed by 60 °C for 4 minutes.

Sequencing reactions could be stored at this point of the protocol at 4°C prior to processing. The sequencing reactions had to undergo a further cleaning step, to remove all the unincorporated labelled nucleotides, as follows: 2 µl of 2.2% (w/v) SDS were added to the mixture and carefully mixed to bring the final concentration to 0.2%. Reactions were briefly spun and then denatured using a thermal cycler at 98°C for 5 minutes followed by 10 minutes incubation at 25°C. The excess dye terminator was then removed by passing the reaction through a Performa DTR gel filtration column (Edge Biosystems CAT N° 42453) followed by a 3-minutes spin at 3200 rpm. Eluates were then submitted to the University of Leicester Protein Nucleic Acid Chemistry Laboratory (PNACL) (http://www2.le.ac.uk/colleges/medbiopsych/facilities-and-services/cbs/protein-and-dna-facility/pnacl) and run through an Applied Biosystems 3730 sequencer. The sequencing data were then analysed using the programme MEGA6.0 (Molecular Evolutionary Genetics Analysis) (Tamura et al., 2013) (http://www.megasoftware.net/).

2.6 Statistical Association Analysis

2.6.1 Analysis of CR1 variants and Parasite Density

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

2.6.2 Analysis of CR1 variants and number of malarial infections

A Poisson regression was constructed using SPSS 20.0 (IBM) to analyse the association with the number of breakthrough infections and the covariates (mother's age, malaria suspicion during pregnancy, sex, birth term, mosquito net, ethnic group, sickle cell,

chloroquine intake during pregnancy and Knops blood group-determining SNPs) in the Tori-Bossito cohort, whereas the covariates were mother infection, sex and SNPs for the Tolimmunpal cohort.

2.6.3 The effect of CR1 Variants on Time to First Malarial Infection

Cox regression is also known as proportional hazards regression, a method for investigating the effect of several variables upon the time a specified event takes to happen. This regression analysis is used to assess the effect of the factors (mother's age, malaria suspicion during pregnancy, sex, birth term, mosquito net, ethnic group, sickle cell, chloroquine intake during pregnancy and Knops blood group-determining SNPs on risk of malaria in the Tori-Bossito cohort, whereas the factors were maternal infection, sex and SNPs in the Tolimmunpal cohort. In order to do this analysis, SPSS 20.0 (IBM) was used.

2.6.4 Malaria Prevalence Data Analysis

The SNP data were collected from 36 different populations across sub-Saharan Africa. For HGDP African samples the ASO assay was used to obtain SNP genotypes, whereas the rest of the SNP data were collected from publically available sources such as 1000 Genomes (more details about the populations are provided in Chapter 2). Malaria prevalence is the proportion of people infected by malaria at a given point in time. Malaria prevalence was estimated by Dr Rita Rasteiro in the year 2000 from Malaria Atlas Project data (http://www.map.ox.ac.uk/map/) for each sampling point, taking an average of a 110-km radius circle around that sampling point. The sPAMM package in R was used for basic linear regression analysis (Rousset and Ferdy 2014) to investigate an association between Sl2 allele frequency and malaria prevalence. sPAMM analysis was done by Dr Edward Hollox.

2.6.5 Linear Regression Analysis (EOAD and LOAD) and Clustering Quality (Q)

For case and control studies (early-onset and late-onset Alzheimer's disease cohorts) linear regression analysis was performed. Logistic regression analysis was performed with SPSS 20.0 (IBM) to assess the probable correlation between *CR1* (LCR) copy number and LOAD or EOAD.

In order to measure the clustering quality (Q), the method of Barnes et al. (2008) was used. They defined Q as a simple averaged value of the signal-to-noise ratios between adjacent copy numbers. The Q calculations were conducted in Plink (http://pngu.mgh.harvard.edu/~purcell/plink/). The command that was used to obtain the Q value is shown in Appendix 10.

2.6.6 Linkage Disequilibrium Analysis

In order to perform the linkage disequilibrium analysis between SNP-SNP or copy number (*CR1*) Plink (http://pngu.mgh.harvard.edu/~purcell/plink/) was used (the commands which were used for the analysis are shown in Appendix 11).

2.6.7 The Power of the LOAD study

The power calculation was done by Dr Edward Hollox (the script is shown in Appendix 12). The power of this study was calculated by using the effect size estimated by the previous study (Brouwers et al., 2012). The script was run ten times using R and the maximum and minimum results were reported.

3 CHARACTERIZATION OF COPY NUMBER VARIATION IN THE HUMAN *CR1* GENE

3.1 Design of Paralogue Ratio Tests Assays for *CR1* CNV

In previous research to study the extent and nature of the copy number variation (CNV) within the *CR1* gene, the four CNV alleles of *CR1* and their frequencies were studied by restriction fragment length polymorphism (RFLP) and Southern blotting (Hollers et al., 1987; Vik and Wong, 1993; Wong et el, 1986; Wong et el., 1991). From these studies, it is known that *CR1* alleles are determined by the addition or deletion of the LCR1 region of *CR1*. Learning the extent and the nature of this CNV, and its relationship to flanking SNP alleles, might provide information to understand more about CNVs (*CR1*) role in diseases such as Alzheimer's disease (AD) and malaria.

Previously, Brouwers et al. (2012) designed an assay to study the CNV of *CR1* (LCR) in late-onset Alzheimer's disease cohorts. In their study, they used the MAQ (Multiplex amplicon quantification) technique with three fluorescently-labelled test amplicons (targeting LCRs) and seven control amplicons (outside the *CR1* locus). Details about measurement of CNV using their approach were not made clear in the paper, nor was the accuracy reliably established. For example, the exact positions for the amplicon targeted regions on LCRs and outside of the *CR1* locus for CNV measurements were not given. Therefore, the method used by Brouwers et al. (2012) or Southern blotting assays would not be suitable for large samples to type AD and malaria cohorts. In order to design a new approach to type the CNV in the *CR1* gene, three different paralogue ratio test (PRT) assays (Armour et al., 2006) were designed.

The first exon of the *CR1L* gene is 3.37 kb proximal to the last exon of *CR1*. The *CR1L* gene shows 92% similarity to *CR1* (UCSC genome browser) and it was also found that a 40-kb genomic region of *CR1L* which contains 10 potential exons had 95% similarity with the amino-terminal coding part of *CR1* (Hourcade et al., 1989). The CR1L protein is expressed in non-human primates as the immune adherence receptor but knowledge about human *CR1L* gene expression is very limited. Expression of the CR1L transcript is limited to haematopoietic and foetal lymphoid tissue in humans (Logar et al., 2004). In addition, it has been suggested that maintenance of CR1-like expression

in non-human primates has been evolutionarily favoured (Birmingham and Hebert., 2001), perhaps in response to malaria.

The three independent PRT assays were intended to gain three separate measurements of LCR-related CNV within the *CR1* gene. It is expected that each PRT assay would give similar copy number (CN). For each PRT assay, the test region was within LCR1, LCR1' and LCR2 of *CR1*, and the reference region was within *CR1L*. From examining the DGV (Database of Genomic Variants) (MacDonald et al., 2014) track within the UCSC Genome Browser (UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) assembly), the *CR1L* is not copy number variable. The Database of Genomic Variants is the most recent repository of CNVs and includes insertions, deletions, inversions and inversion breakpoints in the human genome. The data are obtained from healthy individuals. DGV considers segments of DNA that are larger than 1000 bp, and insertions or deletions of larger that 50 bp are also included.

For the first PRT assay (PRT1), PRT1 test and PRT1 reference region were amplified using the same set of PRT1 primers. It yields a 108-bp product from the *CR1L* gene and three 115-bp products from the *CR1* gene for CR1-B allele. The positions at which the PRT1 primers bind in the *CR1* and *CR1L* genes on the *CR1*-B allele are shown below (Figure 3.11).

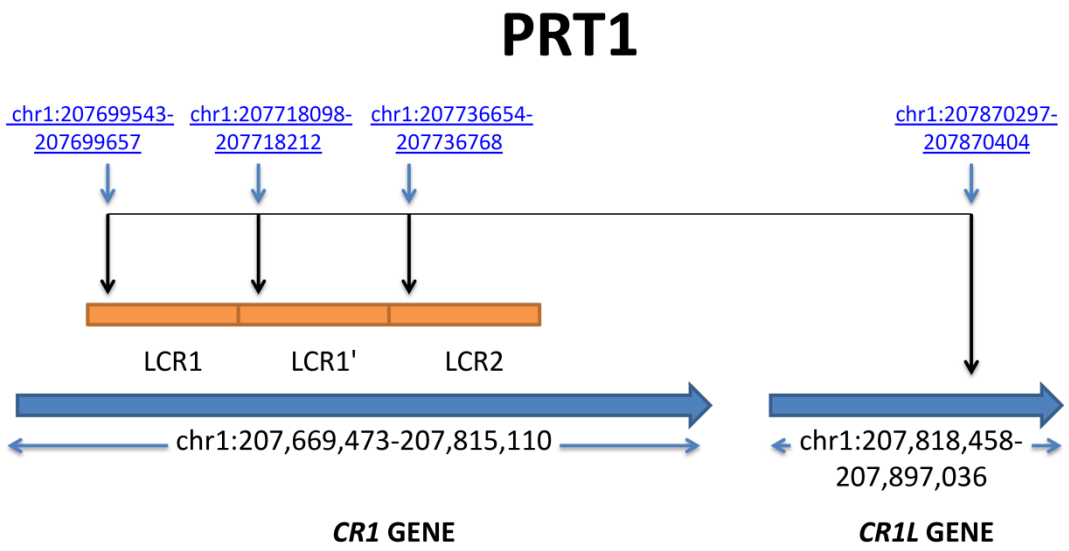


Figure 3.11: The positions of the PRT1 primer set in the *CR1* and *CR1L* genes for *CR1*-B allele. The primers bind to *CR1* gene as the test (LCR1, LCR1' and LCR2) and *CR1L* as the reference

region. The reference sequence for the *CR1*-B allele was collected from UCSC Genome Browser.

Different primer sets were designed and used for the second PRT assay (PRT2). This assay gives a 164-bp product from the *CR1L* gene (reference) and three 187-bp products from the *CR1* gene (test). The positions at which the PRT2 primers bind on the *CR1* and *CR1L* genes in the *CR1*-B allele are shown below (Figure 3.12).

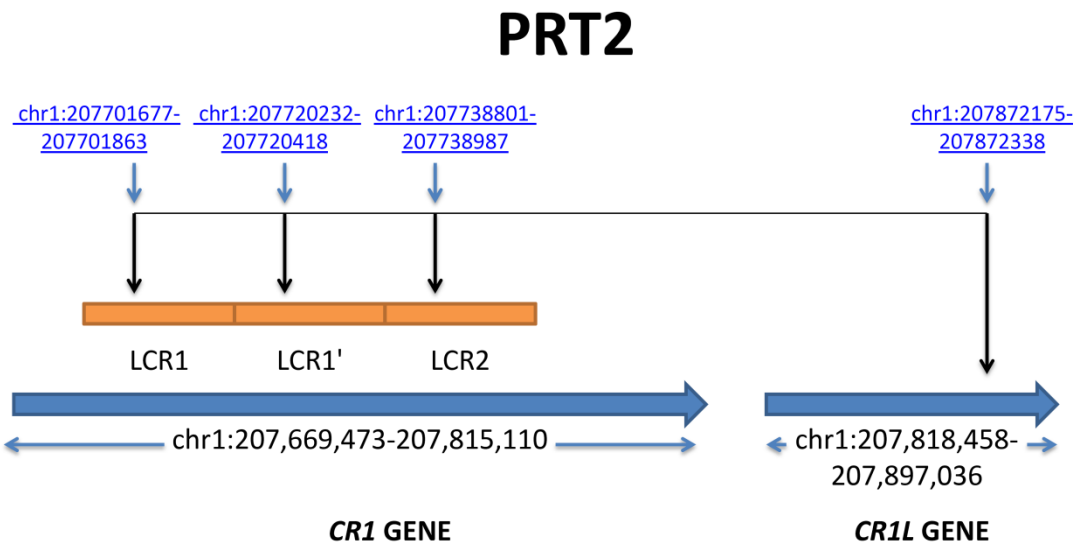


Figure 3.12: The positions of the PRT2 primer set in the *CR1* and *CR1L* genes for *CR1*-B allele. The primers bind to *CR1* as the test (LCR1, LCR1' and LCR2) and *CR1L* as the reference region. The reference sequence for the *CR1*-B allele was collected from UCSC Genome Browser.

For the third PRT assay (PRT3), a 142-bp product from the *CR1L* gene and three 149-bp products from the *CR1* gene were obtained. The positions at which the PRT3 primers bind on the *CR1* and *CR1L* genes on the *CR1*-B allele are shown below (Figure 3.13).

The forward primer of the each of the three primer sets was tagged by one of three different dyes which are HEX, NED and FAM for PRT1, PRT2 and PRT3 forward primers, respectively.

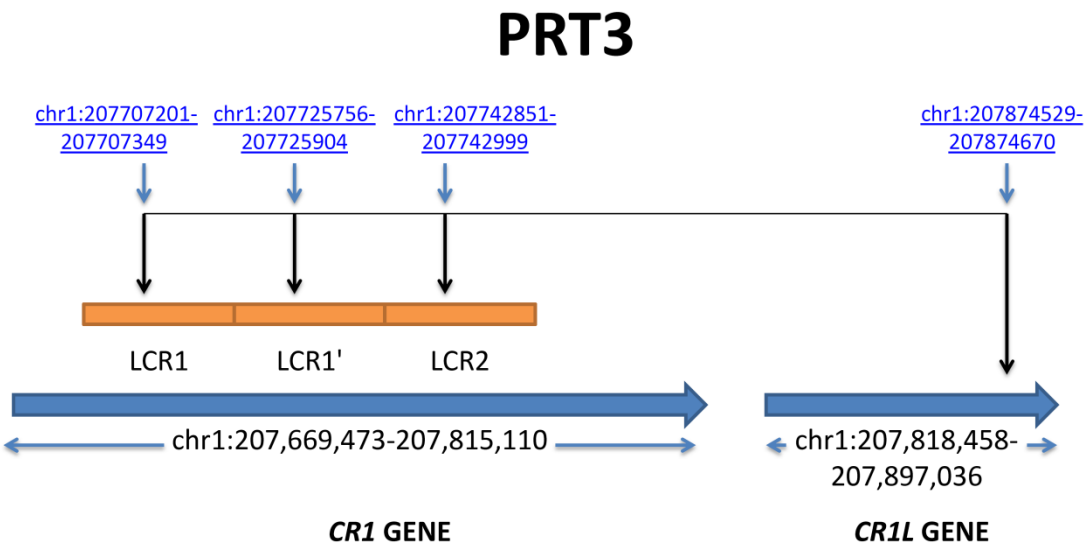


Figure 3.13: The positions of the PRT3 primer set in the *CR1* and *CR1L* genes for *CR1*-B allele. The primers bind to *CR1* as the test (LCR1, LCR1' and LCR2) and *CR1L* as the reference region. The reference sequence for the *CR1*-B allele was collected from UCSC Genome Browser.

After the PCR step, the data were transferred to GeneMapper software for copy number calling (Figure 3.14).

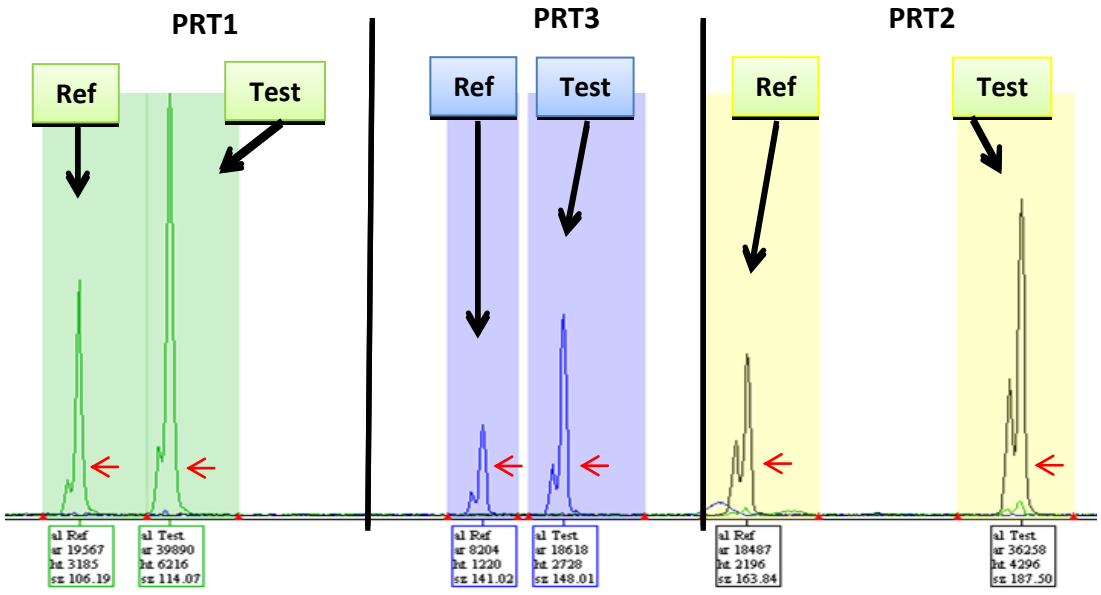


Figure 3.14: Data obtained following capillary electrophoresis by using GeneMapper software. These data are used to decide the LCR-related copy number of *CR1* for different samples. The copy number of NA18555: Average (Test/Reference): PRT1= 2.03; PRT2= 1.96 and PRT3= 2.26 that LCR CN is 2 for the *CR1* gene (before optimization).

The raw paralogue ratios (test/reference) were calculated from each dye for each PRT and corrected for PCR plate batch effects. The internal controls included on each PCR

run of these assays were used as 'gold standards' of known *CR1* (LCR) copy number against which to calibrate the rest of the plate (Table 3.12).

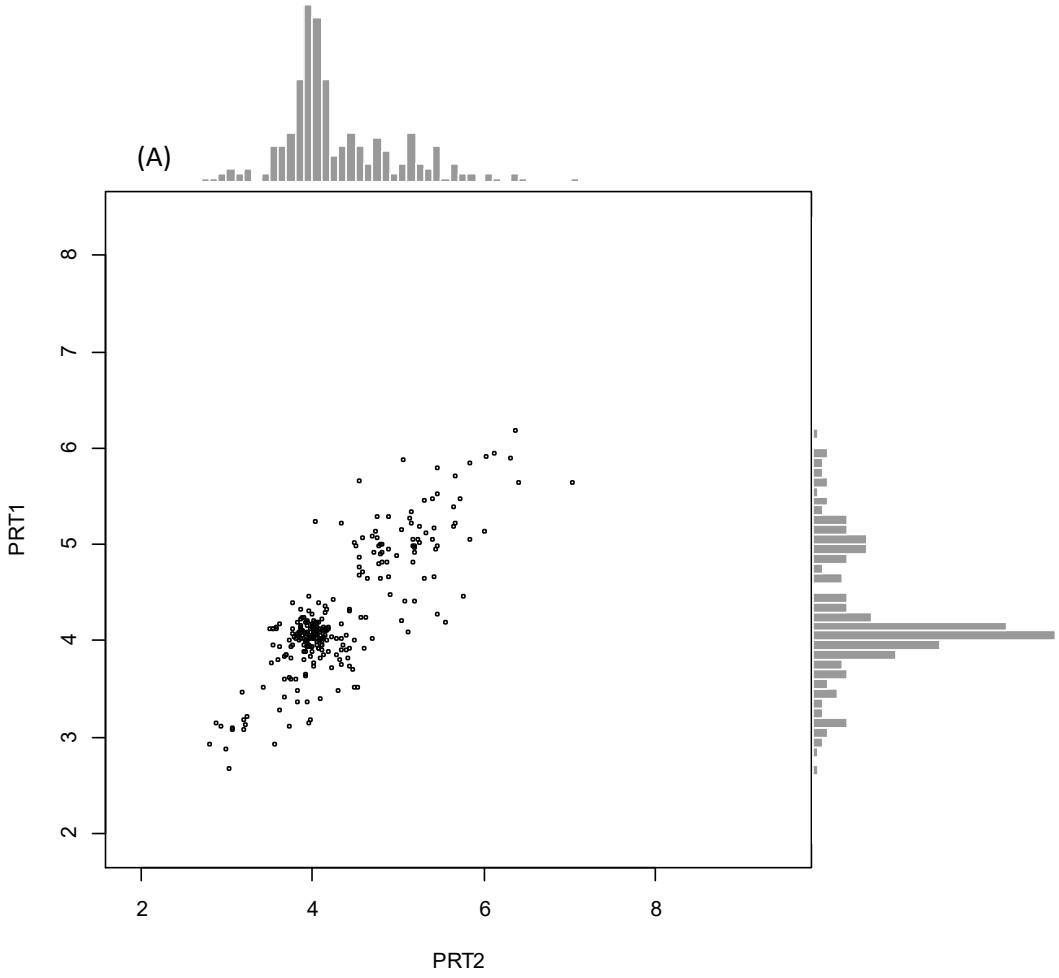
Table 3.12: The internal controls used as gold standards of known *CR1* copy number (LCR) in three independent PRT assays. The numbers of the low-copy repeats (LCRs) *CR1* which were used as reference copy numbers were taken from array-CGH data (Conrad et al., 2010). The LCR *CR1* copy number of C0140 and C0182 were found after the first batch of tests using PRT1-3.

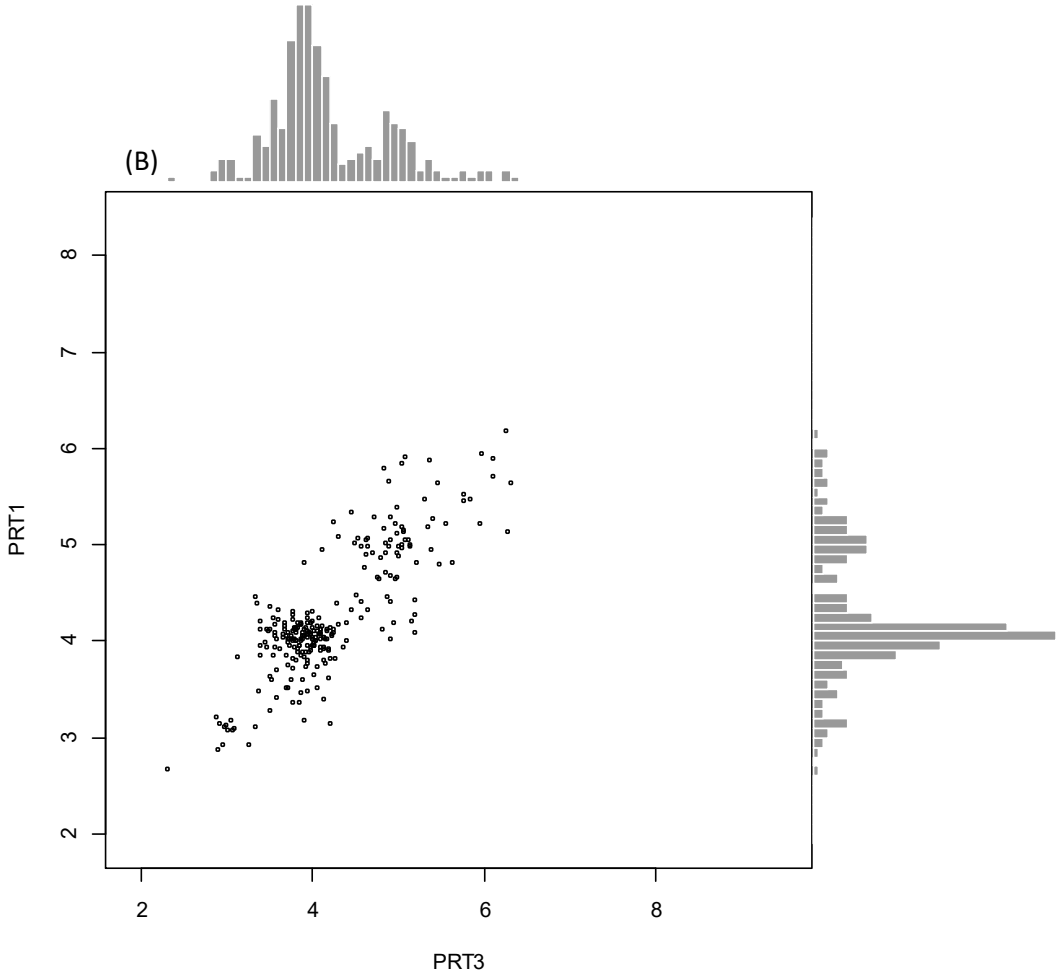
Reference Sample ID	**LCR related Copy Number in *CR1***
NA18507 (International HapMap Project-YRI)	**3**
NA18517 (International HapMap Project-YRI)	**4**
NA18555 (International HapMap Project-CHB)	**4**
HRC1-C0182 (Human Random Control 1)	**5**
HRC1-C0140 (Human Random Control 1)	**5**

3.1.1 Comparison of the three PRT assays

In order to understand the quality of the data and how each assay works with the data, PRT assays were compared with each other as raw copy numbers to see distinct clusters which give the copy number of LCR (*CR1*). It was also important to show that the PRT assays (PRT1-3) are giving the same copy number results, in which case to decide the copy number of LCR *CR1* their average can be used in future CNV studies.

HapMap Phase 1 samples were used to compare three PRT assays. The HapMap phase 1 samples, which were 266 individuals from four geographically diverse populations (CEU, YRI, CHB and JPT) were typed using the three PRT assays (PRT1-3) (Figure 3.15).


(A)
PRT1
PRT2
2
3
4
5
6
7
8
2
4
6
8


(B)
PRT1
PRT3
2
3
4
5
6
7
8
2
4
6
8

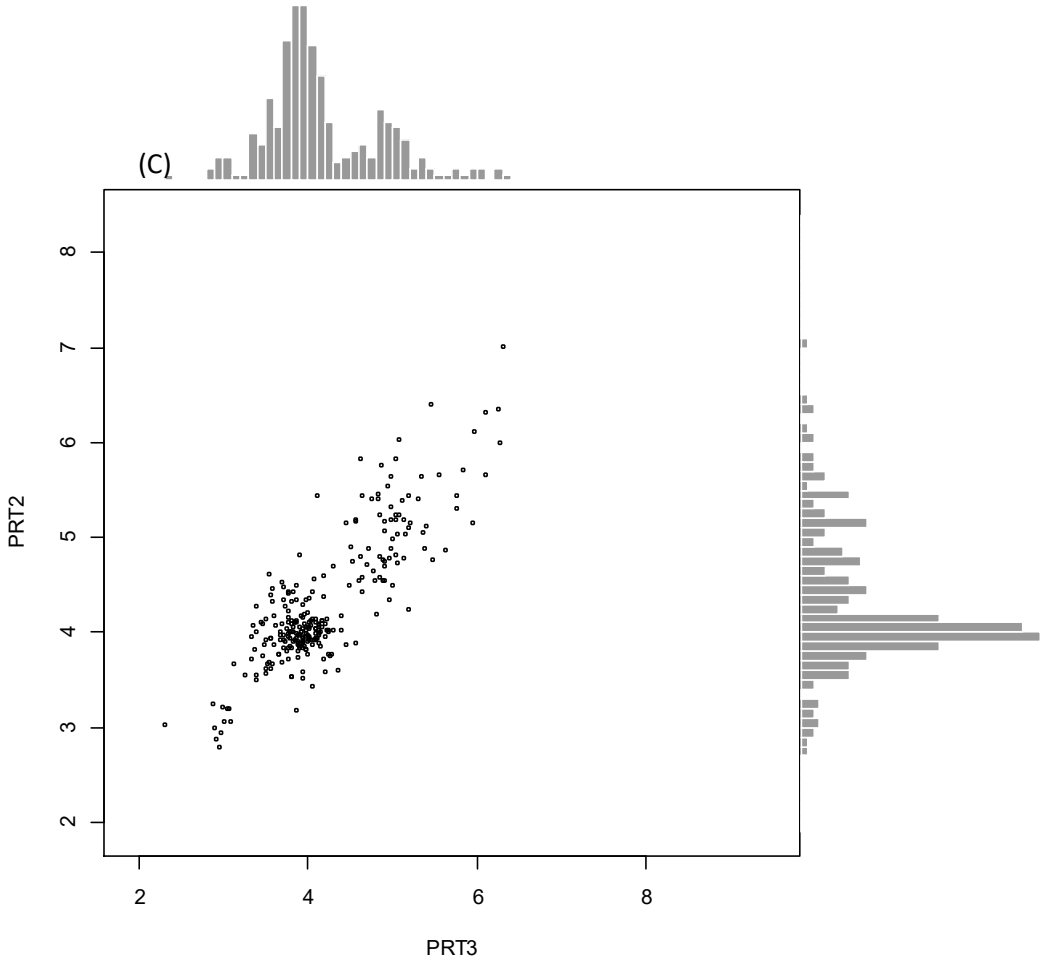


Figure 3.15: The comparison of three PRT assays in HapMap phase 1 (267 individuals from four geographically diverse populations (CEU, YRI, CHB and JPT)). **(A)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT2 data in HapMap phase 1. **(B)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT3 data in HapMap phase 1. **(C)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT2 data and raw data normalized to known copy number of PRT3 data in HapMap phase 1.

The comparision of each assay provides a better understanding about each assay and which assay is better to use for copy number calling individually or together for a unique cohort. It also shows they can be combined meaning that their avarage value can be taken to decide the copy number of LCR CR1 for any cohort.

3.2 Validation of PRT Assays Using arrayCGH Data

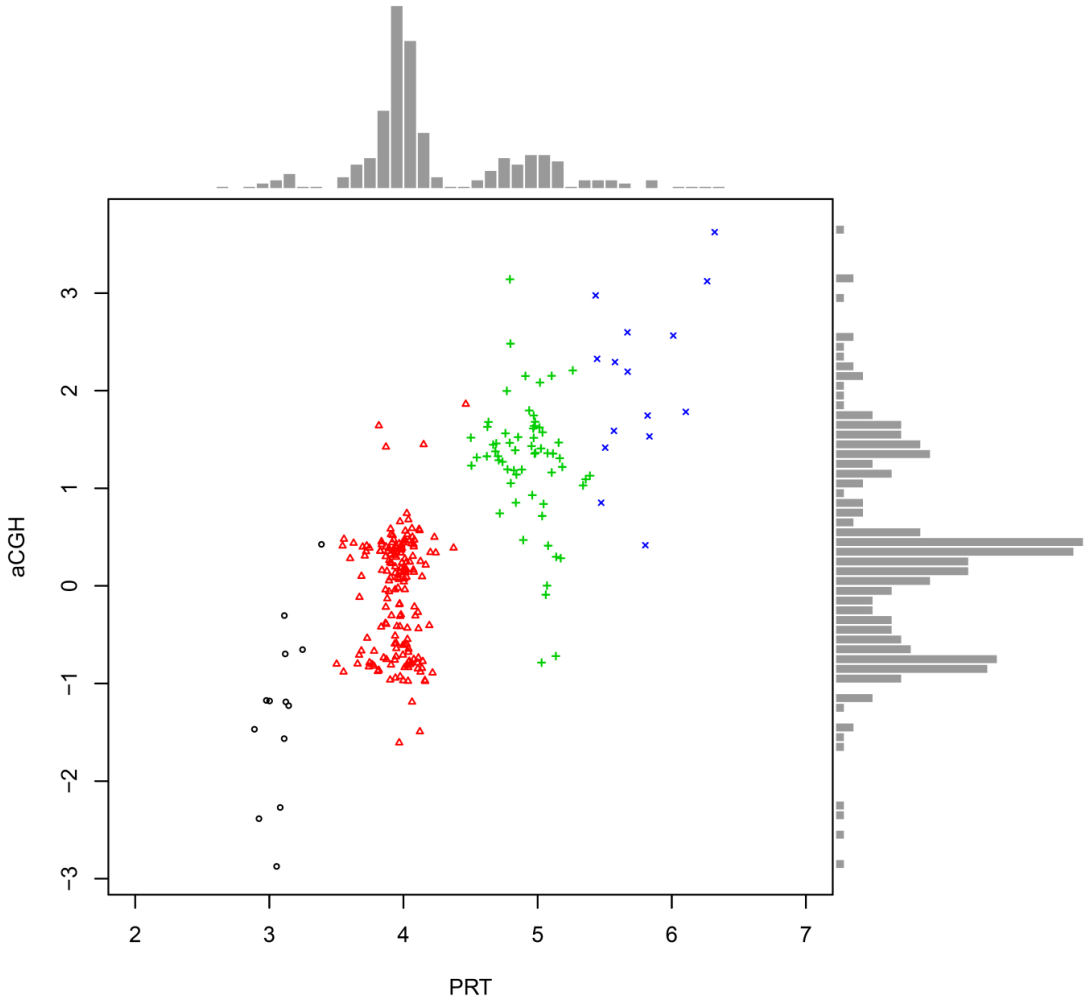


Figure 3.16: Assessment of *CR1* LCR copy number assays (average of 3 PRTs raw ratio of *CR1* LCR) quality in HapMap samples. The scatterplot and associated histogram showing raw PRT values generated by average raw PRT data (PRT1, PRT2 and PRT3) (x-axis) plotted against arrayCGH data (Conrad Agilent aCGH data) (y-axis) for 267 of 269 HapMap Phase 1 DNA samples. According to PRT assays, the black-coloured dots represent the diploid *CR1* LCR copy number of three whereas the red-coloured dots represent the diploid *CR1* LCR1 copy number of four (modal copy number) and the green color dots represent the diploid *CR1* LCR copy number of five whereas the red-coloured dots represent the diploid *CR1* LCR1 copy number of six.

Validation of three PRT assays with an accurate copy number measurement analysis was an essential requirement of this research. ArrayCGH (aCGH) data for *CR1* LCR was provided by Dr Donald F. Conrad (Conrad et al., 2009), and the copy number value of PRT (average of PRT1, PRT2 and PRT3) compared with aCGH data (Figure 3.16). These were analyzed for 267 individuals from the HapMap project (four geographically

diverse populations, CEU, YRI, CHB and JPT (2.1.3 HGDP-CEPH Panel). The aCGH data were compared with mean unrounded PRT ratios for each sample to examine the copy number calling for *CR1* LCR using PRT assays (PRT1, PRT2 and PRT3). For this purpose, aCGH data were compared with average PRT raw ratio of *CR1* LCR. A scatter plot (Figure 3.16) was plotted to determine the reliability of PRT assays and also to finalise which approach may be best for validation of copy number calling of *CR1* LCR. A robust PRT assay should display distinct clusters of copy number (LCR) which should be agree with the allele frequency of CN variants of the *CR1* gene.There is a positive relationship between copy number called using aCGH and that called using PRT, showing that they are both measuring the same CNV. However, PRT was a better assay for copy number calling of *CR1* LCR because it shows clearer clustering.

3.3 Validation of PRT Assays Using Sequence Read Depth in the 1000 Genomes Project Data

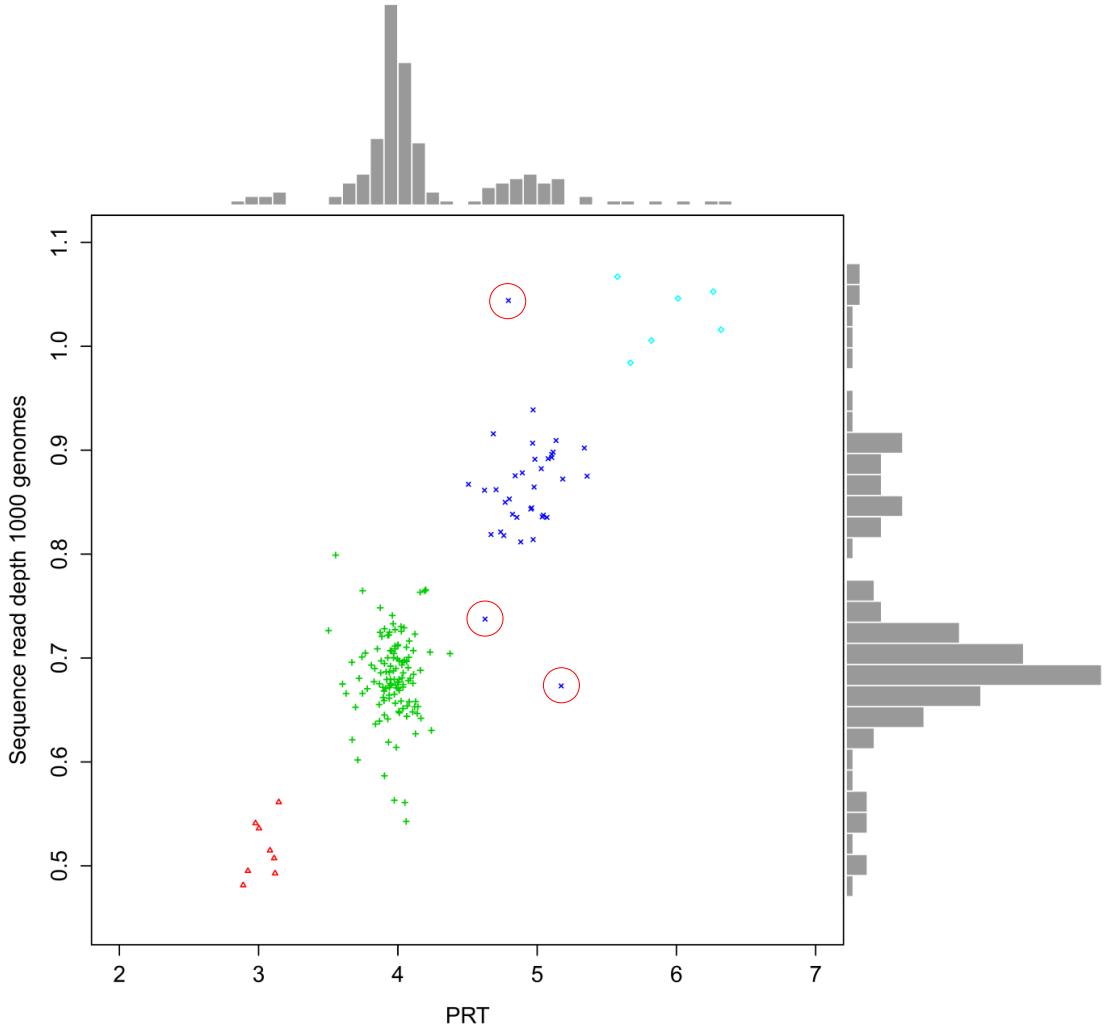


Figure 3.17: Assessment of *CR1* LCR copy number assays (average of 3 PRTs raw ratio of *CR1* LCR) quality in HapMap samples. The scatterplot and associated histogram showing raw PRT value generated by average raw PRT data (PRT1, PRT2 and PRT3) (x-axis) plotted against raw sequencing read depth in the 1000 Genomes Project data (y-axis) for the 174 HapMap Phase 1 DNA samples. The samples were collected from unrelated indivuduals so not all 269 HapMap Phase 1 DNA samples (includes trios) were used. The red dots indicate heterozygous samples for *CR1*-C and *CR1*-A alleles whereas the green dots shows the homozygous samples for *CR1*-A allele. The blue dots indicate heterozygous samples for *CR1*-A and *CR1*-B alleles whereas the light blue dots showhomozygous samples for *CR1*-B allele. The red circles highlight the three samples which are showing unmatched results between PRT and raw sequencing read depth data. These three samples were retyped in the following experiments.

Sequence read depth copy number data for *CR1* LCR was provided by Dr Edward J Hollox. For each sample, the next-generation sequencing data were downloaded from the 1000 Genomes Project. In order to get copy number for each sample, the number

of sequence reads that map to a region (chr1:207,697,239-207,751,921) including CNV (LCR1 and LCR1') was divided by the number of sequence reads that map to a similar-sized region (chr1:207,953,949-208,008,574) but outside the CNV and outside the gene in an area that does not show CNV according to DGV. Sequence read depth was compared with average PRT raw ratio of *CR1* LCR. A scatter plot (Figure 3.17) was constructed to determine the quality of PRT assays and also to finalize which approach might be best for validation of copy number calling of *CR1* LCR. The scatter plot showed that the both PRT and sequencing methods were useful for copy number calling of *CR1* LCR as they both produced similar results showing clustering for each copy number, both sets of copy number data matched with the frequencies of *CR1* alleles in worldwide populations.

Three of the samples did not match between PRT and the sequencing read depth data (highlighted by the red circles, Figure 3.17), and those samples were retyped with three PRT assays (Table 3.17). For the new analysis, the original DNA stocks were used instead of the working DNA stocks. Two of the new results (NA07056 and NA18870) agree with the sequencing results. However, one sample (NA18853) still showed variation from the sequencing result.

Table 3.13: Three HapMap1 samples which showed discrepancies during validation of PRT assays were retyped by PRT1, PRT2 and PRT3 and compared to the sequencing and aCGH data.

Sample	PRT1	PRT2	PRT3	AVERAGE	SEQUENCING	aCGH
NA07056	6.24	6.48	6.24	6.32	1.044	3.141
NA18853	5.16	4.56	4.14	4.62	0.737	1.630
NA18870	4.04	3.41	3.75	3.74	0.673	0.284

3.4 Worldwide Distribution of LCR *CR1* Gene Copy Number

The variation of LCR CN of *CR1* was analysed between different populations by using Paralogue Ratio Test (PRT) to investigate the nature and extent of LCR CNV globally. In order to see the LCR CN distribution in UK samples 40 unrelated samples (HRC1) were analysed by PRT.

Initial analysis showed that 4 (n= 23; 57.5%) was the most common copy number whereas the second commonest copy number was 5 (n= 15; 37.5%). In this study, the copy number of 3 (n= 2; 5.3%) was observed less commonly whereas the copy number of 6 (n=0; 0%) was not observed in 40 samples of HRC1 (Figure 3.18).

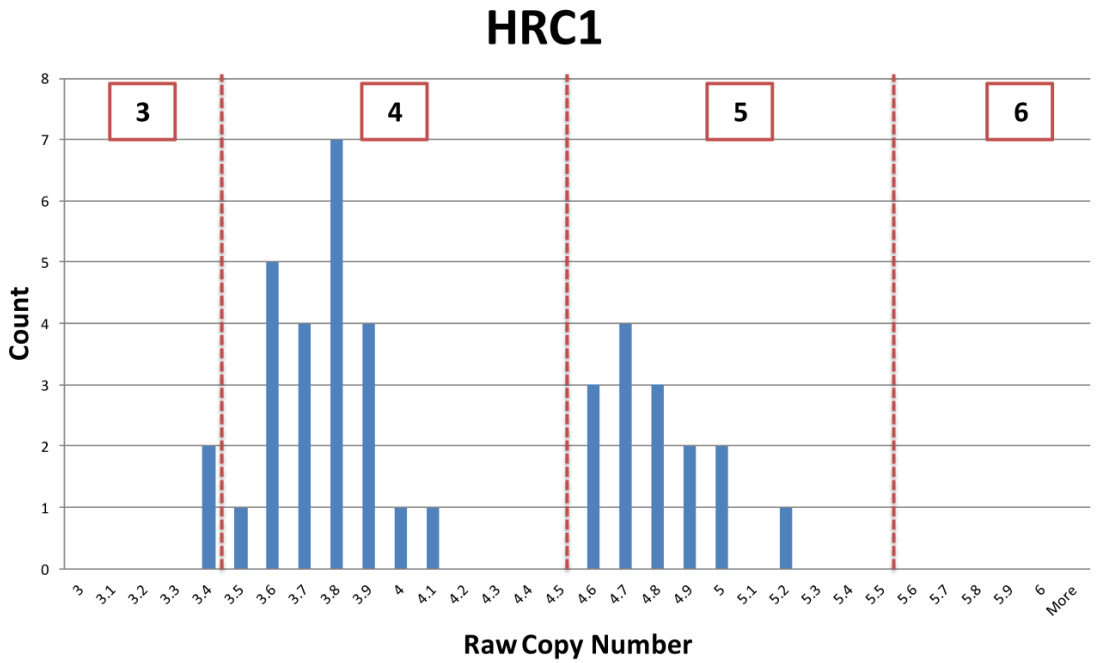


Figure 3.18: Histogram of the raw copy number estimates. The raw copy number data are obtained by taking the average of three normalized raw PRTs (PRT1-3). The raw copy number estimates are in bins of 0.1, with the count of each bin displayed on the y-axis. The numbers at the top indicate the integer diploid copy number of *CR1* LCR.

Other worldwide populations were analysed to determine the distribution of LCR *CR1* CN (Table 3.14) between distinct populations. For this purpose, a variety of populations has been used to assess the distribution of LCR1 *CR1* CN among worldwide populations such as HapMap phase 1 (YRI, JPT, CHB and CEU) and Human Genome Diversity Cell Line Panel (HGDP-CEPH) (seven geographical regions; Africa, Europe, the Middle East, central/South Asia, East Asia, Oceania and America). These samples were typed for LCR *CR1* copy number using the PRT method.

The distribution of diploid LCR *CR1* copy number was calculated for each region. The distribution of integer copy of LCR *CR1* is shown in the Gaussian distribution below, based on the HGDP-CEPH panel in the form of subset H971 (contains no two

individuals with a first-degree relationship (parent/offspring or full siblings)) and the HapMap sample (Figure 3.19). In order to obtain more distinct distribution of LCR *CR1* with higher population variety the HapMap sample copy number results and the HGDP panel results were combined (Figure 3.19).

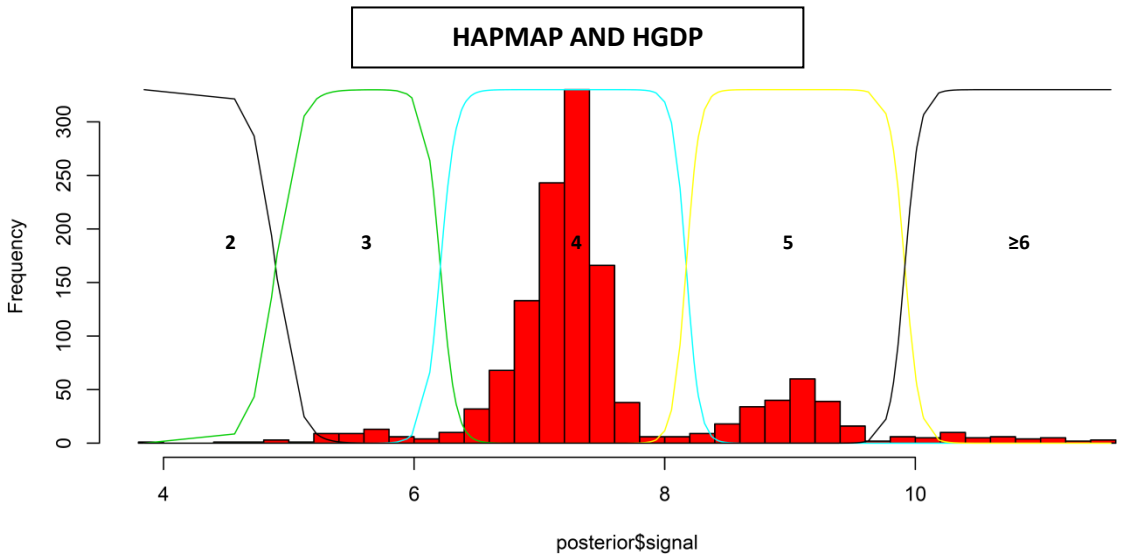


Figure 3.19: Population distribution of LCR *CR1* copy number. Output of the clustering procedure is using the PRT data of LCR *CR1* CN for HGDP and HapMap samples. The coloured lines show the Gaussian distributions for each of the five copy number classes (copy number = 2, 3, 4, 5 or ≥6). The x-axis indicates LCR *CR1* diploid copy number data after division by the standard deviation of the entire dataset.

The frequency of the copy number 5 was highest in Orcadians and it was also high for the African and the European populations (Table 3.15). According to HapMap and HGDP PRT results, the African and European populations showed a greater variety of copy number variation (LCR *CR1*) than the other populations meaning that rarer LCR *CR1* copy numbers such as 2 and >6 (*CR1*-C and *CR1*-D respectively) were observed more often in these populations (Figure 3.20).

The four co-dominant alleles of *CR1* can be related to the LCR *CR1* copy number (Table 3.14).

The copy number frequencies of *CR1* gene for HGDP samples were calculated for each population according to their localization in five different continents (Table 3.15). The copy number frequencies showed variations between different populations.

Table 3.14: Relationship between diploid LCR copy numbers and *CR1* copy number genotypes. Three of the CR1 alleles are also known by other names (B=S, C= F' and A=F).

Genotype	***CR1*-C (*CR1*-F') 1 copy**	***CR1*-A (*CR1*-F) 2 copies**	***CR1*-B (*CR1*-S) 3 copies**	***CR1*-D 4 copies**
***CR1*-C (*CR1*-F') 1 copy**	2	3		
***CR1*-A (*CR1*-F) 2 copies**	3	4	5	
***CR1*-B (*CR1*-S) 3 copies**		5	6	7
***CR1*-D 4 copies**			7	≥8

The frequency of individual alleles can be inferred from the frequency distribution of diploid copy numbers by using the algorithm implemented in the CoNVEM software. CoNVEM is a web-tool which graphically and numerically displays CNV allele and genotype distributions (Gaunt et al., 2010). The most frequent allele of *CR1* is *CR1*-A which matched with copy number 4 is the most frequent CN in the population; this copy number is expected to be mostly homozygous for *CR1*-A, with a small minority of heterozygotes for C and B. The others are combined with the same method according to their frequency in the worldwide population. Therefore, 5 copies of LCR1 *CR1* was the second most frequent CN for all populations whereas 6 copies or higher and 3 copies and lower were observed less (Table 3.16). As 5 LCR *CR1* copy number was the second most frequent in the population, it is believed to be heterozygous for *CR1*-A and *CR1*-B alleles. In addition, a LCR1 *CR1* copy number of 6 is interpreted as homozygous *CR1*-B. This is a simplification, in that a diploid copy number of 6 might be

heterozygous for *CR1*-A and *CR1*-D, for example, but the allele frequency of *CR1*-D is low at 0.002 and therefore this is unlikely. Finally, LCR1 *CR1* copy numbers of 7 and 8 were very rare when compared to the other copy numbers and can be matched with rare alleles of *CR1* showing low allele frequencies in most populations.

According to these results, the allele frequencies of Europeans and Africans showed higher variation than those of Asia, America and Oceania. Although, the *CR1*-A allele was the most common in populations overall, the second most common *CR1* allele, *CR1*-B, showed dramatic variation across different populations (Table 3.16). *CR1*-B allele was not the second most common allele in Oceania, instead this was *CR1*-C (Table 3.15). Also, *CR1*-B allele frequency was high in Africans and Europeans (Table 3.15).

In a study by Dykman et al. (1985), the allele frequencies for *CR1* alleles were found to be 0.01 for *CR1*-C, 0.83 for *CR1*-A, 0.16 for *CR1*-B and 0.002 for *CR1*-D. Moulds et al. (1996) found that the *CR1*-A allele has a frequency of 0.87 whereas *CR1*-B allele frequency is 0.11 and *CR1*-C allele frequency is 0.02 in Caucasians (n=112). These results are matching with the population results of Europeans in this study (Table 3.16). Moulds et al. (1996) also showed that the *CR1* allele frequencies in African-American (n=63) and Mexican (n=66) were 0.82 and 0.89 for *CR1*-A allele, 0.11 and 0.11 for *CR1*-B allele and 0.06 and <0.001 for *CR1*-C allele, respectively. Moulds et al. (1998) showed that the *CR1* allele frequencies in Chinese population (n=100) were 0.96 for *CR1*-A allele, 0.03 for *CR1*-B allele and 0.001 for *CR1*-C allele and none of the individuals was found to carry the *CR1*-D allele. In other research, the *CR1*-A and *CR1*-B allele frequencies were found to be 0.916 and 0.084 respectively in Asian Indians (Katyal et al. 2003). Another study (Panchamoorthy et al., 1993) showed the similar results to Katyal et al. 2003. They showed the *CR1* allele frequencies in Indian population (n=79) were 0.975 for *CR1*-A allele, 0.025 for *CR1*-B allele and none of the individuals was found to express *CR1*-C or *CR1*-D alleles. These results are matching with Asian population frequency results (Table 3.15). Finally, the results showed that the alleles, *CR1*- C and *CR1*-D, were infrequent in all populations and this was confirmed by several studies (Eikelenboom and Stam, 1982; Wong et al., 1983 and Moulds et al., 1996). According to other studies and our results, the *CR1*-A allele was

the most frequent and *CR1*-B was the second most frequent allele in all populations except in Americans (HGDP) (Table 3.16).

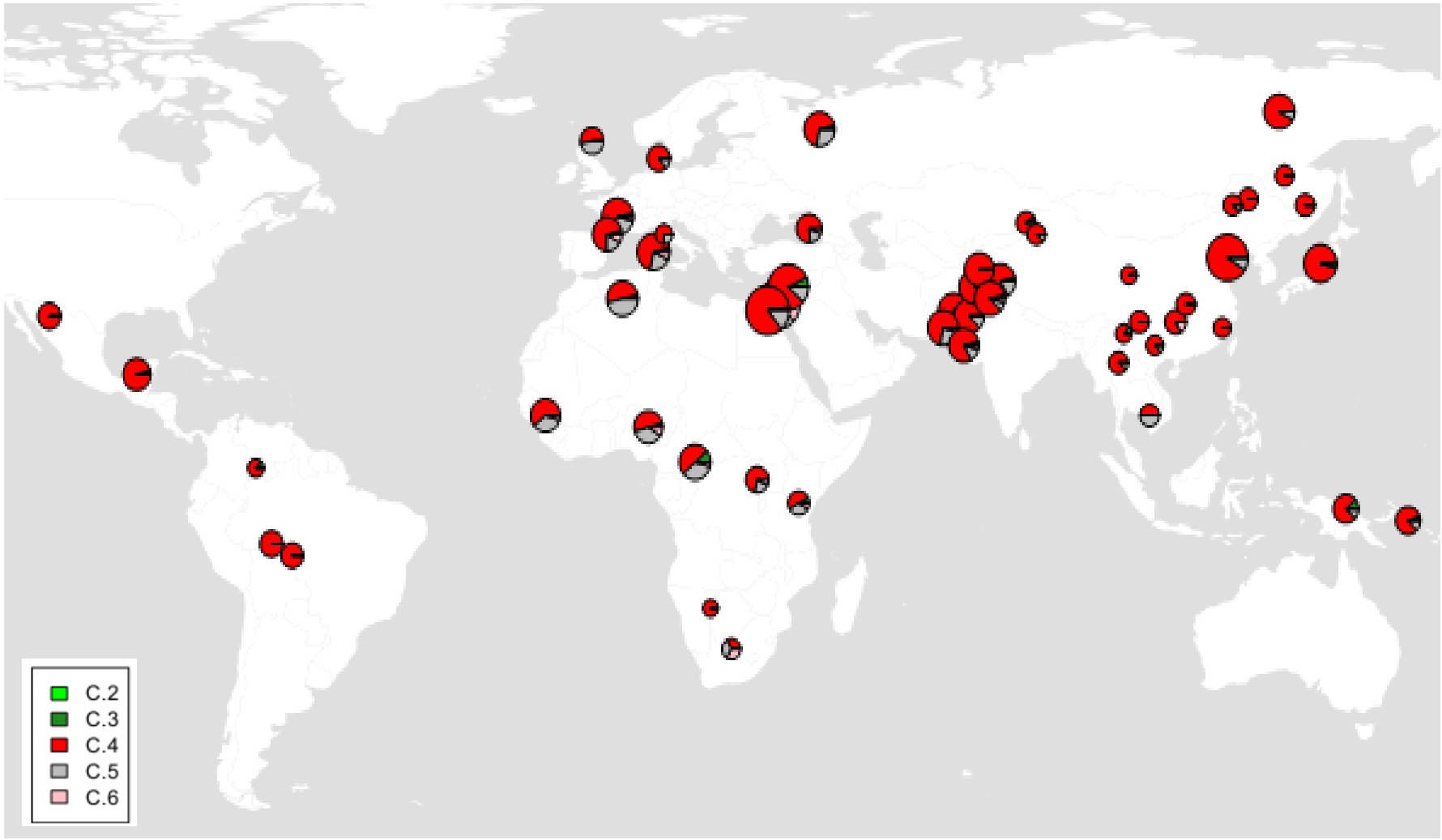

C.2
C.3
C.4
C.5
C.6

Figure 3.20: Population distribution of diploid LCR *CR1* copy number (copy number = 2, 3, 4, 5 or ≥6). Distribution of LCR *CR1* integer copy number in the HGDP populations, pie charts are in proportion to sample size.

Table 3.15: Copy number frequencies for each population. The populations were divided into five major groups according to their localization across five continents and Asia was also subdivided into three groups as western, south & central, and eastern. n= number of samples for each group.

			Copy Number Frequency				
Continent		**Population**	**2**	**3**	**4**	**5**	**≥6**
Europe		Adygei (n=17)	0	0	0.765	0.177	0.0589
		French (n=28)	0	0.0357	0.643	0.286	0.0357
		French_Basque (n=24)	0	0.0417	0.71	0.167	0.083
		North_Italian (n=14)	0	0	0.857	0.143	0
		Orcadian (n=15)	0	0	0.533	0.467	0
		Russian (n=25)	0	0.04	0.68	0.28	0
		Sardinian (n=24)	0	0.0357	0.679	0.214	0.0714
		Tuscan (n=8)	0	0	0.75	0.25	0
Africa		Bantu_Kenya (n=12)	0	0.083	0.5	0.33	0.083
		Bantu_South Africa (n=9)	0	0	0.33	0.33	0.33
		Biaka_Pygmy (n=26)	0	0.115	0.5	0.346	0.0385
		Mandenka (n=23)	0	0	0.61	0.348	0.0435
		Mbuti_Pygmy (n=14)	0	0	0.714	0.214	0.0714
		San (n=6)	0	0	1	0	0
		Yoruba (n=22)	0	0.0455	0.5	0.363	0.091
		Mozabite (n=26)	0	0.0385	0.5	0.4615	0
Asia	**Western**	Bedouin (n=51)	0.0392	0.059	0.608	0.176	0.118
		Druze (n=43)	0	0.0698	0.814	0.116	0
		Palestinian (n=50)	0	0.02	0.84	0.14	0
	Central and South	Balochi (n=24)	0	0	0.83	0.167	0
		Brahui (n=24)	0	0	0.875	0.125	0
		Burusho (n=25)	0	0.04	0.8	0.16	0
		Cambodian (n=10)	0	0	0.5	0.5	0
		Hazara (n=24)	0	0	0.91	0.083	0
		Kalash (n=23)	0	0	1	0	0
		Makrani (n=25)	0	0	0.72	0.24	0.04
		Pathan (n=25)	0	0.04	0.84	0.08	0.04
		Sindhi (n=24)	0	0.042	0.792	0.125	0.042
		Uygur (n=10)	0	0.1	0.8	0.1	0
	Eastern	Dai (n=10)	0	0	0.9	0.1	0
		Daur (n=10)	0	0	1	0	0
		Han (n=45)	0	0	0.91	0.089	0

		Hezhen (n=10)	0	0	1	0	0
		Japanese (n=30)	0	0	0.967	0.033	0
		Lahu (n=8)	0	0	0.875	0.125	0
		Miaozu (n=12)	0	0	0.83	0	0.167
		Mongola (n=9)	0	0	0.89	0.11	0
		Naxi (n=7)	0	0.143	0.857	0	0
		She (n=8)	0	0	1	0	0
		Oroqen (n=9)	0	0	1	0	0
		Tu (n=7)	0	0	1	0	0
		Tujia (n=10)	0	0	1	0	0
		Xibo (n=9)	0	0	0.89	0.11	0
		Yakut (n=25)	0	0	0.92	0.08	0
		Yizu (n=10)	0	0	1	0	0
Americas		Colombian (n=7)	0	0.143	0.857	0	0
		Karitiana (n=14)	0	0	1	0	0
		Maya (n=21)	0	0.0476	0.952	0	0
		Pima (n=14)	0	0	1	0	0
		Surui (n=13)	0	0	1	0	0
Oceania		Melanesian (n=16)	0	0.0625	0.813	0.125	0
		Papuan (n=17)	0	0.118	0.765	0.118	0

Table 3.16: The summary of *CR1* allele frequency results for HGDP samples according to five continents. The allele frequencies were calculated by an expectation-maximization program for determining allelic spectrum from diploid CNV data (CoNVEM). The results showed that the *CR1* allele distributions show variation between different populations. n=number of samples for each group.

Continent	***CR1*-C allele (1 copy)**	***CR1*-A allele (2 copies)**	***CR1*-B allele (3 copies)**	***CR1*-D allele (4 copies)**
Europe n=156	0.02	0.84	0.14	0.0077
Africa n=138	0.029	0.73	0.24	<0.0001
Asia n=577	0.011	0.92	0.058	0.0078
Americas n=69	0.01	0.99	<0.0001	<0.0001
Oceania n=33	0.0485	0.887	0.063	<0.0001

3.5 Breakpoint Analysis for *CR1* (LCR) Duplication

According to a paper published in 1993 by Vik and Wong the *CR1* alleles were produced by unequal crossing over between homologous regions of the *CR1* gene. There could be two possible breakpoints to produce the *CR1*-B allele (Figure 3.21). This

model allows the development of breakpoint junction assays for the detection of the *CR1*-B allele.

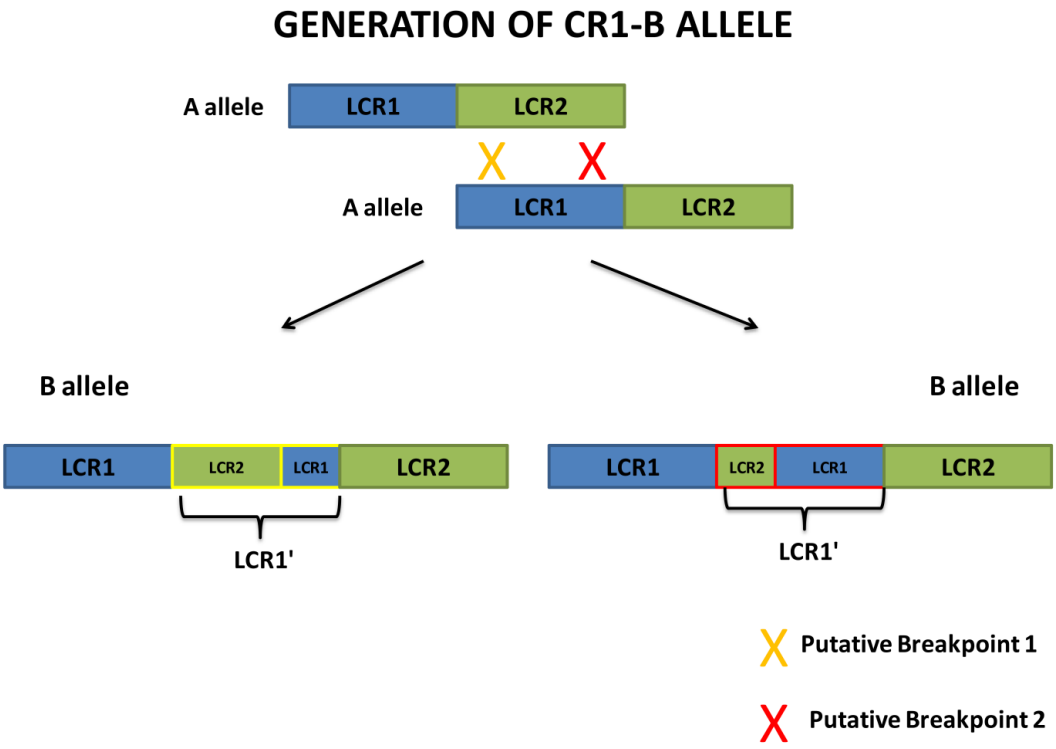


Figure 3.21: The proposed mechanism of generation of CR1-*B* alleles. The unequal crossover between homologous regions of LCR1 and LCR2 can happen in one of two points and give rise to the *CR1*-B allele.

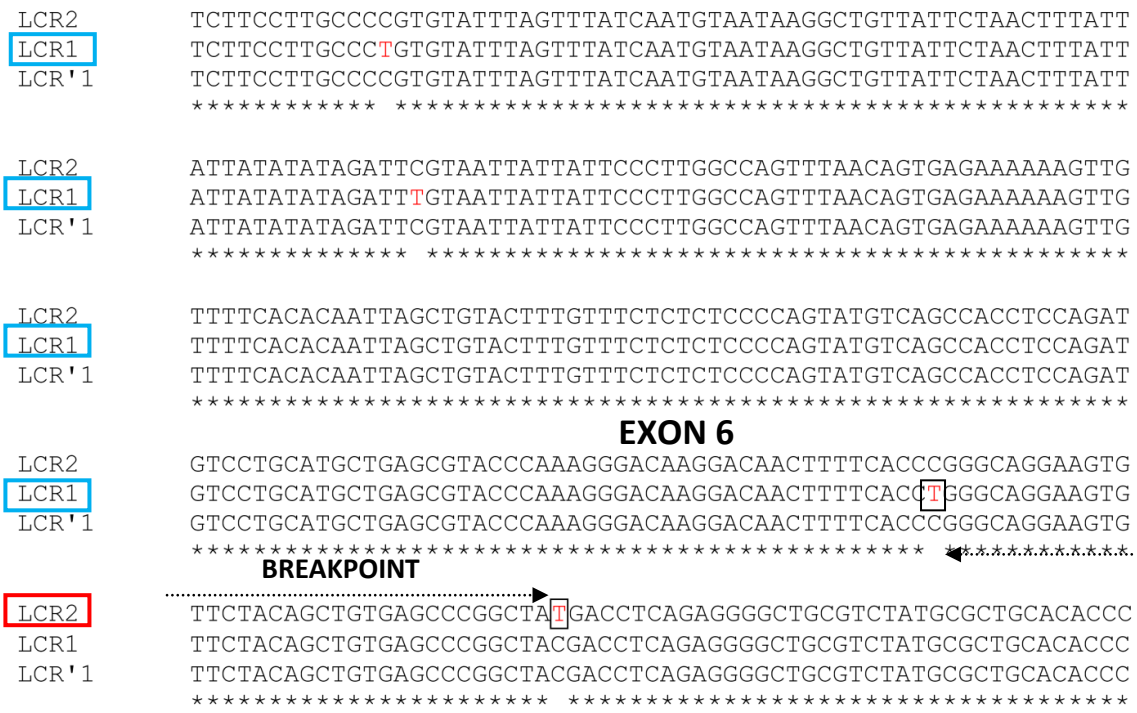


Figure 3.22: The alignment of LCR1, LCR2 and LCR1' DNA sequences of the *CR1*-B allele by Clustal Omega. The putative breakpoint for duplication of *CR1* (LCR) was proposed to be in exon 6 where the single base differences between LCR2, LCR1' and LCR1 switched from LCR1 to LCR2.

In previous study, Vik and Wong (1993) showed that exon 6 contains the breakpoint for generation of the *CR1*-B allele. To confirm this in another independent sample, the genome assembly hg19 was examined, as it shows an example of the *CR1*-B allele. The sequences for LCRs were collected separately from UCSC Genome Browser (Human Feb. 2009 (GRCh37/hg19) Assembly). Then, these regions were aligned using Clustal Omega. In order to find the putative breakpoint for duplication (*CR1*-B allele) the base differences between all three LCRs were checked and the point where the base differences switched from LC1 to LCR2 defined as the breakpoint (Figure 3.22). Because the breakpoint for the *CR1*-B allele identified in the genome assembly and by Vik and Wong (1993) is the same, it suggests that this breakpoint is responsible for most, and possibly all, *CR1*-B alleles.

To develop a junction-fragment assay for the *CR1*-B allele across the exon 6 breakpoint, LNA primers were designed to bind the samples with LCR1' (*CR1*-B) region (Figure 3.23). The LNA primers were chosen to bind around the breakpoint. Therefore, the LNA base was located at the 3' end only.

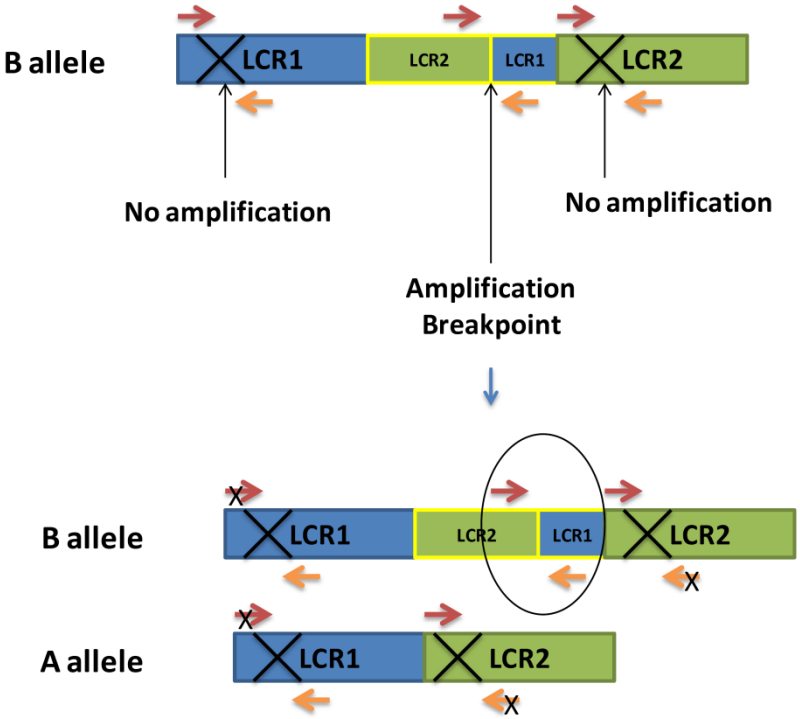


Figure 3.23: The binding regions of forward and reverse LNA primers on LCRs of *CR1*-A and *CR1*-B alleles. LNA primers take advantage of the base differences between LCR1, LCR1' and LCR2 to only give PCR products in the existence of *CR1*-B allele. The forward primers do not bind to the LCR1 region of *CR1* whereas the reverse primers do not bind to the LCR2 region of *CR1*. They both bind the LCR1' region but only samples with *CR1*-B alleles give a product.

The LNA primers were designed to study the existence of *CR1*-B designed around the breakpoint of the *CR1* gene (Figure 3.24).

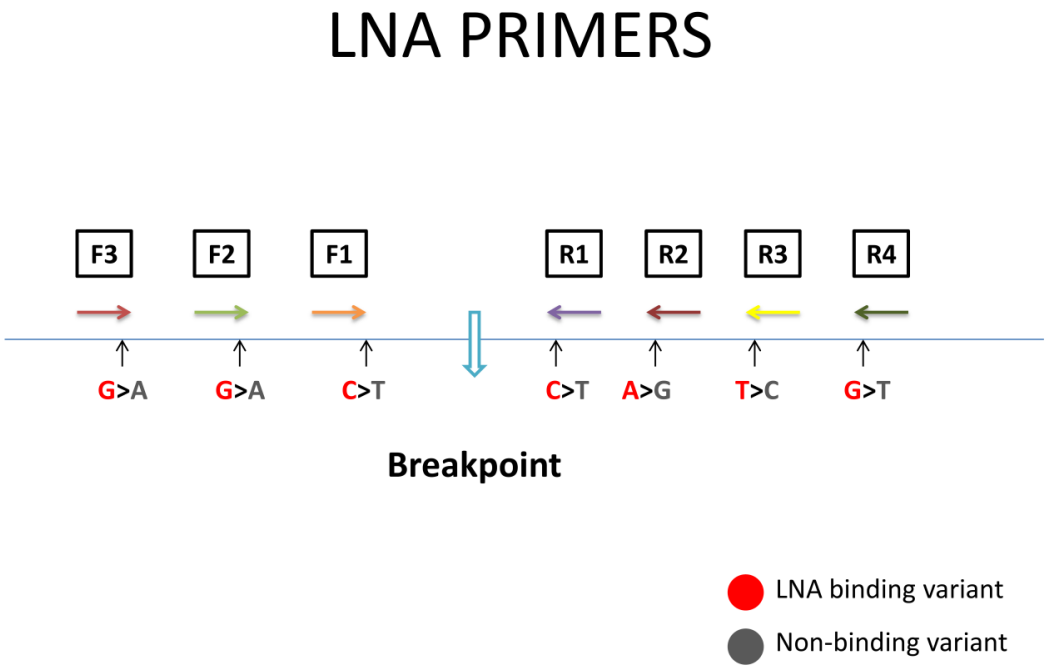


Figure 3.24: The LNA primers designed for detection of the exact breakpoint for the generation of *CR1*-B alleles. The forward LNA primers bind LCR2 and LCR1' because of the base differences between LCR1, LCR1' and LCR2 whereas the reverse LNA primers bind to LCR1 and LCR1'. Therefore, the primers only amplify the LCR1' region. The variants which bind to the LNA primers are labelled as red whereas the non-binding variants are labelled as grey.

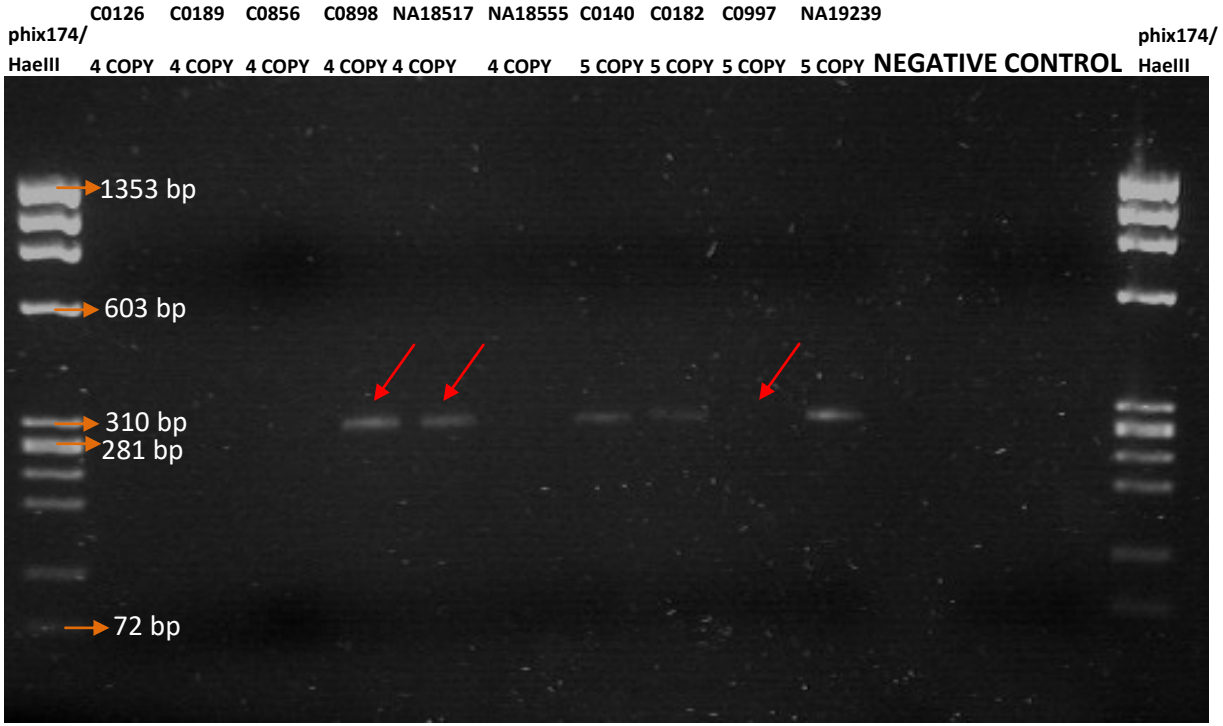


Figure 3.25: *CR1*-B breakpoint-specific PCR analysis using LNA_F1 and LNA_R1 primers. The PCR product of the touchdown PCR was run on a 2% (w/v) agarose gel and using the marker phix174/HaeIII (200 ng loaded). The predicted PCR product size was 287 bp. The primers should give a product for 5 copy number samples (which should carry the *CR1*-B allele) whereas no bands were expected for 4 copy number samples (which will not carry the *CR1*-B allele). According to the gel results were shown above C0898 and NA18517 were not expected to work as they carried 4 copies, whereas C0997 were expected to give a band as it was a 5-copy sample (red arrows). Therefore, the results are not reliable enough to study the breakpoint of *CR1* (LCR).

Several LNA primers were designed targeting the sequence differences between LCR1 and 2 across the breakpoint to try and generate a breakpoint-specific PCR. A series of samples typed by PRT (HRC1) were run on the PCR to assess specificity of the breakpoint-specific PCR. For the first breakpoint analysis, LNA_F1 forward primer and LNA_R1 primer pair were used. The primers were not specific enough to discriminate the difference between *CR1*-A and *CR1*-B (duplication allele). In order to increase the discrimination power of the LNA primers, touchdown PCR was applied for the experiments. Even though the touchdown PCR method was chosen for breakpoint analysis for some of the samples (European and African individuals) it failed to show the existence or non-existence of the *CR1*-B allele in a particular individual meaning that sometimes it gave no bands for a 5-copy individual (assumed to contain 1 *CR1*-B and 1 *CR1*-A alleles) or gave bands for 4-copy individuals (assumed to contain 2 *CR1*-A allele) (Figure 3.25). Therefore, other LNA primers were designed including L2, L3, R2, R3 and R4. The primers L2, R2 and R4 failed to give specific and single PCR product so they were not used in the rest of this study. Three primers (LNA_F2 (forward) and LNA_F3 (forward) and LNA_R3 (reverse) primers) were designed which amplified a wider region around the breakpoint. Only European samples were used for this assay as European populations are more likely to show *CR1*-B alleles with only one breakpoint position. For the small panel of samples analysed, LNA_F3 and LNA_R3 produced a clear band only in 5-copy individuals, suggesting that this was a successful breakpoint-specific PCR for the *CR1*-B allele (Figure 3.26). This was confirmed on a further 24 samples UK samples from the HRC1 collection. Therefore, the third assay (LNA_F3 and LNA_R3) was used to assess and confirm the existence of *CR1*-B allele in further studies described later in this thesis.

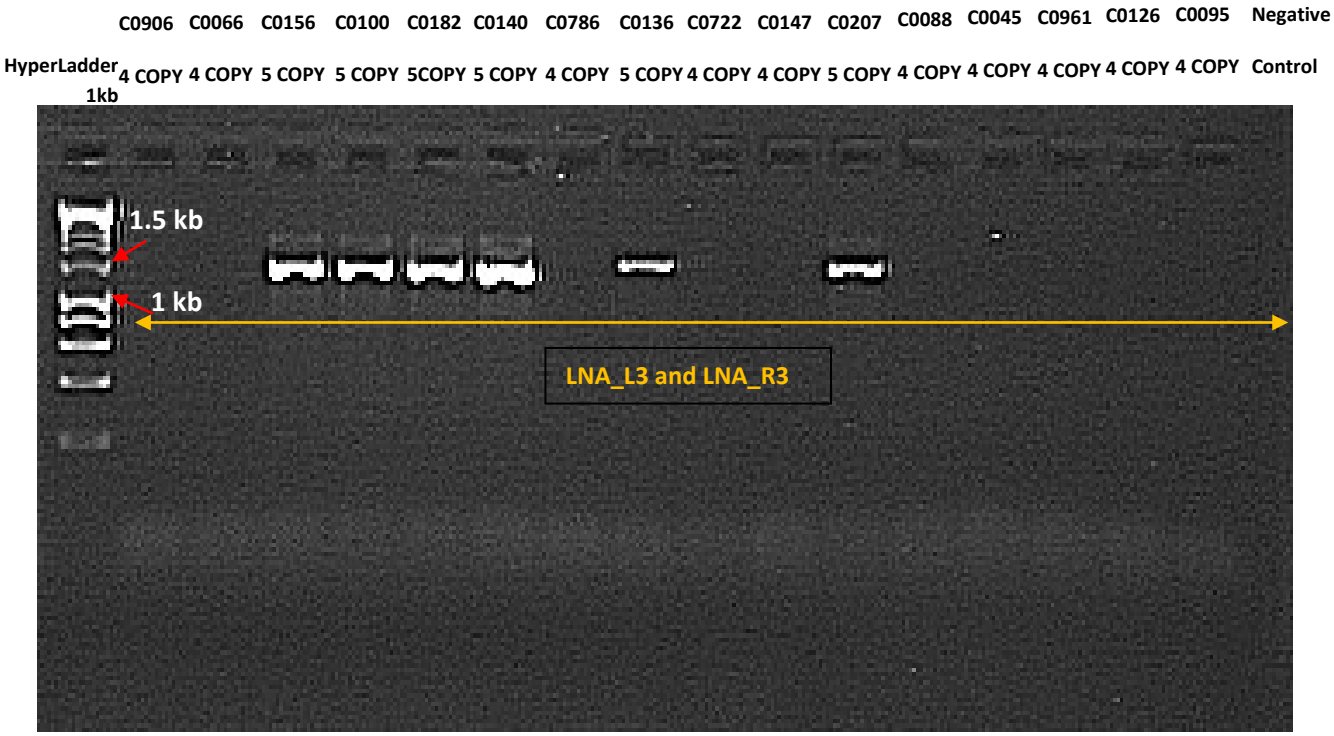


Figure 3.26: *CR1*-B breakpoint-specific PCR analysis using LNA_F3 and LNA_R3 primers. The PCR product of the touchdown PCR was run on a 2% (w/v) agarose gel. The expected PCR product size was 1436 bp for LNA_L3 and LNA_R3. The samples were chosen from HRC1 plate (A1 to H1 and A2 to H2). The primers should give bands for 5-copy number samples (LCR *CR1*) whereas no bands were expected for 4-copy number samples. According to the gel results shown above the primers worked specifically with these samples. Therefore, no bands were observed for the samples with 4 copies.

3.6 Heterogeneity of the PRT3 Assay

During PRT analysis (PRT1, PRT2 and PRT3) of worldwide populations, PRT3 assay showed variation (a drop in copy number as a comparison to PRT1 and PRT2) (Table 3.17) for some of the African samples (HGDP n=11; 1.13% and HapMap3: n=2; 2.1%) (Appendix 14).

Table 3.17: Some of the HGDP (African) and HapMap3 samples showed a drop in PRT3 assay. The PRT copy number data are raw and normalized according to the reference copy numbers.

Sample ID	Plate	PRT1	PRT2	PRT3
HGDP00460	HGDP	4.05	4.22	2.62
HGDP00449	HGDP	3.89	4.17	2.59
HGDP00456	HGDP	5.17	4.98	3.11
HGDP00913	HGDP	3.93	4.10	2.42
HGDP00915	HGDP	3.91	4.05	1.99
HGDP00175	HGDP	4.59	4.61	2.81
HGDP00701	HGDP	3.85	3.70	1.89
HGDP00739	HGDP	3.78	3.79	1.31
HGDP00921	HGDP	4.03	4.16	2.13
HGDP00923	HGDP	3.89	4.13	2.02
HGDP00468	HGDP	3.91	3.86	2.06
NA19152	HAPMAP3	4.06	3.79	2.03
NA19094	HAPMAP3	4.07	3.85	2.08

This variation may be caused by either deletion in *CR1* gene or duplication in the *CR1L* gene affecting the binding site of the PRT3 primers. In order to understand the reason for the variation four sets of new PRT primers were designed. These primers were designed to study the copy number variation of the *CR1L* gene, so the reference region is the *CR1* gene and the test region is the *CR1L* gene (Figure 3.27). Importantly, samples that only had *CR1* copy number of 4 were analysed in these analyses to ensure that variation was due to *CR1L*.

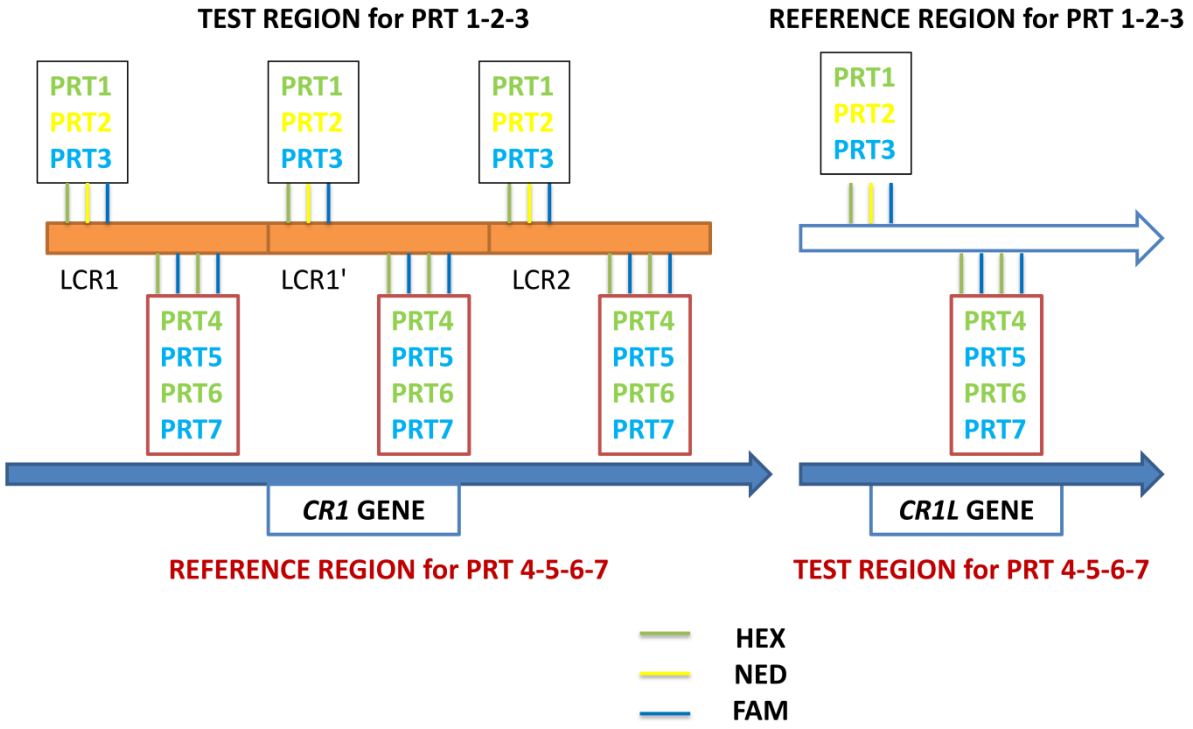


Figure 3.27: The comparison of PRT primers designed to assess the copy number of *CR1* and *CR1L* genes. The first set of primers were PRT1 (HEX), PRT2 (NED) and PRT3 (FAM), designed to understand the copy number variation of LCR of *CR1* in worldwide populations whereas the second PRT primer sets were PRT4 (HEX), PRT5 (FAM), PRT6 (HEX) and PRT7 (FAM), designed to study the copy number variation of *CR1L* in African samples, as the PRT3 assay showed variation during copy number analysis in some of the African samples. The forward primers are labelled at their 5' end with HEX (green), NED (yellow) or FAM (blue) dye.

During the experiments, four samples (HGDP00921, HGDP00923, NA19152, HGDP00468 and NA19094) which showed variation for PRT3 assays were used in order to investigate the cause of the heterogeneity. Two control samples (NA18555 and NA18517) which were used for PRT1, PRT2 and PRT3 assays and gave a copy number of 4 for *CR1* were used for the new PRT assays. The details of the samples are included in Appendix 13. According to the PRT results, there was no significant copy number difference between the samples showing variation for PRT3 assay and the controls to indicate any duplication in *CR1L* or deletion in *CR1* for any of the PRT assays (Table 3.18).

Table 3.18: PRTs (PRT4-5-6-7) were designed to specifically bind *CR1* (LCR1 and LCR2) as reference and *CR1L* as test region. The raw data of the PRT4-5-6-7 results are shown below.

Raw *CR1L* CN→	PRT4	PRT5	PRT6	PRT7	*CR1* CN
HGDP00468	0.47	0.32	0.38	0.24	4
HGDP00921	0.45	0.26	0.41	0.21	
HGDP00923	0.46	0.27	0.42	0.21	
NA19094	0.56	0.25	0.4	0.22	
NA19152	0.44	0.29	0.39	0.22	
NA18517(control)	0.34	0.24	0.45	0.22	
NA18555(control)	0.41	0.25	0.49	0.18	

According to the results shown above, the test region on *CR1L* using PRT4-7 showed no sign of copy number variation on *CR1L* gene. The previous results which implied a copy number variation of *CR1L* (duplication) or *CR1* deletion may have resulted from a SNP in the PRT3 primer binding site or a small deletion of the PRT3 region in the *CR1* gene, or a small duplication in the *CR1L* gene that it was outside the region targeted by the PRT4-7 primers. Therefore, further investigations are needed to explain the heterogeneity of PRT3 assay. It seems likely that a single nucleotide variant underneath the PRT3 primers in *CR1* may be responsible for this.

No evidence of a single nucleotide variant was found underneath the PRT3 primer binding sites in *CR1* (there are some SNPs in the *CR1L* binding site: [rs140123007], but not in *CR1*) by examining the dbSNP track in the genome browser.

3.7 Discussion

The present study shows that PRT1-2-3 assays are robust in accurately calling CNVs of *CR1* (LCR) and can be used to estimate CNVs of *CR1* in large disease association (case-control) studies.

The HGDP and the HapMap samples were typed to study *CR1*'s (CNV) general pattern of global variation across worldwide populations. PRT was a cost effective method and worked efficiently with low quantities of genomic DNA (5-10 ng). The PRT assays gave a good estimation of the *CR1* (LCR) copy number. When they were compared to aCGH and sequencing read-depth data, they agreed with the latter while showing variation from aCGH. Therefore, PRT appeared better than aCGH for determining the LCR *CR1* copy number in HapMap samples. qPCR is not chosen for copy number calling in this study as it is not suitable for use in large scale case-control studies (Haridan et al., 2015), so PRTs were preferably used for large case-control studies instead of qPCR. PRTs also agreed with the *CR1* allele frequency results obtained from previous studies. For example, the results are in agreement with the literature as the allele frequencies of *CR1* matched with the previous *CR1* population studies (Dykman et al., 1985; Katyal et al., 2003). The most common allele was found to be *CR1*-A (2 copies) and the second commonest allele found to be *CR1*-B, even though the frequency of *CR1*-B (3 copies) showed a high variation from population to population. The study of Madi et al. (1991) found that erythrocytes were more efficient in binding immune complexes when carrying B or D alleles compared to those having *CR1*-C or *CR1*-A structural alleles of *CR1*. Therefore, having the *CR1*-B allele or *CR1*-D may be favoured in some of the populations. During the process of fighting against pathogens (infectious diseases) in harmful environments, *CR1*-B allele frequency can increase in those populations due to the selective pressure that these pathogens place upon the population. Because of that, the possibility of *CR1*-B allele association to malaria is examined in the next chapter.

In addition, the results of this study also confirmed the findings of Vik and Wong (1993). That study found that exon 6 of *CR1*-A allele contains the breakpoint for generation of the *CR1*-B allele. Also, the LNA primers designed for this investigation can be used to assess the existence of the *CR1*-B allele in European samples.

Taken together, the PRT and breakpoint-specific assays can be used to genotype large disease cohorts in European samples.

In African populations, the PRT3 assay should be interpreted carefully. A small number of the African samples from HGDP and HapMap3 have shown a drop of *CR1* LCR copy number in the PRT3 assay which differs from the PRT1 and PRT2 assays. No explanation was found for the heterogeneity of PRT3 assay. It could result from mutation or polymorphism in the binding region of PRT3 primers or a small deletion in *CR1* or duplication in *CR1L*. Sequencing could be used around the primer binding regions (or even more widely) in both the *CR1* and *CR1L* genes for future studies.

4 ASSOCIATION OF COPY NUMBER VARIATION WITHIN *CR1* WITH MALARIA

4.1 Introduction

Malaria continues to be a serious public health problem. Globally, malaria deaths increased from 995,000 in 1980 to 1,817,000 in 2004 and then decreased to 1,238,000 in 2010 (Murray et al., 2012). Particularly in Africa, malaria deaths increased from 493,000 in 1980 to 1,613,000 in 2004 and then decreased around 30% to 1,133,000 in 2010. The increase through the 1980s and 1990s to a peak in 2004 was due to increasing chloroquine resistance, and first-line antimalarial drug resistance. The other explanation for this increase could be the interaction between HIV infection and malaria, which leads to more malaria cases when compared to non-HIV infection and malaria. After a global peak in 2004, there was a significant decrease in malaria deaths. This can be attributed to an improvement in the healthcare infrastructure in Sub-Saharan Africa. Investments from establishments such as the Global Fund to fight AIDS, Tuberculosis and Malaria have brought about 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 and the artemisinin-combination treatment and vector control strategies are critical for addressing growing antimalarial drug resistance (Murray et al., 2012).

Each year around the world, between 75,000 and 200,000 infant deaths are related to malaria infection in pregnancy. In endemic areas, pregnant women are more susceptible to malaria and the frequency and the severity of disease is higher in comparison to adults or non-pregnant women. Malaria in pregnant women is related to a high frequency and density of *P. falciparum* parasitaemia and causes high rates of maternal morbidity including fever and severe anaemia, abortion and placental malaria (Uneke, 2007). The prevalence of malaria in pregnant women is affected by maternal age, gravidity, use of prophylaxis, nutrition, host genetics, level of anti-parasite immunity, parasite genetics and transmission rates. The infants born to mothers with placental malaria are at an increased risk of anaemia, malaria infection rates and mortality during their first year of life (Tako et al., 2005). There is a variety of

factors affecting the host's response to infection such as the intensity and seasonality of malaria transmission, the virulence of the parasite which can depend on its genetic polymorphism, and also 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 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 tolerant phenotype to blood-stage antigens which could persist during childhood (Le Port et al., 2012).

According to the World Health Organization (WHO), children under 5 years of age (WHO, 2016a) are one of the most vulnerable groups affected by malaria (Figure 4.28). In 2015, 438,000 malaria deaths have been estimated around the world and 69% of these were children under 5 years of age. Additionally, they reported that newborns and infants less than 12 months of age (WHO, 2016b) are one of the most vulnerable groups affected by malaria and they are at risk of rapid disease progression, severe malaria and death. Infants at 3 months of age become vulnerable to *P. falciparum* malaria as their acquired immunity from their mothers begins to decrease.

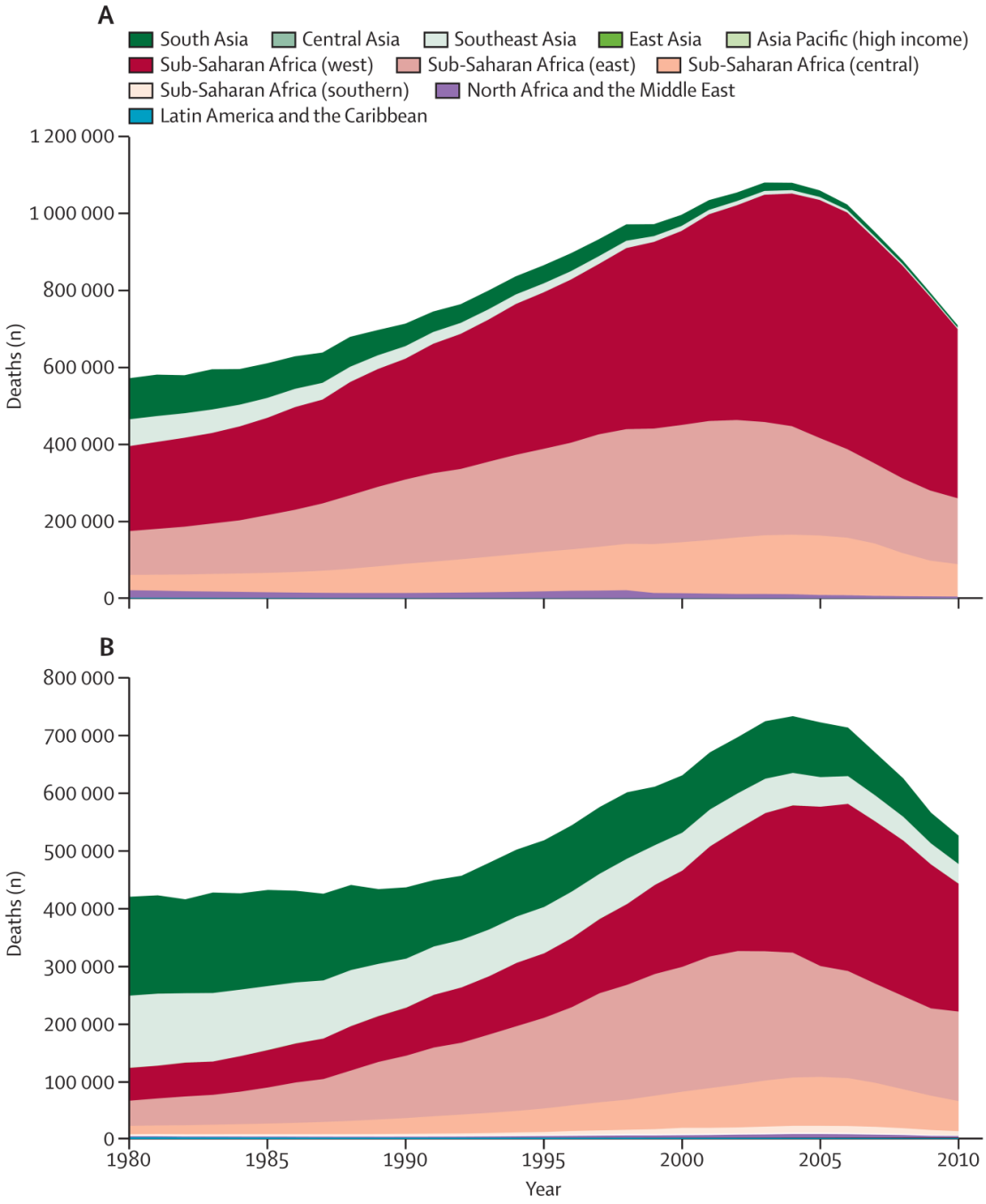


Figure 4.28: (A) Malaria deaths by global burden of disease study region for children younger than 5 years and (B) individuals aged 5 years or older, 1980 to 2010. The under 5 years old age group is much smaller than the over 5 years old age group, but has more deaths. Therefore, the chance of death in children under 5 years old is higher (infants <5 years old) but decreases as age increases (>5 years old). Taken from Murray et al. (2012).

A study by Griffin et al. (2014) showed that malaria imposes its greatest burden in infants and young children in highly endemic areas whereas in areas of lower transmission many cases occur in older children and adults. It means there is a shift in cases to older ages with declining transmission. For example, between the time

periods 1996-2010 and 2003-2007 in south western Senegal and western Gambia respectively, a substantial reduction in malaria incidence was observed to be accompanied by an increase in the mean age distribution of clinical cases. Malaria cases in the under-fives fell from 34% to 5% between 1996-2010 in Senegal, while the mean age of paediatric malaria admissions in Gambia rose from 3.9 years to 5.6 years.

In high transmission areas of malaria where the parasite prevalence in 2-10 years old is measured at around 60%, 57% of the cases are predicted to be children less than 5 years of age. In lower transmission areas of malaria where the parasite prevalence in 2-10 year olds is measured at around 20%, 21% of the cases are predicted to be children less than 5 years of age, with another 22% in children aged 5-10 years. In low transmission areas of malaria where the parasite prevalence in 2-10 year olds is measured at around 5%, 61% of the cases are predicted to be children more than 15 years of age with another 10% in children aged less than 5 years (Griffin et al., 2014).

Therefore, studies of the genetic susceptibility of infants to malaria infection are warranted. In this chapter I analyse the relationship between variation in the *CR1* gene, which encodes a receptor for *P.falciparum* (see 1.7 CR1 and Malaria in introduction) and susceptibility to malaria in two infant cohorts from Benin in West Africa.

Aims of this chapter:

- To investigate an association between *CR1* CNV and malaria.
- To assess the role of Knops blood group antigen-determining SNPs in infant malaria infections, and ask if they provide protection against malaria.
- To investigate natural selection on the Swain-Langley Sl2 blood group allele of *CR1*.

4.2 Description of the Malaria Cohorts Used

Two cohorts were used and they were obtained from two longitudinal multidisciplinary projects. The Tori-Bossito cohort and Tolimmunpal (Projet Tolérance Immunitaire Associée au Paludisme) cohort were financed by l'Agence Nationale de la Recherche (ANR) (http://www.agence-nationale-recherche.fr/en/). These were

projects managed by Atlantique Department of southern Benin by David Courtin and André Garcia. The malaria data which were analysed in this project were from the PALNOUGENENV cohort in Tori-Bossito (following 600 children from birth till 18 months) and the Tolimmunpal cohort at Allada (following 400 children from birth till 24 months and their mothers during pregnancy). The locations of the malaria samples are shown below (Figure 4.29).

4.3 Tori-Bossito Cohort

The Tori-Bossito project was conducted from June 2007 until January 2010. The aim of the project was to investigate the relation between placental malaria infection and peripheral parasitaemias in infants during the first years of their lives while taking into account environmentally related 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. Tori Bossito cohort is a cohort of 656 infants (only 600 infants from the study of Le Port et al. (2012) were provided to us for our study) followed by a parasitological (symptomatic and asymptomatic parasitaemia) and nutritional follow-up from birth to 18 months. Ecological, entomological and behavioural data were collected along the duration of the study (Le Port et al., 2012). Our collaborators sent us 583 of these DNA samples which were collected from a rural area in Benin with two seasonal peaks in malaria transmission. These samples were successfully typed for *CR1* LCR copy number using the PRT method (2.3.1 Paralogue Ratio Test (PRT) for *CR1* and *CR1L* Genes) and were also genotyped for SNPs that define Knops blood group antigens.

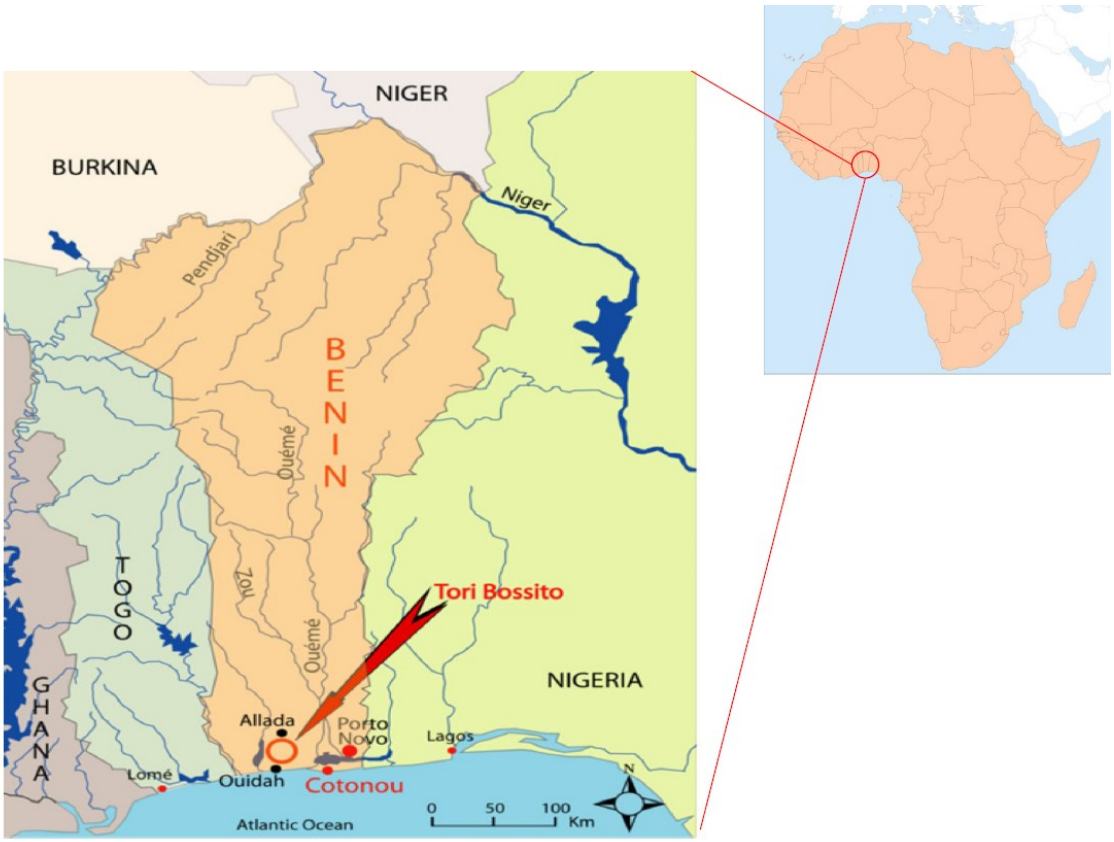


Figure 4.29: Geographical locations of Tori-Bossito and Allada (Benin) are shown. These are two locations where the malaria samples were collected. The red arrow indicates the location of Tori-Bossito, Benin. Adapted from Le Port et al. (2012) and http://www.freeworldmaps.net/.

4.3.1 CR1 copy number typing on the Tori-Bossito Cohort

The Tori-Bossito cohort was successfully typed for *CR1* LCR1 CN using the PRT method. Three samples of this cohort have been excluded from the total samples as they have some uncertainties about the basic information (such as sample name) recorded for them. Therefore, the total sample number typed for CNV with PRT assays was 580.

The histogram analysis of raw PRT ratios shows clustering of values and indicated a total of four clusters (Figure 4.30). The first cluster indicated a diploid copy number 3 or fewer (2 copy number is rare in all populations) and the remaining clusters were assigned diploid copy numbers of 4, 5 and 6 or more (7 or 8 copy numbers are rare in all populations) respectively (Figure 4.30). The copy number counts and frequencies for the Tori-Bossito cohort are shown below (Table 4.19). The allele frequency for the Tori-Bositto cohort for each *CR1* allele is shown below (Table 4.20).

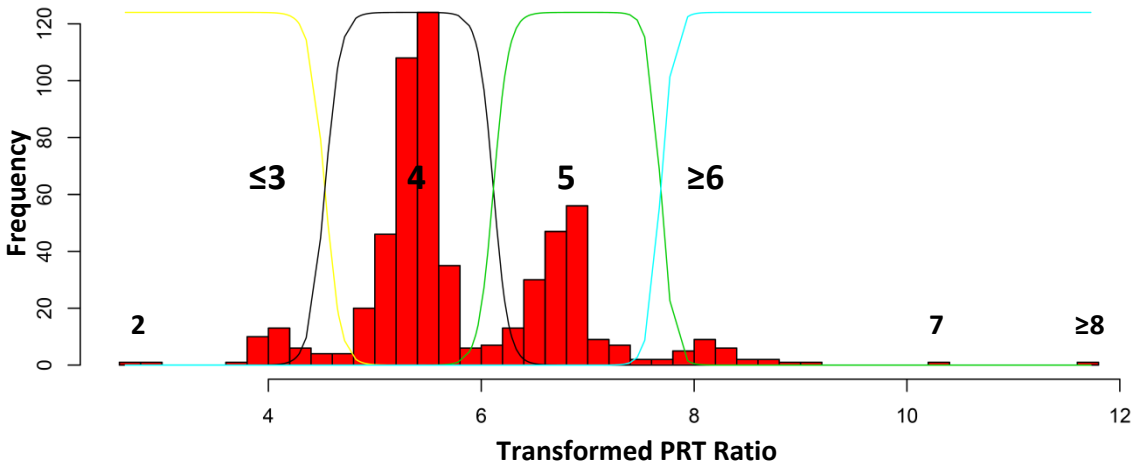


Figure 4.30: Population distribution of LCR *CR1* copy number in Tori Bossito cohort. The coloured lines show the Gaussian distributions for each of the four copy number classes (copy number = ≤3, 4, 5 or ≥6). The x-axis indicates LCR *CR1* diploid copy number data after division by the standard deviation of the entire dataset. The rare copy numbers of *CR1* carrying the *CR1*-D allele are labelled with smaller numbers as 7 and ≥8 whereas C/C genotype (*CR1*-C allele) is labelled as 2 (diploid copy number). They represent a small portion of the Tori-Bossito cohort**.**

Table 4.19: The summary of diploid copy number frequencies of *CR1* (LCR) in the Tori-Bossito cohort. The counts for each copy number are also shown.

	Diploid Copy Number						
	2 copies	**3 copies**	**4 copies**	**5 copies**	**6 copies**	**7 copies**	**≥8 copies**
Number of observations	2	37	342	168	28	1	2
Frequency	0.00345	0.0638	0.59	0.29	0.0483	0.00172	0.00345

Table 4.20: The summary of *CR1* allele frequency results for Tori-Bossito cohort. The allele frequencies were calculated by an expectation-maximization program for determining allelic frequency spectrum from diploid CNV data (CoNVEM) (http://apps.biocompute.org.uk/convem/). In the table, n indicates the sample number.

Population	***CR1*-C** **F' allele** **(1 copy)**	***CR1*-A** **F allele** **(2 copies)**	***CR1*-B** **S allele** **(3 copies)**	***CR1*-D** **D allele** **(4 copies)**
Tori-Bossito **n=580**	0.0419	0.7582	0.1885	0.0102

4.3.2 Genotyping Knops Blood Group SNPs in Control Samples

The single nucleotide variants responsible for the Knops blood group variation are all within a single exon (exon 29 of *CR1*-A allele, exon 39 of *CR1*-B allele) of the *CR1* gene (Figure 4.31). I developed a PCR-ASO (allele-specific oligonucleotide) assay to genotype the single nucleotide variants responsible for the Knops a/b (rs41274768), Swain-Langley 1/2 (rs17047661), KAM+/- (rs6691117) and McCoy a/b (rs17047660) antigens. The control samples which were used in this experiment were NA18507, NA18555, NA18517, NA19239 and NA18572. Their genotype data for the four SNPs were collected from the 1000 Genomes Project browser (http://www.internationalgenome.org/). Two HRC1 samples (C0182 and C0140) were used as European sample controls.

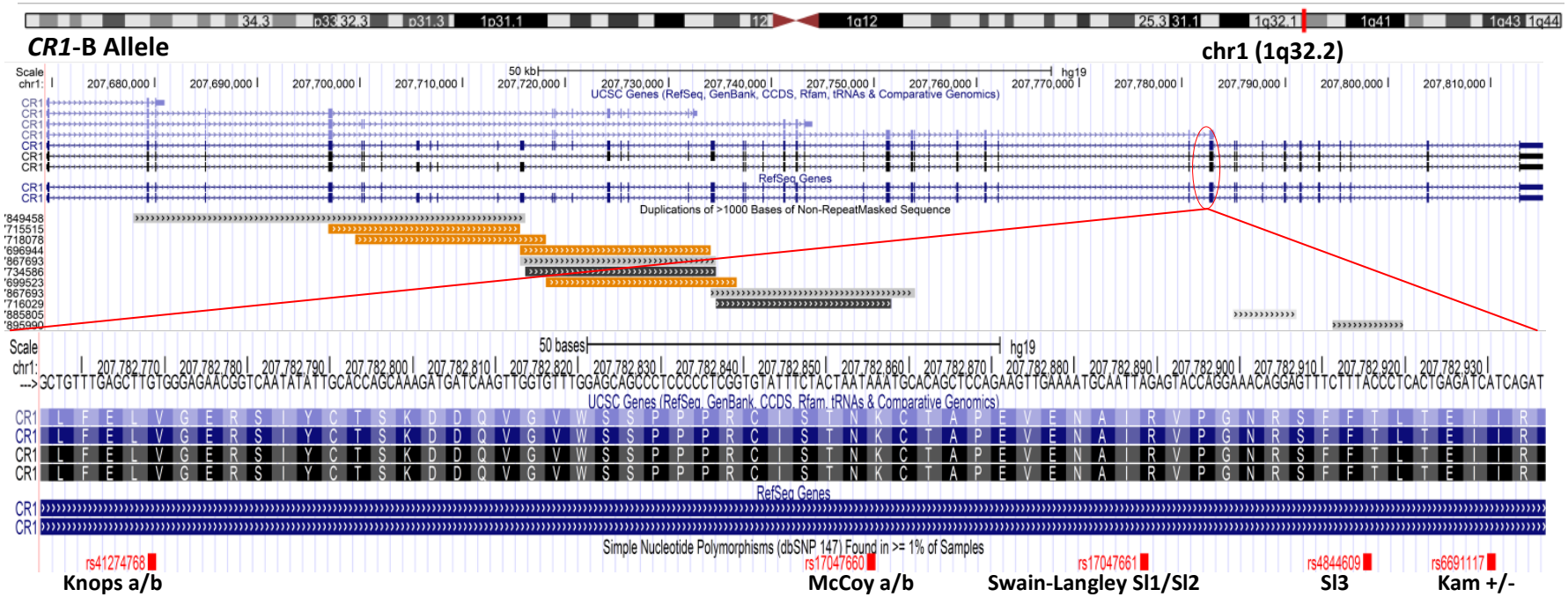


Figure 4.31: The single nucleotide variants responsible for the Knops blood group variation are all within a single exon (Exon 29/39) of the *CR1* gene. rs4844609 also determines the Knops blood group antigen Sl3 and was not tested for this research.

Additionally, in order to confirm the dot-blotting results (ASO assay), the control samples which were used during the ASO assay were sequenced by the Sanger sequencing method. The DNA region was amplified by ASO primer to confirm the genotypes of four SNPs which are determining the Knops blood group antigens (Table 4.21).

Table 4.21: The Sanger sequencing results for the controls: The DNA region which contains 5 SNPs (rs41274768, rs17047660, rs17047661, rs6691117 and rs4844609 (not used in the ASO assay but within the DNA region)) determining Knops blood group antigens of seven control samples were sequenced. The genotype data were compared and matched with the ASO assay genotyping results.

Samples ID	rs41274768	rs17047660	rs17047661	rs6691117	rs4844609
NA18507	GG	AG	AG	GG	TT
NA18555	GG	AA	AA	AA	TT
NA18517	GG	AA	AG	GG	TT
C0140	GG	AA	AA	AG	TT
C0182	GG	AA	AA	AG	TA
NA18572	GG	AA	AA	AG	TT
NA19239	GG	AA	AG	GG	TT

The sequencing results of NA18507 (control sample) confirmed the ASO assay result (Figure 4.32). The results showed that rs17047661 is heterozygous (AG) for control sample NA18507. The sequencing results of controls confirmed the ASO assay results (Figure 4.33).

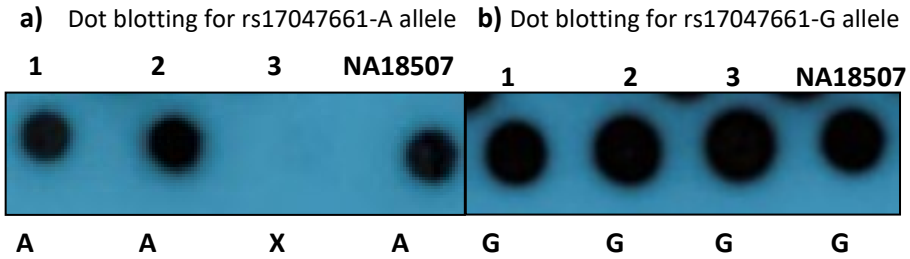
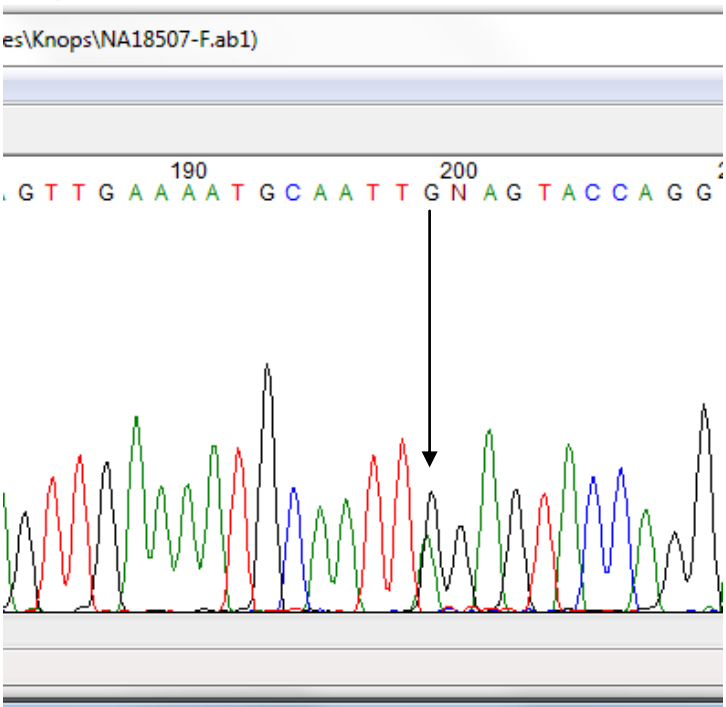
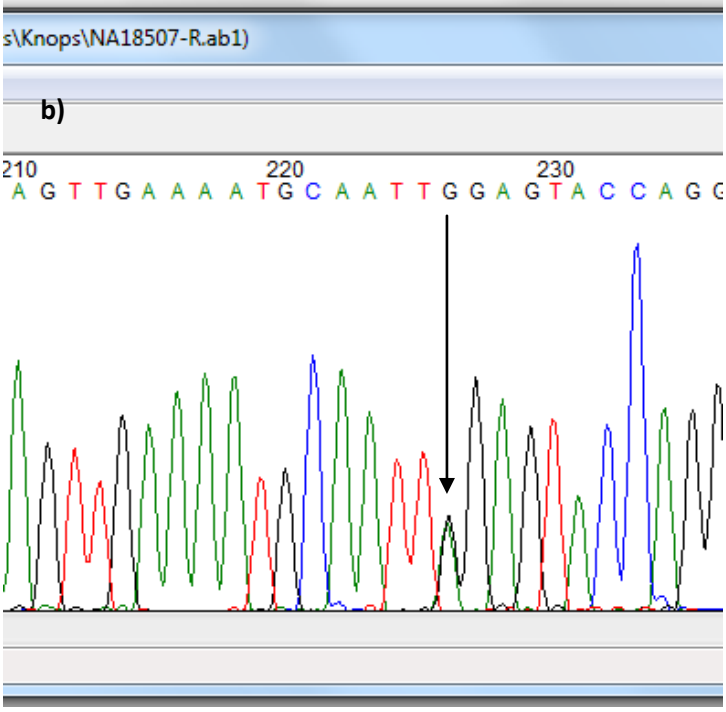


Figure 4.32: An example of dot blotting result of ASO assay: three samples from the Tori-Bossito cohort and a control sample (NA18507) are shown to determine the alleles of rs17047661 polymorphism for each sample. a) Dot blotting results of four samples for rs17047661-A allele, b) Dot blotting results of four samples for rs17047661-G allele. According to the above results samples 1, 2 and NA18507 are AG for rs17047661 whereas sample 3 is GG for rs17047661.

NA18507

a)


es\Knops\NA18507-F.ab1)
190
200
G T T G A A A A T G C A A T T G N A G T A C C A G G


s\Knops\NA18507-R.ab1)
210
220
230
A G T T G A A A A T G C A A T T G G A G T A C C A G G

b)

Figure 4.33: The sequencing result of rs17047661 for control sample NA18507: The picture shows the alignment of forward and reference sequences (obtained from UCSC Genome browser). In the image, the two windows show the forward (a) and reverse (b) DNA strands sequencing results: (a) is from forward primer. The SNP (rs17047661) is located at base 199 (forward) gives two peaks in the same position as the SNP is heterozygous (A/G). (b) is from reverse primer but the sequence trace has been reverse complemented by MEGA. The SNP (rs17047661) located at position 226 (reverse) gives two peaks in the same position again as the SNP is heterozygous (A/G). This confirms that rs17047661 is heterozygous (A/G) for control sample NA18507.

4.3.3 Genotyping Knops blood group SNPs in the Tori-Bossito Cohort

The genotype data for 582 samples (genotype results of one sample, C349, were not able to be collected) were obtained by ASO assay (Table 4.22). One SNP (rs41274768) was not polymorphic in this cohort. There was no significant departure from HWE (Hardy-Weinberg equilibrium) for the other three SNPs.

Table 4.22: The summary of genotype counts, allele frequencies, and test for departure from Hardy-Weinberg equilibrium for Knops blood group antigen-determining SNPs (Tori-Bossito Cohort). In the table, N indicates the sample number. rs41274768 was not polymorphic.

			Genotype frequency and counts			Allele frequency and Counts		Departure from HWE	
Variant	**Major Allele**	**Minor Allele**	**Major Homozygote**	**Heterozygote**	**Minor Homozygote**	**Major Allele Frequency**	**Minor Allele Frequency**	**d^2/exp**	**P value**
rs41274768	G	A	GG N=563 -	AG N=0 -	AA N=0 -	G 1 N=1126	A 0 N=0	*	*
rs17047660	A	G	AA N=336 0.601	AG N=201 0.34842	GG N=26 0.0505	A 0.7753 N=873	G 0.2247 N=253	0.345	0.557
rs17047661	G	A	GG N=318 0.559	AG N=206 0.377	AA N=39 0.0636	G 0.7478 N=842	A 0.2522 N=284	0.506	0.48
rs6691117	G	A	GG N=495 0.866	AG N=64 0.129	AA N=4 0.0048	G 0.9306 N=1054	A 0.0694 N=72	0.219	0.639

4.3.4 The Effect of CR1 Variants on the Time to First Malarial Infection

Cox regression analysis was conducted in order to investigate the effect of several variables upon the time to first malarial infection in infants (Tori-Bossito cohort) (Table 4.23). The first malarial infection was defined by first parasite density after the malaria attack. The Tori-Bossito cohort was composed of 580 samples but as some of the data were missing for few samples they were not included in the analysis. Therefore, the total number of the samples used for the analysis was 494. In this cohort, rs41274768 was not polymorphic. We tested for association of the other three polymorphisms, and *CR1* copy number, with different parameters such as sex, maternal age, and use of mosquito net, sickle cell, ethnicity and malarial treatment as covariates. For rs17047661 only, a significant relationship with time to first infection was found.

Table 4.23: The Cox regression analysis 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**
Mother Age	-0.010	0.010	0.907	1	0.341	0.991	0.971	1.010
Malaria suspicion during pregnancy (No=0; Yes=1, reference)	0.032	0.132	0.058	1	0.809	1.032	0.797	1.338
Sex (M=1, reference; F=0)	0.008	0.112	0.005	1	0.945	1.008	0.808	1.256
Birth term	-0.002	0.029	0.005	1	0.943	0.998	0.944	1.055
Mosquito net (No=0; Yes=1, reference)	0.264	0.120	4.889	1	**0.027**	1.302	1.030	1.646
Ethnic group (Tori=0; Fon=1; other=2)			5.513	2	0.064			
Ethnic group (Tori vs other/Fon (reference))	0.350	0.164	4.569	1	**0.033**	1.418	1.029	1.954
Ethnic group (Fon vs Tori/other(reference))	0.475	0.229	4.290	1	**0.038**	1.608	1.026	2.521
Sickle cell (Yes=0; No=1, reference)	0.776	0.346	5.035	1	**0.025**	2.172	1.103	4.277
Chloroquine intake during pregnancy (Yes=0; No=1; Unknown=2)			1.528	2	0.466			
Chloroquine intake during pregnancy (Yes vs No/unknown(reference))	1.229	1.015	1.465	1	0.226	3.418	0.467	25.005
Chloroquine intake during pregnancy (No vs Yes/unkown(reference))	1.246	1.010	1.522	1	0.217	3.476	0.480	25.164
rs17047661 (G) (AA=0, AG=1 and GG=2)	-0.234	0.089	6.896	1	**0.009**	0.791	0.665	0.942

A protective effect of use of mosquito nets and presence of sickle cell anaemia were observed. For the mosquito net, absence of a mosquito net leads to an HR (hazard ratio) of 1.302, because having the mosquito net (Yes,1) is the reference. Presence of sickle cell leads to an HR of 2.172, because not having the sickle cell trait (No,1) is the reference (Table 4.23). Therefore, individuals with sickle cell or individuals using mosquito nets are protected against malaria infection. Ethnicity showed a protective effect as well. For ethnicity, having the ethnicity of either Tori (p=0.033) or Fon (p=0.038) leads to an HR of 1.418 and 1.608, respectively. In our results, it was found that the G allele of rs17047661 is protective against malaria. The Cox regression detected a statistically significant and protective effect of rs17047661 (p=0.009) with HR of 0.791 (95% CI: 0.665 to 0.942) (Table 4.23). The individuals with GG genotypes had a delay on the disease (malaria) compared to the individuals with AA and AG genotypes (Figure 4.34). There was no association between *CR1* copy number (OR= 0.941, 95% CI: 0.808 to 1.097, p=0.439) and malaria in Tori-Bossito cohort and neither rs17047660 (OR= 1.113, 95% CI: 0.920 to 1.347 and p=0.269) nor rs6691117 (OR= 0.931, 95% CI: 0.688 to 1.260 and p=0.643) showed association with malaria. The results are summarized below (Table 4.23).

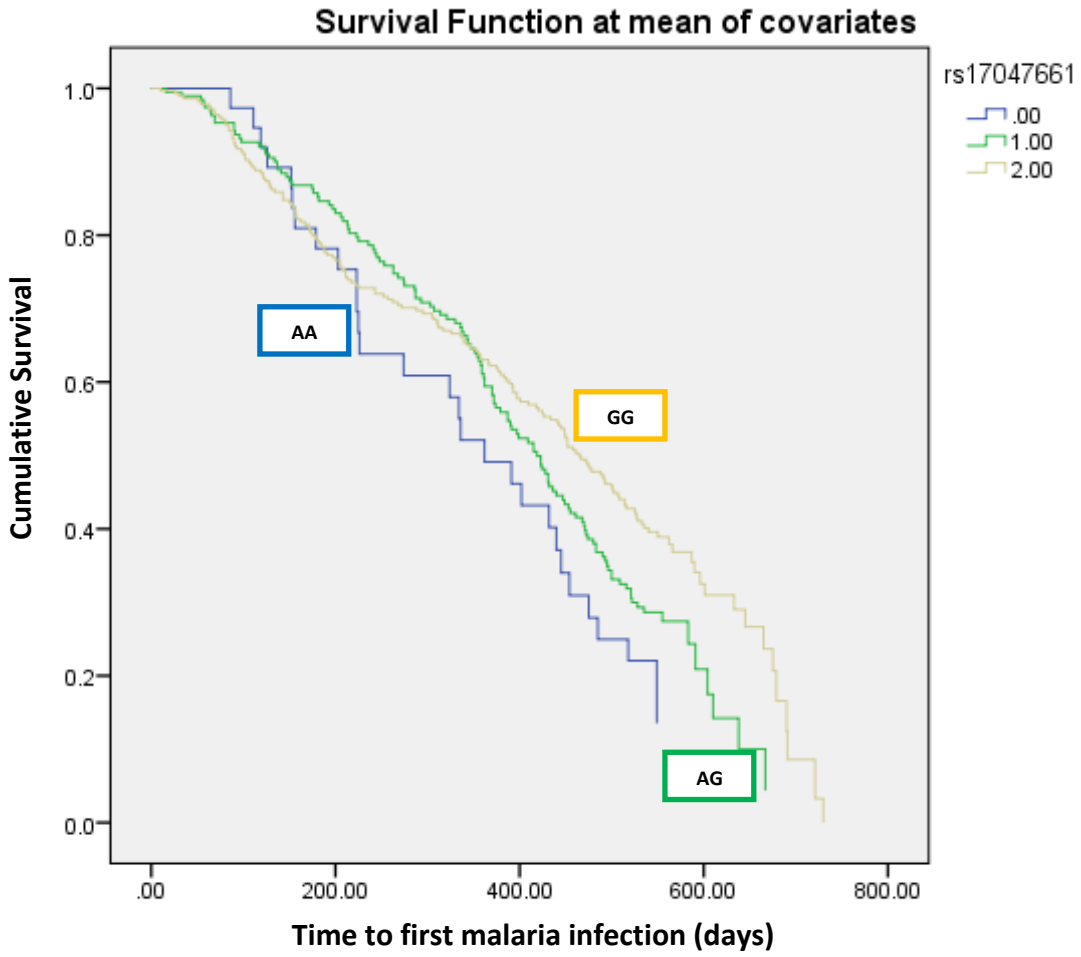


Figure 4.34: Survival curve from a Cox regression analysis in the Tori-Bossito cohort has revealed that the patients with G alleles (Sl2 allele) are protected against malaria and the patients with A alleles are more likely to develop infection. 2.00 is GG genotype, 1.00 is AG genotype and .00 is AA genotype for rs17047661.

4.3.5 Analysis of CR1 Variants and Number of Malarial Infections

An alternative outcome variable was used in the next set of association studies: 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 risk of malaria: mother's age, malaria suspicion during pregnancy, sex, birth term, mosquito net, ethnic group Tori, Chloroquine intakes during pregnancy, sickle cell, and genetic polymorphism. Four polymorphisms were analysed in turn: *CR1* copy number variation and three SNPs (SNPs: rs17047660, rs17047661 and rs6691117). Of the four polymorphisms, only rs17047661 showed a statistically significant result (Table 4.24).

Table 4.24: For Poisson generalized linear model the same covariates 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.

Parameter Estimates							
Parameter	**B**	**Std. Error**	**95% Wald Confidence Interval**		**Hypothesis Test**		
			Lower	**Upper**	**Wald Chi-Square**	**df**	**Sig.**
Mother Age	-0.013	0.0072	-0.027	0.001	3.351	1	0.067
Malaria suspicion during pregnancy (0) (No=0; Yes=1)	1.206	0.4685	0.288	2.125	6.630	1	**0.010**
Malaria suspicion during pregnancy (1, reference) (No=0; Yes=1)	0[a]						
Sex (0) (M=1, reference ; F=0)	0.094	0.0765	-0.056	0.244	1.523	1	0.217
Sex (1, reference)	0[a]						
Birth term	0.011	0.0192	-0.027	0.049	0.317	1	0.574
Mosquito net (0) (No=0; Yes=1, reference)	0.260	0.0818	0.100	0.420	10.105	1	**0.001**
Mosquito net (1, reference)	0[a]						
Ethnic group (Tori vs other/Fon (reference))	0.276	0.1156	0.049	0.502	5.688	1	**0.017**
Ethnic group (Fon vs Tori/other(reference))	0.317	0.1556	0.012	0.622	4.141	1	**0.042**
Ethnic group (Tori=0; Fon=1; other=2, reference)	0[a]						
Sickle cell (1) (Yes=0; No=1, reference)	0.923	0.2115	0.509	1.338	19.058	1	**≤0.0001**
Sickle cell (1, reference)	0[a]						
Chloroquine intake during pregnancy (Yes vs No/unknown (reference))	0.406	0.4710	-0.517	1.329	0.744	1	0.388
Chloroquine intake during pregnancy (No vs Yes/unkown (reference))	0.388	0.4692	-0.531	1.308	0.685	1	0.408
Chloroquine intake during pregnancy (Yes=0; No=1; Unknown=2, reference)	0[a]						
rs17047661 (G) (AA=0, AG=1 and GG=2)	-0.158	0.0595	-0.275	-0.042	7.082	1	**0.008**

The results show that the absence of a mosquito net will increase the number of infections by 1.297 whereas the absence of the sickle cell will increase the number of infections by 2.517 (Table 4.24). Interestingly, ethnic group and rs17047661 have a significant effect on the number of breakthrough infections in this cohort. The value for rs17047661 was -0.158 (95% CI: -0.275 to -0.042). Therefore, we expect that for every G allele, the number of breakthrough infections will reduce by 0.853. There was no significant correlation between *CR1* copy number (B= 0.002, 95% CI: -0.101 to 0.104, p=0.975) and the number of breakthrough infections in the Tori-Bossito cohort and either rs17047660 (B= 0.090, 95% CI: -0.039 to 0.220, p=0.172) or rs6691117 (B= -0.048, 95% CI: -0.250 to 0.155, p=0.644) showed no association with the number of breakthrough infections. The results are summarized below (Table 4.26).

4.3.6 Analysis of CR1 Variants and Parasite Density

The estimation of parasite density of malaria for malaria patients has a critical importance as parasite resistance to available therapy is increasing and that makes it harder to obtain promising results from clinical trials and drug studies (Gyasi et al., 2012). The genotyping data which were collected using the ASO assay, were used in the parasite density analysis after Cox regression and Poisson regression analysis (Table 4.25). Parasite density adjustment data (the data was adjusted for seasonal mosquito levels), which was used as dependent variable for this analysis, was organized by David Courtin.

Table 4.25: The parasite density analysis was used to assess the effect of the following factors on risk of malaria: mother's age, malaria suspicion during pregnancy, sex, birth term, mosquito net, ethnic group (Tori, Fon or others), Chloroquine intake during pregnancy, sickle cell, *CR1* copy number variation (duplication (*CR1-B*) and deletion (*CR1*-C)) and polymorphisms (SNPs: rs41274786 (not polymorphic), rs17047660, rs17047661 and rs6691117) that determine Knops blood group antigens. **Bold Results**=significant, **sig.**=significance and **B**=the coefficient for the constant.

Parameter	**B**	**Sig.**	**95% Confidence Interval**	
			Lower Bound	**Upper Bound**
Sex (0) (M=1; F=0)	0.452	0.685	-1.73	2.64
Sex (1)	0[b]			
Ethnic group (0) (Tori=0; Fon=1; other=2)	2.79	0.067	-0.198	5.77
Ethnic group (1)	6.26	**0.006**	1.82	10.7
Ethnic group (2)	0[b]			
Sickle cell (1) (Yes=0; No=1; Unknown=2)	-0.361	0.933	-8.75	8.02
Sickle cell (2)	0[b]			
Chloroquine intake during pregnancy (0) (Yes=0; No=1; Unknown=2)	4.90	0.491	-9.073	18.9
Chloroquine intake during pregnancy (1)	4.092	0.563	-9.77	17.96
Chloroquine intake during pregnancy (2)	0[b]			
Mosquito net (0) (No=0; Yes=1)	1.58	0.204	-0.854	4.005
Mosquito net (1)	0[b]			
Malaria suspicion during pregnancy (0) (No=0; Yes=1)	1.19	0.368	-1.40	3.780
Malaria suspicion during pregnancy (1)	0[b]			
Time after infection	-0.290	**0.009**	-0.51	-0.0718
Mother Age	-0.0142	0.890	-0.215	0.186
Birth Term	-0.225	0.428	-0.780	0.331
Time to infection	-0.0169	**≤.0001**	-0.0231	-0.011
rs17047661	-0.067	0.941	-1.841	1.71

Neither any of the other SNPs nor *CR1* copy number were associated with parasite load. The parameters which have a major effect on parasite load were time to infection (reflecting how old the child is when it catches malaria), time after infection and ethnicity. In summary, the Swain-Langley Sl2 allele (rs17047661-G) appears to provide protection against early acquisition of malaria and subsequent number of

malarial infections, but not parasite density following infection. There was no significant association between *CR1* copy number (B= -0.137, 95% CI: -1.33 to 1.60, p=0.855) and parasite density following infection in Tori-Bossito cohort and either duplication (B= 0.114, 95% CI: -1.79 to 2.013, p=0.907) or deletion (B= -2.11, 95% CI: -5.99 to 1.784, p=0.288) or rs17047660 (B= -0.447, 95% CI: -2.38 to 1.41, p=0.651) or rs6691117 (B= 0.893, 95% CI: -2.11 to 3.90, p=0.560) showed no association with parasite density following infection (Table 4.26).

Table 4.26: The summarized data for Knops blood group determining SNPs and *CR1* CNV for three analyses. P is the p value and B is the size effect.

Polymorphism /CNV	**Time to Infection**	**Number of Infections**	**Parasite Density**
rs17047660	P=0.269 B=0.107	P=0.172 B=0.090	P= 0.651 B=-0.447
rs17047661	P=0.009 B=-0.234	P= 0.008 B=-0.158	P=0.941 B=-0.067
rs6691117	P= 0.643 B=-0.072	P= 0.644 B=-0.048	P=0.560 B=0.893
***CR1* CN**	P= 0.439 B=-0.060	P= 0.975 B=0.002	P= 0.855 B=-0.137

4.3.7 Linkage Disequilibrium Analysis for rs17047661 in YRI

In this project, I measured pairwise linkage disequilibrium (LD) of rs17047661 with all other SNPs that had an allele frequency >5% in the Yoruba population from Nigeria (YRI, 1000 Genomes Project) across the *CR1* gene. The YRI population was used for the analysis. It may have a simillar patterns of LD with Tori-Bossito cohort. These populations are geographically close to each other, and if we assume similar population histories then they might be expected to have similar patterns of LD; notably, they also both speak languages belonging to the Yoruba family (https://www.ethnologue.com/language/yor). The Sl2 allele of Swain-Langley was found to be protective for malaria. It is important to understand which SNPs are in LD with it as if the SL2 allele is in LD with another SNP, then the protective effect of the

allele could be because it is in LD with another variant and on the same haplotype which is actually the protective one. The results clearly show that rs17047661 is not in strong LD (r2>0.8) with any other common alleles across the *CR1* gene (Figure 4.35).

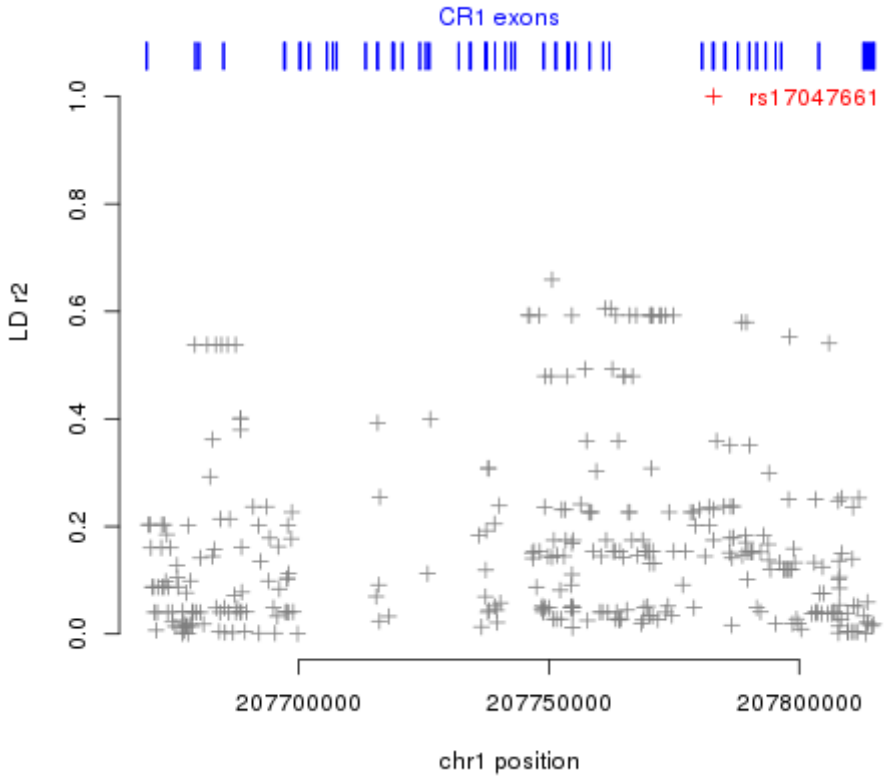


Figure 4.35: Pairwise LD with rs17047661 in YRI population: Linkage disequilibrium analysis for rs17047661 revealed that rs17047661 is not in strong LD (r^2>0.8) with any other common alleles across the *CR1* gene in YRI. The blue vertical lines indicate to the *CR1* exons all accross the C*R1* gene and grey crosses stand for all different SNPs accross the *CR1* gene. The red cross shows the location of rs17047661 in *CR1* by coordinate (and also exon position) on the x-axis whereas on the y-axis it shows the LD value with other SNPs. The population data were collected by Dr Edward Hollox.

4.3.8 LD Analysis of Knops Blood Group SNPs and CR1 CNV

In order to understand if Knops blood group antigen-defining SNPs are in LD with *CR1-B* and *CR1-C*, I assessed LD analysis for Tori-Bossito cohort (Table 4.27). According to the result, rs17047660, rs17047661 and rs6691117 were in low LD with both *CR1-B* and *CR1-C*. As rs41274768 was not polymorphic, no results were obtained for this SNP. Therefore, if Sl2 is a protective allele against malaria, it may be in LD with another variant that is providing protection, but not the *CR1* CNV.

Table 4.27: Linkage disequilibrium analysis in CEU, JPT/CHB, YRI and the Tori-Bossito cohort. The GWAS studies related to the SNPs are (1) Kullo et al., 2011 and (2) Naitza et al*., 2012*. *=not polymorphic. Haplo=Haplotype and Freq=Frequency; the samples which were homozygous for the *CR1*-B allele are coded as AA (variant could be T/A) whereas homozygous for the *CR1*-C allele are coded as GG (variant could be C/G) during LD analysis.

Variant	**Disease and/or Trait**	**Risk allele**	**LD with *CR1-B* or *CR1-C***	**CEU**		**JPT/CHB**		**YRI**		**Tori-Bossito**	
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**	**R^2**	**D'**
rs12034383 (A/G)	Erythrocyte sedimentation rate (ESR) **(1)**	G	*CR1-B* (A) (A/T)	0.337	0.856	0.022	1.000	0.010	0.473		
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**		
				AG	0.237	AA	≤0.0001	AA	0.016		
				TG	0.194	TA	0.483	TA	0.122		
				AA	0.021	AG	0.023	AG	0.208		
				TA	0.548	TG	0.494	TG	0.654		
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**		
			CR1-C (G) (G/C)	0.034	1.000	0.011	1.000	0.022	0.184		
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**		
				GG	≤0.0001	GA	≤0.0001	GA	0.028		
				CG	0.431	CA	0.483	CA	0.110		
				GA	0.043	GG	0.011	GG	0.067		
				CA	0.526	CG	0.506	CG	0.795		
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**		
rs12034598 (A/G)	ESR **(2)** (inflammation)	A	*CR1-B* (A) (A/T)	0.024	0.630	0.006	0.998	0.050	1.000		
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**		
				AG	0.014	AG	≤0.0001	AG	≤0.0001		
				TG	0.133	TG	0.213	TG	0.144		
				AA	0.245	AA	0.022	AA	0.229		
				TA	0.609	TA	0.764	TA	0.627		
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**		

			CR1-C (G) (G/C)	0.008	1.000	0.042	1.000	0.017	0.983		
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**		
				GG	≤0.0001	GG	0.011	GG	0.013		
				CG	0.147	CG	0.202	CG	0.128		
				GA	0.043	GA	≤0.0001	GA	0.080		
				CA	0.810	CA	0.787	CA	0.778		
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**		
rs7527798 (C/T)	ESR **(1)**	not given	*CR1-B* (A) (A/T)	0.059	0.709	0.001	1.000	0.002	0.141		
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**		
				AC	0.019	AC	≤0.0001	AC	0.009		
				TC	0.231	TC	0.034	TC	0.017		
				AT	0.240	AT	0.023	AT	0.216		
				TT	0.510	TT	0.943	TT	0.759		
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**		
			CR1-C (G) (G/C)	0.006	0.215	≤0.0001	1.000	0.008	0.182		
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**		
				GC	0.018	GC	≤0.0001	GC	0.007		
				CC	0.232	CC	0.034	CC	0.019		
				GT	0.025	GT	0.011	GT	0.088		
				CT	0.725	CT	0.955	CT	0.886		
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**	**R^2**	**D'**
rs41274768	Kn^a /Kn^b		*CR1-B* (A) (A/T)	0.046	1.000	*	*	*	*	*	*
				Haplo	**Freq**						
				AA	0.017						
				TA	≤0.0001						
				AG	0.259						
				TG	0.724						

				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**	**R^2**	**D'**
			CR1-C (G) (G/C)	*	*	*	*	*	*	*	*
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**	**R^2**	**D'**
rs17047661	Sl1/Sl2		*CR1-B* (A) (A/T)	*	*	*	*	0.085	0.807	0.033	0.636
								Haplo	**Freq**	**Haplo**	**Freq**
								AA	0.013	AA	0.018
								TA	0.297	GA	0.175
								AG	0.211	AT	0.235
								TG	0.479	GT	0.572
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**	**R^2**	**D'**
			CR1-C (G) (G/C)	*	*	*	*	0.163	0.836	0.051	0.675
								Haplo	**Freq**	**Haplo**	**Freq**
								GA	0.084	AG	0.028
								CA	0.226	GG	0.009
								GG	0.011	AC	0.225
								CG	0.679	GC	0.738
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**	**R^2**	**D'**
rs6691117	KAM+/- and lower ESR (G allele) **(1)**		*CR1-B* (A) (A/T)	0.033	0.644	0.005	1.000	0.001	0.188	≤0.0001	0.031
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**
				AG	0.017	AG	≤0.0001	AA	0.016	AA	0.014
				TG	0.167	TC	0.223	TA	0.069	GA	0.179
				AA	0.246	AA	0.018	AG	0.213	AT	0.050
				TA	0.570	TA	0.759	TG	0.702	GT	0.757
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**	**R^2**	**D'**
			CR1-C (G) (G/C)	0.010	1.000	0.043	1.000	0.031	0.186	0.002	0.952
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**
				GG	≤0.0001	GG	0.012	GA	0.022	AG	≤0.0001

				CG	0.184	CG	0.211	CA	0.063	GG	0.036
				GA	0.044	GA	≤0.0001	GG	0.071	AG	0.064
				CA	0.772	CA	0.777	CG	0.844	GC	0.899
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**	**R^2**	**D'**
rs17047660	McCoy[a]/ McCoy[b]		*CR1-B* (A) (A/T)	*	*	*	*	0.101	1.000	0.037	0.731
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**
								AG	≤0.0001	GA	0.012
								TG	0.254	AA	0.181
								AA	0.229	GT	0.212
								TA	0.517	AT	0.595
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**	**R^2**	**D'**
			CR1-C (G) (G/C)	*	*	*	*	0.035	0.997	0.011	1.000
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**
								GG	≤0.0001	GG	≤0.0001
								CG	0.254	AG	0.037
								GA	0.093	GC	0.224
								CA	0.653	AC	0.740

4.3.9 LD Analysis of Erythrocyte Sedimentation Rate-associated SNPs and CR1 CNV

Erythrocyte sedimentation rate is a test that measures the degree of inflammation existing in the body by measuring the rate of sedimentation of erythrocytes in a sample of blood. Extreme elevation of ESR was observed in malaria-infected patients (*P.falciparum*) (Ifeanyichukwu et al., 2015). Because of this relation to malaria, ESR level measurement can be critical for malaria-infected patients. For example, ESR can be considered as complementary marker of malaria infection in cases of unsuccessful microscopic detection of the malaria parasite in blood smears (Boampong et al., 2010). To understand if GWAS sentinel SNPs (rs12034383, rs12034598 and rs7527798), within the 250-kb region of the *CR1* gene (Figure 4.36), and Knops blood group antigen-defining SNPs (rs41274768, rs17047660, rs17047661 and rs6691117) are in LD with *CR1-B* and *CR1-C*, I undertook LD analysis in the HapMap samples (Table 4.27). According to the result, rs41274768, rs17047660, rs17047661 and rs6691117 and the GWAS sentinel SNPs were in low LD with *CR1-B* and *CR1-C*.

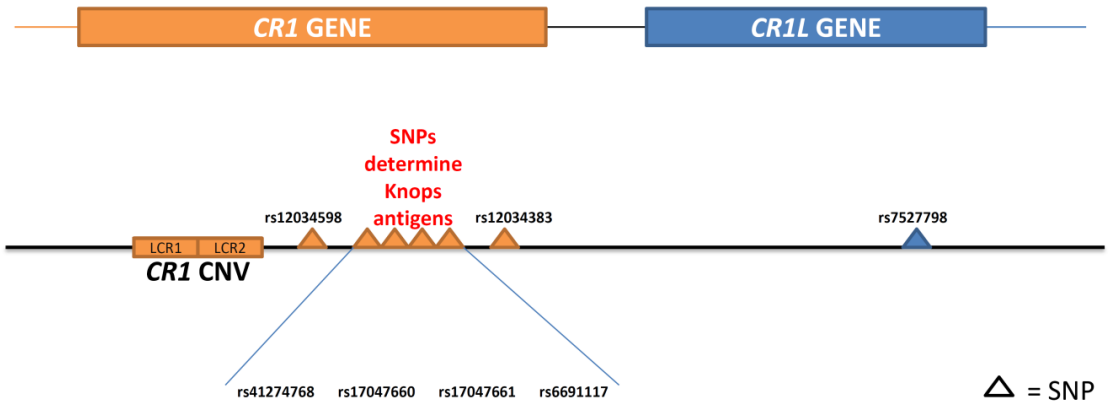


Figure 4.36: The locations of GWAS SNPs within and around *CR1* are shown with Knops blood group-defining SNPs. Only rs7527798 is located in *CR1L;* the rest of the SNPs are located in *CR1*.

The results show that LD of SNPs shown in GWAS studies with ESR is not due to *CR1-B*/*CR1*-C. These SNPs may be in LD with another variant related with ESR and malaria but not related to *CR1-B*/*CR1*-C. Knops blood group-defining SNPs were not in LD with *CR1-B*/*CR1*-C either, meaning that any signal of malaria protection from the Knops blood group-defining SNPs is not due to *CR1*-C or *CR1*-B. The results have conclusively shown that the Sl2 allele of Swain-Langley is not in LD with the CNV (Table 4.27) nor

with the SNPs (Figure 4.35), therefore it is very likely to be causing the protective effect against malaria.

4.4 Tolimmunpal Cohort

The aim of the Tolimmunpal project was to study the environmental, biological and genetic factors involved in the development of immune tolerance associated with malaria for the protection of pregnant women and young children. This project was conducted in southern Benin, in the municipality of Allada, an area in the Atlantic department. The town of Allada is bounded to the north with the community of Toffo and the municipality with Tori-Bossito in the south. Tolimmunpal (follow-up of 400 children from birth to 24 months) is a continuation of the Malaria in Pregnancy Preventive Alternative Drugs project (follow-up of 1179 women during their pregnancies) and made it possible to obtain data based on malaria infection during pregnancy (contact of the foetus with soluble *P. falciparum* antigens) of mother and also children during the first 24 months of their lives. Our collaborators provided us with 278 DNA samples of the Tolimmunpal cohort.

4.4.1 Cox Regression Analysis for the Tolimmunpal Cohort

The Tolimmunpal cohort that was used in this project was composed of 278 infants. The SNP data of the cohort were provided by David Courtin. Two individuals had no SNP information, so the samples which were analysed were 276 in total. After analysis of the Tori-Bossito cohort, the study was repeated with a different and smaller cohort from the same region (Benin) to confirm the recent findings from Tori-Bossito. The main reason to repeat the cohort with another was for replication with a different cohort, especially for rs17047661 (Swain-Langley). However, the Tolimmunpal cohort did not have the same databased variables as the Tori-Bossito so that some of the covariates which were used in the previous analysis could not be used in the Tolimmunpal analysis. For example, the Tollimmunpal cohort did not have any information about chloroquine intake during the mothers' pregnancies which could easily affect the association studies, producing misleading data. The covariates which existed in the Tori-Bossito cohort but were missing for the Tolimmunpal cohort were mother's age, malaria suspicion during pregnancy, birth term, mosquito net, ethnic group, sickle cell, LCR *CR1* copy number and chloroquine intakes during pregnancy. The

genotyping data for three SNPs (rs41274786, rs17047660 and rs6691117) were missing for the Tolimmunpal cohort as well.

Table 4.28: The Cox regression analysis for the Tolimmunpal cohort, with days until disease. The Cox proportional hazards model was used to assess the effect of the following factors on risk of malaria: mother infection (malaria 1=infected; 0=no infection), sex (1=male; 0=female) and polymorphism (SNP: rs17047661) determining Swain-Langley Knops blood group antigens. The dependent variable was disease progress period days in the analysis. There was no significant protective effect of any of the covariates. Dependent: Time to first malaria infection (days). **df**= degree of freedom, **sig.**=significance, **CI**=the confidence interval, **Exp(B)**=the exponentiation of the B coefficient, **SE**=standard errors associated with 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
Mother infection (0=No and 1=Yes)	-0.263	0.147	3.21	1	0.073	0.769	0.577	1.025
Sex (Male=1 and Female=0)	-0.143	0.136	1.11	1	0.292	0.867	0.664	1.131
Genotype (rs17047661: AA=0; AG=1 and GG=2)	-0.029	0.105	0.077	1	0.781	0.971	0.791	1.193

According to the results, none of the covariates showed a protective effect. The protective effect of the G allele (rs17047661) was not observed in the Tolimmunpal cohort (Table 4.28). The survival curve showed the results more clearly (Figure 4.37).

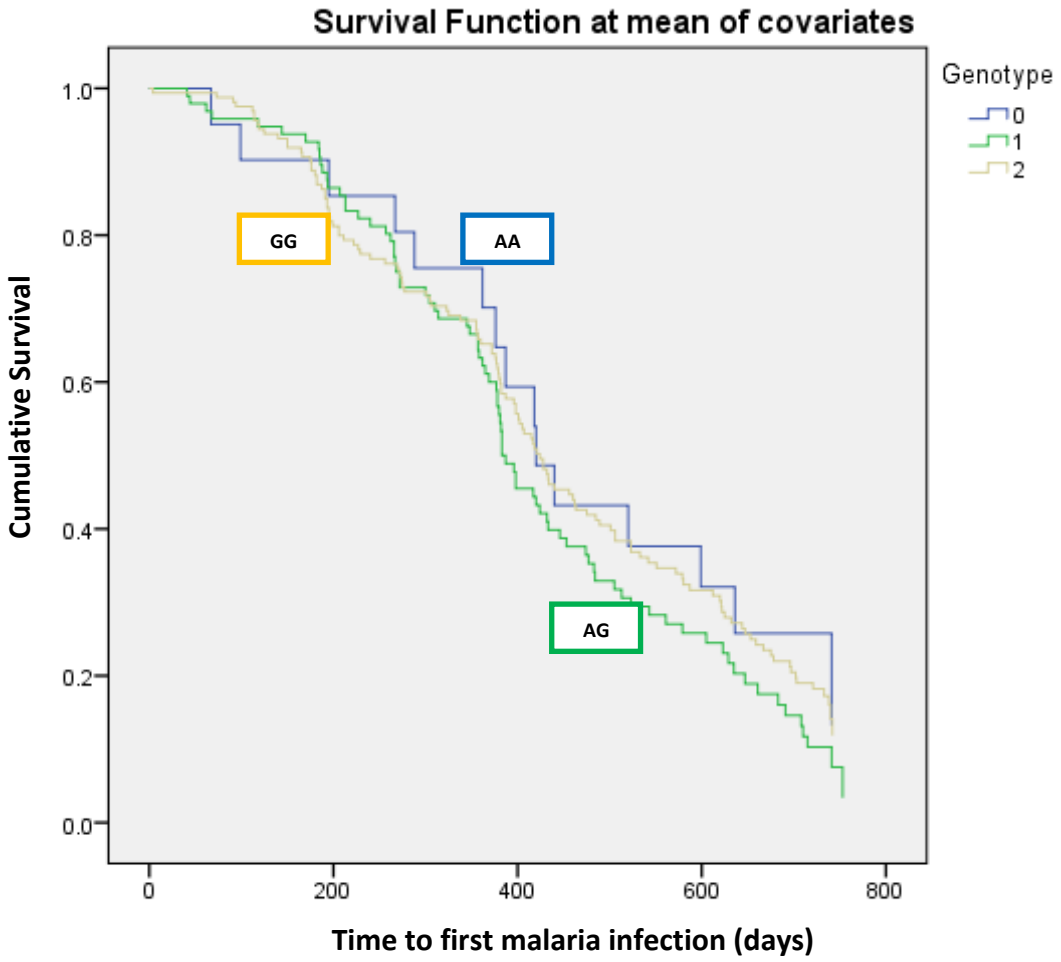


Figure 4.37: Survival curve from a Cox regression analysis in the Tolimmunpal cohort showed that the patients with the G allele (Sl2 allele) are not protected against malaria as they had the malaria infection at same time period as the patients with the A allele. 2.00 is GG genotype, 1.00 is AG genotype and .00 is AA genotype for rs17047661. Data were analysed using SPSS.

4.4.2 Poisson Regression Analysis for the Tolimmunpal Cohort

The same covariates used in Poisson regression analysis were used for Cox regression analysis. The results revealed that rs17047661, mother infection and sex (infant) did not have a significant effect on the number of breakthrough infections in this cohort (Table 4.29). Therefore, there was no significant association between rs17077661 (G allele) and the number of breakthrough infections in this cohort.

Table 4.29: The Poisson regression analysis for the Tolimmunpal cohort. The same covariates used in Poisson regression analysis were used for Cox regression analysis with the difference that the dependent variable was number of malaria infections. **df**= degree of freedom, **sig**.=significance, **Std.Error**=standard error and **B**=the coefficient for the constant.

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
Mother infection Mother infection=0 (0=No and 1=Yes)	0.016	0.0691	-0.120	0.151	0.051	1	0.822
Mother infection Mother infection=1 (0=No and 1=Yes)	0[a]						
Sex Sex=0 (Male=1 and Female=0)	-0.035	0.0623	-0.157	0.087	0.312	1	0.576
Sex Sex=1 (Male=1 and Female=0)	0[a]						
Genotype (rs17047661: AA=0; AG=1 and GG=2)	-0.087	0.0485	-0.182	0.008	3.22	1	0.073

Dependent Variable: Number of malaria infections

4.5 Geographical Distribution of the Sl2 allele and Malaria Prevalence

The association of rs17047661-G allele (Sl2 allele) with protection against malaria made it possible to speculate whether there was an association between the frequency of this allele and malaria prevalence across Africa. If correct, this would support a role for natural selection increasing the frequency of the Sl2 allele in malarial regions by protecting against malaria. I collated our own and published data for 36 populations across sub-Saharan Africa (Table 4.30). Population SNP data for the malaria prevalence analysis were obtained from different sources: (1) online available SNP data from the 1000 Genomes Project, (2) study of Gurdasani et al., 2015 (AGVD), (3) study of Patin et al., 2014 (European Variation Archive (EMBL-EBI, access number: EGAD00010000496)), (4) Tori-Bossito data from Le port et al., 2012 (SNPs were typed by our ASO assay), (5) Tolimmunpal SNP data provided by David Courtin, (6) HGDP's

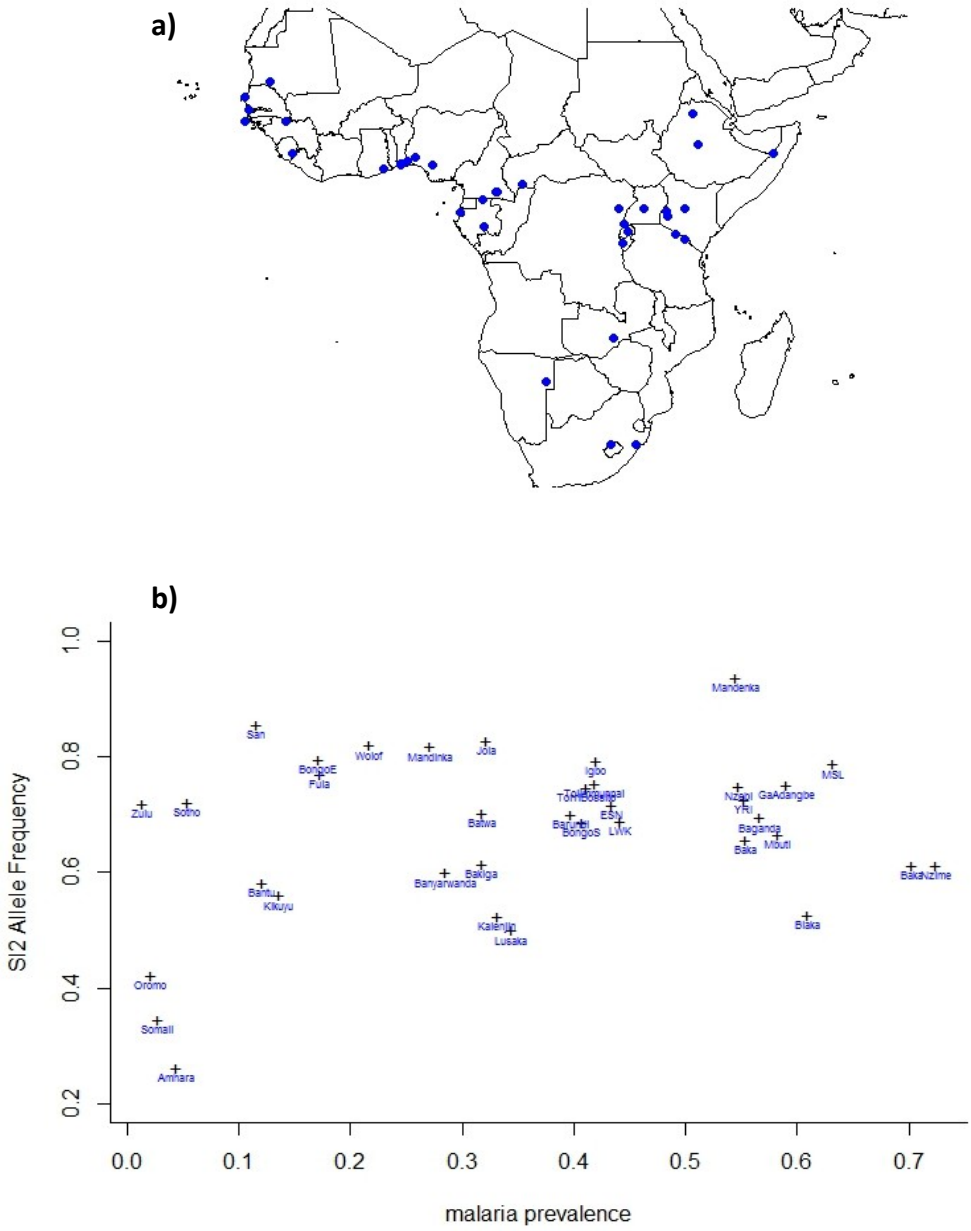
(African(CEPH-HGDP)) and (7) Zambia's SNP data (25 samples) were obtained using Sanger sequencing (ASO primers) with the help of Daniel A. Pritchard (BSc student).

In addition, malarial frequency against Sl2 allele frequency was plotted for each population (Figure 4.38). Malaria prevalence was estimated by Dr Rita Rasteiro in the year 2000 from Malaria Atlas Project data for each sampling point, taking an average of a 110-km radius circle around that sampling point. We found no association of Sl2 allele frequency and malaria prevalence ($p>0.05$) using both simple linear regression and a regression model that allowed for spatial relatedness between populations (sPAMM package in R, Rousset and Ferdy 2014). sPAMM analysis was done by Dr Edward Hollox.

Table 4.30: The samples which were used in malaria prevalence analysis for populations are shown below. The major allele for all of the populations was G (rs17047661) except for three populations, which were Amhara, Somali and Oromo. The minor allele frequencies and allele frequency for G allele for each population were shown below as well. 1000G refers to 1000 Genomes Project data whereas AGVP refers to The African Genome Variation Project. There were 7 sources ((1) 1000 Genomes Project data, (2) AGVD, (3) EMBL-EBI (Hunter), (4) Tori-Bositto cohort (Courtin), (5) Tolimmunpal cohort (Courtin), (6) HGDP (African) and (7) Zambia. The sample size of each population and geographic location information are shown below.

Samples	Population	A1	A2	MAF (minor)	Allele freq (G)	Sample Size	Source of Sampling site	Geographical Coordinates		Malaria Prevalence
								Longitude	Latitude	
1	ESN Esan Nigeria	A	G	0.28	0.72	99	1000G- capital, Edo state	6	6.5	0.434
1	MSL Mende Sierra Leone	A	G	0.21	0.79	85	1000G - Kenema	-11.2	7.9	0.633
2	Baganda (Uganda)	A	G	0.30	0.70	100	Murdock, 1967	32	1	0.567
2	Banyarwanda (Republic of Congo)	A	G	0.4	0.6	100	location of Rwanda capital	30.06	-1.94	0.285
2	Barundi (Burindi- east and central Africa)	A	G	0.3	0.7	97	location of Burundi capital	29.4	-3.4	0.398
2	Amhara	G	A	0.262	0.262	42	Murdock, 1967	38	13	0.0444
2	Somali	G	A	0.346	0.346	39	Murdock, 1967	48	8	0.0275
2	Oromo	G	A	0.423	0.423	26	Location of capital Addis Ababa	38.74	9.03	0.0213
2	Fula (West Africa)	A	G	0.23	0.77	74	Murdock, 1967	-14	17	0.173
2	Ga-Adangbe (Ghana-Togo)	A	G	0.248	0.752	99	Murdock, 1967	0	6	0.590
2	Igbo (Nigeria)	A	G	0.207	0.793	99	Murdock, 1967	3	7	0.42
2	Jola	A	G	0.171	0.829	79	Murdock, 1967	-17	12	0.322
2	Kalenjin	A	G	0.475	0.525	100	Murdock, 1967 (Nandi)	35	0	0.333
2	Kikuyu	A	G	0.439	0.561	99	Murdock, 1967	37	1	0.137
2	LWK Luye Webuye Kenya	A	G	0.311	0.689	74	1000G	34.8	0.6	0.441
2	Mandinka	A	G	0.182	0.818	88	Location of capital	-16.6	13.4	0.272
2	Sotho	A	G	0.279	0.721	86	Murdock, 1967	28	-29	0.0548
2	Wolof	A	G	0.18	0.82	78	Murdock, 1967	-17	15	0.218
2	YRI Yoruba Ibadan	A	G	0.2727	0.7273	99	1000G	3.9	7.4	0.553
2	Zulu	A	G	0.28	0.72	100	Murdock, 1967	31	-29	0.0138
3	Baka_Gabon	A	G	0.344	0.656	162	Patin et al., 2014	12.13	2.13	0.554
3	Bakiga	A	G	0.386	0.614	35	Patin et al., 2014	29.62	-0.98	0.319

3	Nzebi	A	G	0.25	0.75	20	Patin et al., 2014	9.45	0.4	0.548
3	Baka_Cameroon	A	G	0.388	0.612	58	Patin et al., 2014	14.07	3.08	0.703
3	Nzime	A	G	0.387	0.613	53	Patin et al., 2014	13.83	3.12	0.724
3	Batwa	A	G	0.297	0.703	32	Patin et al., 2014	29.62	-0.98	0.319
3	Bongo S	A	G	0.312	0.688	24	Patin et al., 2014	12.43	-1.34	0.408
3	Bongo E	A	G	0.204	0.796	22	Patin et al., 2014	36	-2.27	0.172
4	Torri Bossito cohort	A	G	0.2522	0.7478	563	Torri Bossito	2.1	6.5	0.411
5	Tolimmunpal cohort	A	G	0.246	0.754	276	Allada	2.15	6.65	0.419
6	Bantu_N.E.	A	G	0.417	0.583	24	CEPH-HGDP	37	-3	0.121
6	Biaka_Pygmy	A	G	0.472	0.528	72	CEPH-HGDP	17	4	0.61
6	Mandenka	A	G	0.06	0.94	48	CEPH-HGDP	-12	12	0.545
6	Mozabite	G	A	0.1552	0.8448	58	CEPH-HGDP	3	32	0
6	Mbuti_Pygmy	A	G	0.333	0.667	30	CEPH-HGDP	29	1	0.584
6	San	A	G	0.143	0.857	14	CEPH-HGDP	20	-21	0.117
7	Lusaka	A	G	0.5	0.5	25	Zambia	28.283	-15.417	0.345



(The estimated proportion of 2-10 years olds in the general population that are infected with *P.falciparum* at any one time averaged over the 12 months of 2010)

Figure 4.38: Results of the malaria prevalence data for sub-Saharan African samples. (a) The locations of populations. (b) The results of malaria prevalence data for each population revealed that Sl2 allele frequency was not in association with malaria prevalence ($p>0.05$) using both simple linear regression and a regression model that allowed for spatial relatedness between populations.

Therefore, we have found no evidence of association of Sl2 allele frequency with malaria prevalence across Africa.

In a recent study by Bhatt et al. (2015) it was shown that with the help of malaria interventions, *P.falciparum* infection prevalence in endemic Africa was halved and the incidence of clinical disease fell by 40% between 2000 and 2015. Based on the data collected by Bhatt et al. (2015) in 2000, Dr. Edward Hollox generated a malaria prevalence map (Figure 4.39). The map shows the *Plasmodium falciparum* parasite rate (PfPR) for the year 2000 predicted at 5x5-km resolution. As in the study, the map used 2-10 year-old children as this age range is associated with a plateau in age prevalence relationship and can therefore act as a standardised comparison.

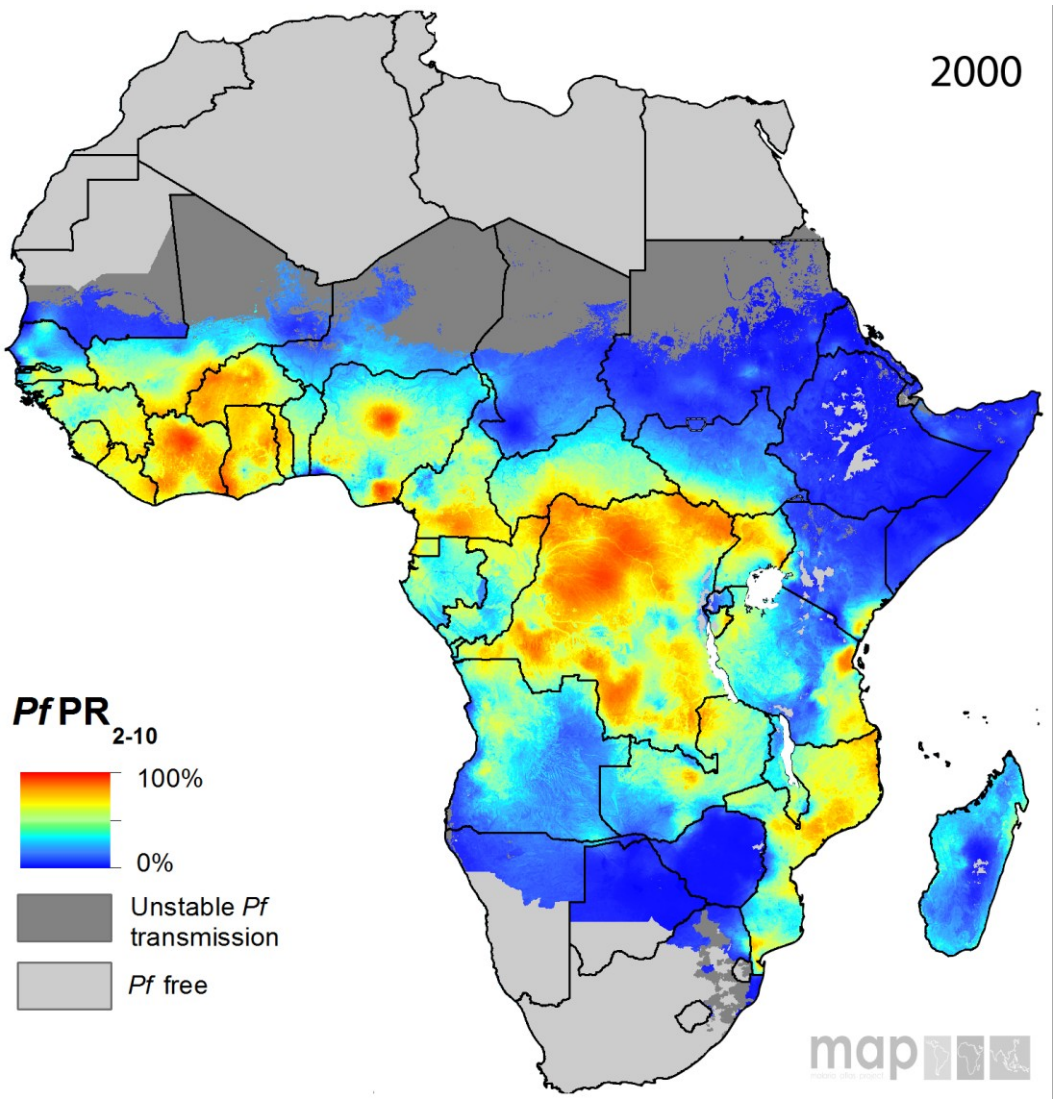


Figure 4.39: *P. falciparum* infection prevalence in endemic Africa, 2000. PfPR for 2–10-year old children for the year 2000 predicted at 5x5-km resolution. The white region indicates water.

4.6 Pathogen Diversity Index

The pathogen diversity index for *CR1* was provided by Dr Manuela Sironi (Scientific Institute IRCCS E. Medea, Bioinformatic Laboratory, Italy), based on pathogen richness data evaluated by gathering information on the number of different pathogen species/ genera present in different geographical areas of the world corresponding to countries (matched with HGDP populations) from the Gideon database. In published research, the micro-pathogens group was one of the major groups of the pathogens and includes viruses, bacteria, fungi, and protozoa. The second major group was macro-pathogens referring to insects, arthropods, and helminths (Fumagalli et al., 2009). The correlation analyses were adapted according to the average copy number of *CR1* (LCR), duplication frequency of *CR1* (*CR1-B*) (because of relation to malaria), deletion

frequency of *CR1* (*CR1*-C) and sample size. The result of the correlation analysis showed there is only a correlation between duplication frequency (*CR1-B*) and bacterial diversity (Table 4.31). There was no correlation between protozoa (including malaria) diversity and duplication frequency.

Table 4.31: The results of partial Mantel analysis with the pairwise distance matrix (Partial Mantel Test done with the R package) by using geographic distance between populations (HGDP-CEPH) (Pozzoli et al., 2011). The regression analysis showed that only bacterial diversity showed significant correlation with duplication frequency. This method accounts for geographic/genetic distances between populations.

Average CN (LCR *CR1*)	***CR1*-B Frequency**	***CR1*-C Frequency**
Protozoa		
Mantel statistic R: -0.0173	Mantel statistic R: -0.0845	Mantel statistic R: -0.000584
Significance: 0.552	Significance: 0.916	Significance: 0.443
Bacteria		
Mantel statistic R: 0.166	Mantel statistic R: 0.177	Mantel statistic R: -0.0228
Significance: 0.0012	Significance: 0.0013*	Significance: 0.649
Viruses		
Mantel statistic R: 0.112	Mantel statistic R: 0.158	Mantel statistic R: -0.0279
Significance: 0.0679	Significance: 0.033	Significance: 0.601
Helminths		
Mantel statistic R: 0.0276	Mantel statistic R: 0.0901	Mantel statistic R: -0.010
Significance: 0.154	Significance: 0.011	Significance: 0.595

According to the above results, populations that experience a high diversity of bacterial pathogens show a high frequency of *CR1*-B.

4.7 Discussion

In recent studies, it has been suggested that *CR1* plays a critical role in the pathogenesis of *P. falciparum* malaria (Stoute, 2011). Although several SNPs have been tested in malaria cohorts, including those involved in the Knops blood group and a SNP affecting CR1 expression levels on the erythrocyte surface, there were no studies looking at *CR1* CNV. For this reason, I have tested for an association between CNV of *CR1* and malaria. We have also tested the role of Knops blood group antigen SNPs in malaria. No association was found between malaria and *CR1* CNV in the Tori-Bossito

cohort. However, we found that the Sl2 allele (rs17047661-G) appears to provide protection against early acquisition of malaria and subsequent number of malarial infections. The Tori-Bossito study was repeated with another population (Tolimmunpal cohort) in order to confirm the Sl2 allele association with malaria. In this cohort, the Sl2 allele does not provide protection against early acquisition of malaria and subsequent number of malarial infections. The Tolimmunpal cohort is smaller and does not have as many covariates as the Tori-Bossito study. For example, the Tolimmunpal cohort did not have any information about chloroquine intake during mother's pregnancy which could easily affect the association studies. If the mother had placental malaria and had taken chloroquine, it would mask the real effect of the genetic variants. Therefore, more data are needed to say the individuals with GG genotype have a better probability of surviving during disease progression. Alternatively, it is possible that there is not just one protective trait against malaria; humans may have many different ways to deal with the infection. Therefore, the mechanisms by which one population fights against malaria may not be necessarily be the same as those of a different population (Faik et al., 2009). The analysis could potentially be conducted in a much larger cohort in order to discover the slighter effect of *CR1* CNV.

An early study suggested that low CR1 expression and the *CR1* polymorphism (Sl(a^-) or Sl2)) are associated with reduced rosetting of erythrocytes (Rowe et al., 1997). This then raised the question of the Sl2 allele's relation to malaria. Another study worked on four variant versions of CR1 among global populations, focusing on CPPs 15-25 (capture the Knops blood group polymorphisms in CPPs (consensus or complement control protein/SCR) 24-25 and the key region for rosetting in CPPs 15-17) of CR1s ectodomains (Tetteh-Quarcoo et al., 2012). Between these *CR1* variants, no difference in complement regulator function was found nor any interactions with *P.falciparum* proteins. No difference was observed across the four *CR1* variants for inhibition of erythrocyte invasion by *P. falciparum* or for rosette disruption. Another study produced similar results to ours by showing that children with the Sl2/2 genotype were less likely to have cerebral malaria than individuals with Sl1/1 in the Kenyan population (Thathy et al., 2005). Particularly, they showed that the combination of genotypes such

as individuals with the Sl2/2McCoy$^{a/b}$ genotype were less likely to have cerebral malaria than individuals with Sl1/2McCoy$^{a/a}$ genotype. However, they could not find any association between the African alleles (Sl2 and McC^{b}) and severe malaria-associated anaemia. This study was not replicated in the Gambian and Southern Ghanaian populations with cerebral malaria cases, severe malaria cases and non-malaria controls (Zimmerman et al., 2003; Hansson et al., 2013). There are appreciable variations between the three studies such as cohort sizes and age differences which could cause conflicting results between them. Therefore, it is possible that the Sl2 allele is in LD with a protective variant in the Kenyan but not Gambian or Ghanaian individuals and also that Sl2/2 and McCoy$^{a/b}$ evolved in relation to malaria transmission, and possibly that certain combinations of these alleles provides a survival advantage to these Kenyan populations. On the other hand, the difference between Swain-Langley Knops blood group antigens (Sl1 (A) and Sl2 (G)) causes a non-conservative amino acid substitution (Arginine and Glycine, respectively) as mentioned in as mentioned Table 1.2. This may lead to an effect on the response to malarial infections. An association study showed that two Knops blood group antigens (Kn^{b} and KAM^{+}) confer susceptibility to *P. falciparum* in the Brazilian Amazon population (Fontes et al., 2011). On the other hand, Gandhi et al. (2009) suggested that the process of rosetting occurs independently of the effect of Knops polymorphisms (rs41274768, rs17047660 and rs17047661) and may be controlled by other *CR1* polymorphisms. All these studies are summarized below (Table 4.32).

Table 4.32: Recent studies conducted in order to interpret the role of Knops blood polymorphisms in malaria.

	Thathy et al.,2005		**Zimmerman et al., 2003**	**Hansson et al., 2013**	**Gandhi et al., 2009**	**Fontes et al., 2011**	**Tori-Bossito**	**Tolimmunpal**
Country	Kenya		Gambia	Ghana	India	Brazil (Amazon)	Benin	Benin
	Nyanza	Kisii						
Sample size	**Severe malaria** Case=137 Cont=137 (ages 14 months)	**Cerebral malaria** Case=23 Cont=23 (ages 29-30 months)	**Cerebral malaria** Case=331 (ages avg. 3.78 years) **Severe malaria** Case=152 (ages avg. 2.19 years) **Non-malaria** cont=390 (ages avg. 3.43 years)	**Uncomplicated Malaria** Case=89 **Severe Anaemia** Case=57 **Cerebral Malaria** Case=121 Cont=275 (ages btw 0-15 years)	**Uncomplicated Malaria (*P.falciparum*)** Case=100 Cont=100	**Malaria (*P.falciparum* and *P.vivax*)** Case=92 (P. falciparum (n = 25), P. vivax (n = 34) and P. falciparum plus P. vivax (n = 33)) Cont=27	563	276
SNPs tested							rs41274768 rs17047660 **rs17047661** rs6691117	rs17047661
Knops Blood Groups tested		Sl2/2 and Sl2/2 McCoy$^{a/b}$	Sl2 McC^{b}	McC^{b}, Sl2, Kn^{b} and KAM^{-}	Kn^{a}/Kn^{b} McC^{a}/McC^{b} Sl1/Sl2	Kn^{a}/Kn^{b} McC^{a}/McC^{b} Sl1/Sl2 Sl4/Sl5 KAM^{-}/KAM^{+}	-	-
Risk allele or genotype	Sl1/Sl1 genotype				-	Kn^{b} and KAM^{+}	Sl1 allele rs17047661-A	
Phenotype	Cerebral Malaria		Severe Malaria	Severe Malaria	Malaria	Malaria	Time to	Time to Infection

measured				infection	infection	Infection	
Strength of association (Odds ratio (OR) or beta)	p= 0.02 (95% CI 0.04 to 0.72) OR = 0.17 and p= 0.02 OR = 0.18, (95% CI = 0.04 to 0.77)	No effect	No effect	No effect	OR=6.68 P=0.04	P= 0.009 (95% Cl 0.665 to 0.942) OR= 0.791 (Sl2 allele)	No effect
Covariates Analysed	Combination allele (Sl and McCoy)and ethnic group	Polymorphisms, haemoglobin genotypes, Sex, ethnic group and location of residence	Variants, parasitaemia, age and sex	Variants	Variants, sex and age	Mother age, malaria suspicion during pregnancy, sex, birth term, mosquito net, ethnic group, sickle cell, chloroquine intake during pregnancy and SNP allele count	Mother malaria infection, sex and rs17047661 allele count

As previously reported, a nonsynonymous SNP rs6691117 in the *CR1* gene was found to be associated with ESR. The minor allele G of rs6691117 was associated with lower ESR (Kullo et al., 2011). Additionally, other SNPs within or surrounding *CR1* (rs12034383, rs12034598 and rs7527798) were associated with ESR (Kullo et al., 2011; Naitz et al., 2012). From our results we did not find any association between rs6691117 and malaria and these SNPs were not in high LD with nor *CR1*-B or *CR1*-C of *CR1* in HapMap populations. These results show that the SNPs' association with ESR does not confirm any association with malaria. Also, Sl2 allele of *CR1* was not in high LD with any other SNPs in and around *CR1* for the YRI population, which is likely to have the same LD pattern as the Tori-Bossito and Tolimmunpal populations. Therefore, Sl2 is likely to be a true protective allele against malaria.

Human populations show genetic diversity as a result of geographic dispersion. A recent genomewide study revealed that rs17047661 in the *CR1* locus was found to be among the most differentiated sites between Europe and Africa (the Sl2 allele (G) frequency is 0.01 in Europeans but 0.71 in Africans (Möller et al., 2016; Gurdasani et al., 2015). We have found no evidence of association of Sl2 allele frequency with malaria prevalence across Africa. It shows the limitations of correlating Sl2 allele frequency change, which occurs slowly relative to malaria prevalence. The first explanation for this may be a true lack of association of Sl2 allele frequency with malaria prevalence within the African populations studied. The second explanation for our results may be lack of malaria prevalence data which was obtained from the Malaria Atlas Project which provides malaria prevalence data only since 2000. Therefore, the malaria prevalence data which had driven the evolution towards protective alleles for malaria because of its survival advantage may be unnoticed that could not be applied to our analysis. Another reason for this might be migration between African populations. For example, some of the South African Bantu populations (Zulu, Sotho) live in a low malaria environment but have high frequency of Sl2, perhaps because allele frequencies have been maintained at a high level by drift since the Bantu migration from malaria-endemic West Africa (Gurdasani et al., 2015).

I tested the departure from HWE for all of the SNPs (Knops blood group-determining) and there was no departure from HWE for any of the SNPs. This test was important to

understand whether the variants were subjected to selection because of malaria. Therefore, it is not possible to interpret that the Sl2 allele is selectively favoured because of resistance to malaria.

Finally, in order to search for evidence of selection of the duplication (*CR1-B*) for malaria, the pathogen diversity index within HGDP population was examined and the results showed that there was no correlation between protozoa (including malaria) diversity or viruses or helminths and duplication frequency of *CR1*. Besides, bacterial diversity showed correlation with duplication (*CR1*-B) meaning that populations that experience a high diversity of bacterial pathogens show a high frequency of *CR1*-B. Therefore, bacterial pathogens may be driving *CR1*-B allele frequency change but not malaria.

5 TESTING THE ASSOCIATION OF COMPLEMENT C3B/C4B RECEPTOR 1 (*CR1*) COPY NUMBER VARIATION WITH ALZHEIMER'S DISEASE

5.1 Alzheimer's disease

Studies conducted by the Alzheimer's Society in 2014 revealed that there would be 850,000 people living with dementia in UK by 2015 and this will cost the UK £26 billion a year. This means an average annual cost of £32,250 per person (Alzheimer's Society, 2014-2015). According to the Alzheimer's Association (Alzheimer's Association, 2017) an estimated number of 5.5 million Americans (two-thirds of them women) are living with Alzheimer's dementia and an estimated 5.3 million of these are aged 65 or older whereas 200,000 are under 65, having younger-onset Alzheimer's.

Familial Alzheimer's disease (AD), which is a result of hereditary mutations and is an early onset form that can occur as early as 40 years of age, comprises 2% of diagnosed cases, whereas the majority of the patients suffer from sporadic AD which is subdivided into early (under 65 years of age) and late onset (the rest of the cases), which comprise 3-5% and 95-97% of diagnosed cases, respectively (Folch et al. 2015).

Thus, a small portion of AD occurs before the age of 65 and is called early-onset Alzheimer's disease (EOAD) (Seltzer and Sherwin, 1983). EOAD has a tendency to follow Mendelian patterns of inheritance (Goate, et al., 1991) whereas late-onset Alzheimer's disease (LOAD), also called sporadic AD, is more common, diagnosed after age 65 and caused by several genetic, epigenetic and environmental factors (Gatz, et al., 2006; Zawia, et al., 2009). EOAD has more severe outcomes compared to LOAD, such as dramatic neural loss and synaptic fall-out (Nochlin, et al., 1993). Different genetic factors underline AD, for example rare mutations in *APP1* and *APP2* cause autosomal dominant familial EOAD whereas variants in the eapolipoprotein E *(APOE*) gene are risk factors for both EOAD and LOAD but are more important risk factors for LOAD (Atkins, et al., 2012; Atkins and Panegyres, 2011). Another genetic difference between EOAD and LOAD is the higher proportion of the *APOEε2* allele of *APOE* in EOAD than in LOAD patients (Panegyres and Chen, 2013). Therefore, early and late onsets of Alzheimer's disease have differences in progression of the disease that may have different causes (Rogaeva, 2002). According to GWAS, the *CR1* gene is found to

be one of the most important genetic susceptibility loci to LOAD after the *APOE* gene *(*Zhu et al., 2014).

5.2 Early-onset Alzheimer's Disease

5.2.1 Study Rationale

A previous study found that the *CR1*-B allele (duplication of LCR1 region) of *CR1* increased AD risk by 30% in a French late-onset Alzheimer's disease (LOAD) cohort (Brouwers et al., 2012). They suggested that this is caused by the presence of *CR1*-B because it increases the number of C3b/C4b binding and cofactor activity sites. My own analysis of *CR1*-B and LOAD follows later in this chapter. However, there has been no previous study to investigate any association between EOAD and *CR1*. Because of the stronger phenotype of the EOAD we might expect a stronger association between *CR1*-B and EOAD compared to LOAD.

5.2.2 Estimation of CR1 copy number in EOAD

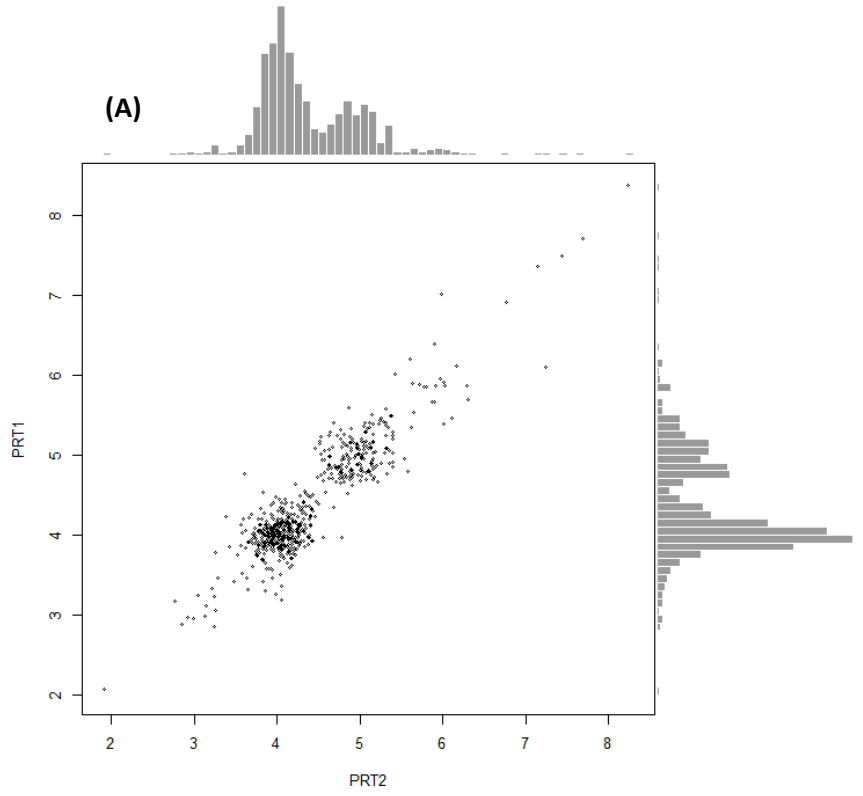
This case-control study used DNA samples which were collected from EOAD patients and controls, and measured the copy number of LCR *CR1*. This cohort consists of 633 individuals of European origin who lived in Nottingham, Bristol, Manchester, Oxford, Bonn and Southampton and was made available to me by our collaborator Professor Kevin Morgan (University of Nottingham, UK). This cohort is composed of 449 cases and 184 controls and details are shown below (Table 5.33). Three independent PRT assays (PRT1, PRT2 and PRT3) were used to determine the *CR1* LCR copy number in these samples.

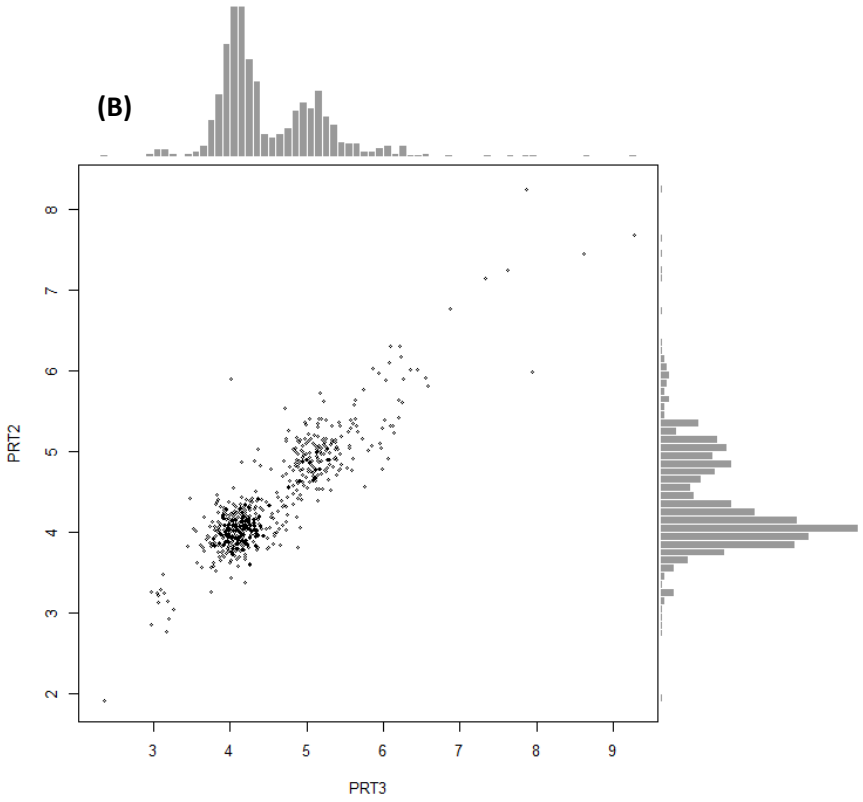
Table 5.33: The sample size and average age details for cases and controls with information about the locations of collection of the samples in the EOAD cohort.

EOAD	Number of samples	Female	Male	Bristol	Nottingham	Manchester	Oxford	Bonn	Southampton	Age Onset Average
Cases	449	219	230	31	45	346	27	-	-	56.88
Controls	184	92	92	-	35	-	-	132	17	74.51
Total	633	311	322	31	80	346	27	132	17	61.26

Firstly, in order to understand the quality of the data, the raw copy numbers generated by the three different PRT assays (PRT1-3) were compared using scatter plots to observe distinct clusters (Figure 5.40). This comparison of each assay provides a better understanding of the data quality and also which assay or assays are good to use for copy number calling individually or together for a unique cohort.

EOAD





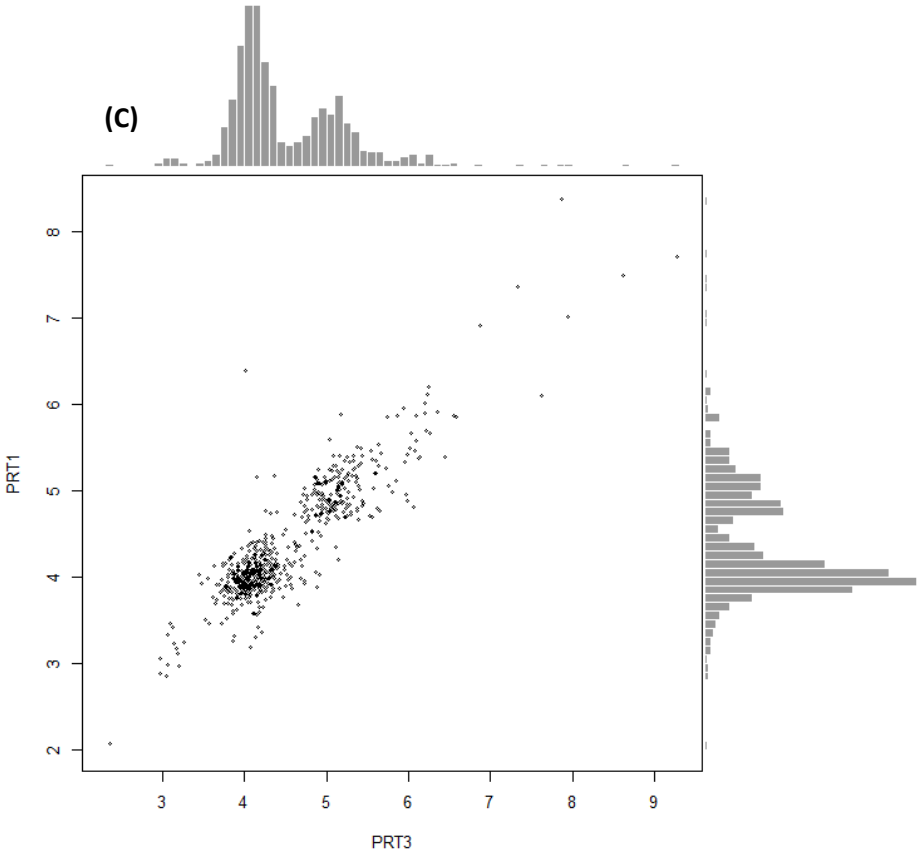


Figure 5.40: Raw data comparison of PRT1-3 for the EOAD cohort. **(A)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT2 data in the EOAD cohort. **(B)** Scatter plot shows the comparison of raw data normalized to known copy number of PRT2 and raw data normalized to known copy number of PRT3 data in the EOAD cohort. **(C)** Scatter plot shows the comparison of raw data normalized to known copy number of PRT1 and raw data normalized to known copy number of PRT3 data in the EOAD cohort.

According to the scatter plots (Figure 5.40), PRT1, PRT2 and PRT3 assays give similar copy numbers and provide separate and well-defined clusters in this EOAD case-control cohort.

Therefore, these three independent PRT assays were used to determine the integer copy number of *CR1* LCR copy number in the EOAD cohort. The average copy number of PRT1, PRT2 and PRT3 for each sample is used to measure the integer copy number of *CR1* (LCR) (Figure 5.41).

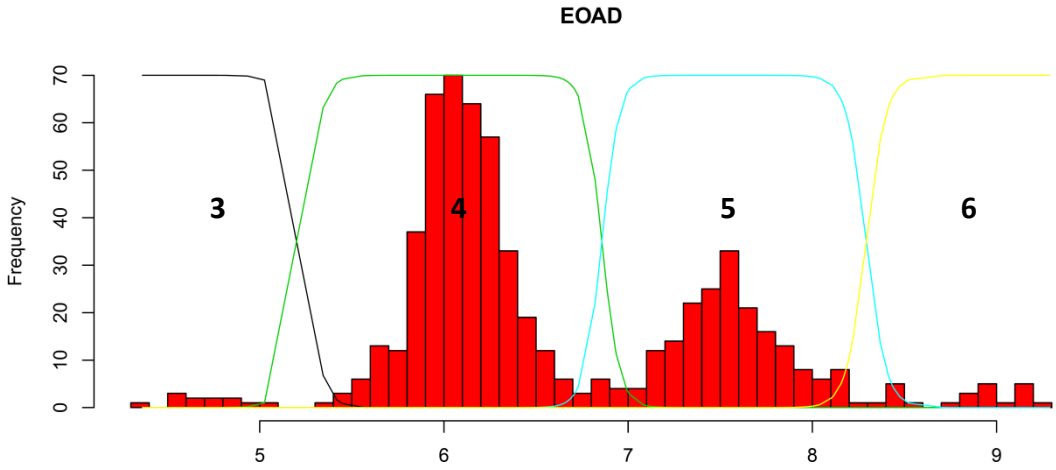


Figure 5.41: Population distribution of diploid LCR *CR1* copy number in EOAD samples. The coloured lines show the Gaussian distributions for each of the four copy number classes (diploid copy number = ≤3, 4, 5, 6 and ≥7). The x-axis indicates LCR *CR1* diploid copy number data after division by the standard deviation of the entire dataset.

The histogram of PRT ratio data indicates four clusters. This histogram shows the average of three PRT assays after division by the standard deviation of the entire dataset. The number of clusters was used to measure integer copy number of *CR1* (LCR) using CNVtools. The diploid copy numbers of the samples were used to relate them to genotypes as described in section 3.4 Worldwide Distribution of LCR CR1 Gene Copy Number). The clusters were counted as 3, 4, 5 and 6. The higher copy number of more than six copies (7, 8 and 9) and the copy number of two were rare, so they were not put on the histogram but for further analysis were added manually afterwards.

I have previously shown (section 3.4 Worldwide Distribution of LCR CR1 Gene Copy Number) that the four co-dominant alleles of the *CR1* can be assessed according to the LCR *CR1* copy number (Table 5.34). The most frequent allele of *CR1* is *CR1*-A which matched with our modal copy number 4, the most frequent CN in the EOAD cohort. The others are combined with the same method according to their frequency in EOAD. Therefore, 5 copies of LCR CR1 (*CR1*-B) was the second most frequent CN for whole cohort whereas 6 copies or higher and 3 copies and lower were observed less frequently in the EOAD cohort (Table 5.34).

Table 5.34: LCR *CR1* copy numbers can be used to determine the *CR1* genotype in EOAD cohort. As the molecular weight of the protein product of each allele is known, LCR copy numbers can be used to determine the likely *CR1* genotype of each individual. The numbers represent the copy number of LCR *CR1*. The LCR *CR1* copy number of 4 was the most frequent CN in EOAD cohort and this copy number should be homozygote for *CR1*-A allele most frequent allele in the EOAD cohort. The LCR *CR1* copy number of 5 was second most frequent in EOAD individuals and it should be heterozygote for *CR1*-A and *CR1*-B alleles. The LCR *CR1* copy numbers of 7 and 8 were rare when compared to the other copy numbers and they have matched with rare alleles of *CR1* which shows low allele frequency in EOAD cohort.

***CR1* Copy Number Genotype**				
Genotype	***CR1*-C (F') (1 copy)**	***CR1*-A (F) (2 copies)**	***CR1*-B (S) (3 copies)**	***CR1*-D (4 copies)**
***CR1*-C (F') (1 copy)**	2	3		
***CR1*-A (F) (2 copies)**	3	4	5	
***CR1*-B (S) (3 copies)**		5	6	7
***CR1*-D (4 copies)**			7	≥8

Barnes et al. (2008) showed that existing strategies to detect meaningful CNV associations with binary disease phenotypes are limited by differential errors and poor clustering quality in a case-control setting. They developed methods to overcome these issues. The clustering quality or degree of clustering (Q) is a simple averaged value of signal to noise ratios between adjacent copy numbers (Barnes, et al., 2008). They showed that discrete copy number clusters (Q value 4 or more (better)) are more likely to achieve the maximum theoretical power whereas with decreasing clustering quality, the power drops dramatically. They also showed that there is a much greater loss of power when the different measurement properties between cases and controls are included in the model. In this study, the Q value was calculated according to Barnes, et al. (2008) for this case and control study (Appendix 10) using R (CNV tools). Barnes, et al. (2008) stated that the clustering quality should be at least 4 or more than

4 in order to achieve a power of 80% (probability that $p< 10^{-4}$) or higher. Therefore, having a clustering quality value of 5 or more is better than 4. Thus, the copy number data which gave highest Q value after clustering were used for this study. The Q value was 5.56; therefore the fitting of the data was good according to the criteria of Barnes et al. (2008) (Table 5.45).

These results show that PRT assays gave discrete copy number data with high clustering quality in the EOAD cohort as the Q value was 5.56. Therefore, our robust PRT assays provided reliable copy number data to assess the diploid copy number of *CR1* (LCR). These PRT data are used later for association study in EOAD cohort to find possible association between EOAD and the *CR1*-B allele.

Finally, the integer copy numbers which are obtained from the histogram (Figure 5.38) are used to estimate the duplication allele frequency. This value is used in further analysis to test for an association between EOAD and the *CR1*-B allele.

5.2.3 Association of Copy Number Variation of CR1 in the EOAD Cohort

In order to determine the contribution of the *CR1*-B allele to EOAD logistic regression analysis was used. APOE- ε4 allele count, *CR1*-B allele count (0, 1 and 2) and sex are used as covariates in this association study (Table 5.35). A linear additive model was assumed for *CR1*-B and APOE-ε4 allele assuming that each extra *CR1*-B and APOE-ε4 allele in the genotype has the same effect on risk. APOE-ε4 allele count is used as a covariate because it has been known that APOE-ε 4 allele is a risk factor for both EOAD and LOAD. Sex is used as a covariate in order to understand if sex has a role in EOAD. Finally, *CR1*-B allele is used as covariate to determine the contribution of *CR1*-B allele count to EOAD (Table 5.36).

Table 5.35: The *CR1*-B allele count for cases and controls was assessed according to the diploid copy number of LCR *CR1* obtained from PRT1-3 assays in the EOAD cohort. The data is used for the association study between *CR1*-B allele and EOAD.

Diploid Copy Number of LCR *CR1* (EOAD)	**Genotypes**	**Copies Duplication**	**Cases**	**Controls**
>6	*CR1*-B/B and higher	2	22	9
5	*CR1*-B/C	1	137	55
<4	*CR1*-A/A and lower	0	290	120

Table 5.36: The result of the association studies of *CR1*-B allele in EOAD samples shows no associations between *CR1*-B and EOAD.

Parameter	**Exp(B)**	**Std. Error**	**95% Wald Confidence Interval**		**Hypothesis Test**		
			Lower	**Upper**	**Wald-chi Square**	**df**	**P-value**
ApoE-ε4_allele	1.25	0.172	0.910	1.58	52.52	1	4.25×10^{-13}
(*CR1*-B allele)	0.049	0.159	-0.261	0.360	0.097	1	0.755
Sex=Female	0.019	0.184	-0.342	0.380	0.011	1	0.918
Sex=male	-	-	-	-	-	-	-

Dependent Variable: Cohort (Case/Control), **Factor:** Sex and **Covariates:** Sex, ApoEε4_allele_count and CR1-B

The results above suggested that there was no significant association between *CR1*-B and EOAD. The *CR1*-B allele was not found to be associated with EOAD in this study because the p-value for duplication was 0.755 and not significant. However, the APOE-ε4 allele is associated with EOAD. The result was highly significant for the APOE-ε4allele and it has already been known that APOE-ε4 allele is a risk allele for EOAD and shown in previous studies (Atkins, et al., 2012; Atkins and Panegyres, 2011). Sex did not show any association with EOAD.

Table 5.37: The results of the association studies of LCR *CR1*-C in EOAD samples showed no associations between *CR1*-C and EOAD.

Parameter	Exp(B)	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald-chi Square	df	P-value
ApoE-ε4_allele	1.26	0.172	0.918	1.59	53.1	1	$3.21x10^{-13}$
***CR1*-C**	0.647	0.63	-0.594	1.89	1.04	1	0.307
Sex=Female	0.027	0.185	-0.335	0.388	0.021	1	0.886
Sex=male	-	-	-	-	-	-	-

Dependent Variable: Cohort (Case/Control), **Factor:** Sex and **Covariates:** Sex, ApoEε4_allele_count and CR1-C

I also investigated whether the *CR1*-C allele was associated with EOAD. The results above suggested that this allele was not found to be associated with EOAD in this study because the p-value for duplication was 0.307 and not significant (Table 5.37).

According to genome-wide association studies (GWAS) it is known that rs3818361, rs6656401 and rs6701713 are located in the *CR1* gene and they are risk loci for Alzheimer's disease (Hollingworth, et al., 2011; Lambert, et al., 2009; Naj, et al., 2011) (Table 5.38).

Table 5.38: Correlation studies (GWAS) between AD and SNPs in 250kb region of the *CR1* gene.

Variants	Nearest genes	GWAS Disease Association	Allele	Odd Ratio or beta	References
rs6656401	*CR1*	Alzheimer's Disease	A	1.22	**Lambert et al., 2009**
rs3818361			A	1.22/1.18	**Lambert et al., 2009, Hollingworth et al., 2011**
rs6701713			A	1.16	**Naj et al., 2011**

GWAS previously have used late-onset Alzheimer's disease cohorts for their analysis, and no previous GWAS has shown whether these loci are risk locus for EOAD or not. Therefore, we investigated the association of these AD risk loci with EOAD in this cohort. The genotype data for these SNPs were given to us by our collaborator Prof Kevin Morgan.

Table 5.39: The results of the association studies of an AD risk variant in EOAD samples show no associations between rs3818361 and EOAD.

Parameter	Exp(B)	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald-chi Square	df	P-value
ApoEε4_allele	1.22	0.181	0.867	1.58	45.4	1	$1.61x10^{-11}$
rs3818361	0.114	0.180	-0.239	0.468	0.401	1	0.527
Sex=Female	0.121	0.195	-0.261	0.504	0.386	1	0.535
Sex=Male	-	-	-	-	-	-	-

Dependent Variable: Cohort (Case/Control), **Factor:** Sex and **Covariates:** Sex, ApoEε4_allele_count and rs3818361

The results above suggest that rs3818361 is not associated with EOAD in this study because the p-value for duplication is 0.527 and not significant (Table 5.39).

Table 5.40: The results of the association studies of an AD risk variant in EOAD samples show no associations between rs6656401 and EOAD.

Parameter	Exp(B)	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald-chi Square	df	P-value
ApoEε4_allele	1.23	0.182	0.869	1.57	45.6	1	$1.48x10^{-11}$
rs6656401	-0.004	0.18	-0.356	0.348	$4.58x10^{-4}$	1	0.983
Sex=Female	0.124	0.195	-0.258	0.507	0.406	1	0.524
Sex=Male	-	-	-	-	-	-	-

Dependent Variable: Cohort (Case/Control), **Factor:** Sex and **Covariates:** Sex, ApoEε4_allele_count and rs6656401

The results above suggest that rs6656401 is not found to be associated with EOAD in this study because the p-value for duplication is 0.983 and not significant (Table 5.40).

Table 5.41: The results of the association studies of an AD risk variant in EOAD samples show no associations between rs6701713 and EOAD.

Parameter	Exp(B)	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald-chi Square	df	P-value
ApoEε4_allele	1.22	0.182	0.868	1.58	45.4	1	$1.58x10^{-11}$
rs6701713	-0.046	0.175	-0.389	0.297	0.069	1	0.793
Sex=Female	0.123	0.195	-0.260	0.505	0.397	1	0.529
Sex=male	-	-	-	-	-	-	-

Dependent Variable: Cohort (Case/Control), **Factor:** Sex and **Covariates:** Sex, ApoEε4_allele_count and rs6701713

The results above suggest that rs6701713 is not associated with EOAD in this study because the p-value for duplication was 0.793 and not significant (Table 5.41).

In conclusion, there was no evidence for either *CR1*-B or *CR1*-C association with EOAD. There was also no evidence for LOAD GWAS SNPs association with EOAD. Therefore, *CR1* is not a risk locus for EOAD.

5.2.4 Linkage Disequilibrium Analysis in EOAD

The matched *CR1* LCR copy number data and SNPs (Figure 5.42) data were known for this cohort therefore the pattern of LD across the region was analysed.

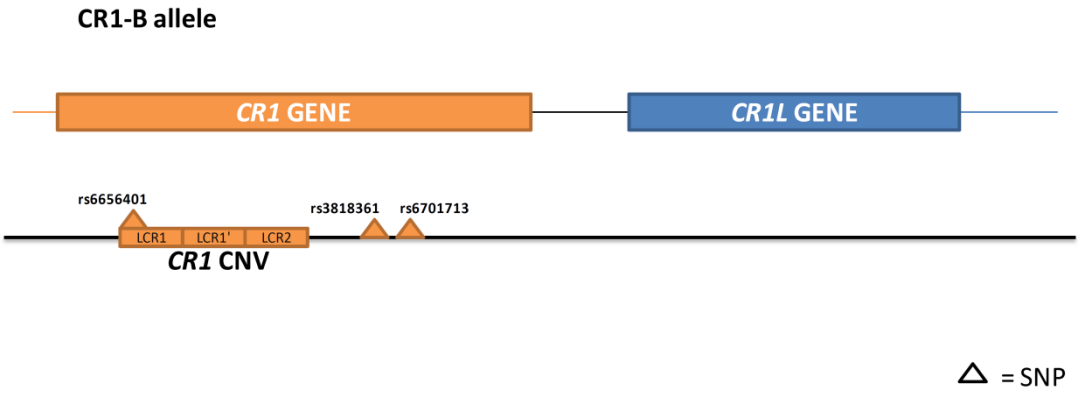


Figure 5.42: The locations of LOAD risk SNPs (rs6656401, rs3818361 and rs6701713) within *CR1* are shown in *CR1*-B allele.

The SNPs were in moderate LD with *CR1*-B (Table 5.42).

Table 5.42: The risk variants (SNPs) for Alzheimer's disease are not in strong LD with *CR1*-B.

Risk Variant for AD (SNP)	**R^2**	**D'**	**Haplotype** *CR1*-B (A) A/T	**Frequency**	**Found Risk Allele**
rs3818361	0.474	0.711	**AT**	0.146	T
			TT	0.054	
			AC	0.044	
			TC	0.756	
rs6656401	0.440	0.674	**AA**	0.139	G
			TA	0.055	
			AG	0.049	
			TG	0.757	
rs6701713	0.461	0.709	**AA**	0.146	G
			TA	0.058	
			AG	0.044	
			TG	0.752	

In addition, these risk variants are in strong LD with each other (Table 5.43).

Table 5.43: The results of the LD analysis in EOAD which shows the LD between AD risk variants with each other.

LD between SNPs	R^2	D'	Haplotype	Frequency
rs3818361 and rs6656401	0.929	0.992	**AT**	0.193
			GT	0.011
			AC	0.001
			GC	0.795
rs3818361 and rs6701713	1.000	1.000	**TA**	0.206
			CA	≤0.0001
			TG	≤0.0001
			CG	0.794
rs6656401 and rs6701713	0.929	0.992	**AA**	0.194
			GA	0.011
			AG	0.001
			GG	0.795

5.3 Late-onset Alzheimer's Disease

5.3.1 Rationale of study

In the light of previous work (Brouwers et al., 2012) we wanted to investigate *CR1*-B and LOAD with three different PRT assays in a larger LOAD cohort to investigate the role and contribution of duplications (LCR1) to LOAD. In the previous *CR1* CNV association study a French LOAD cohort was used with 1393 cases (69.5±8 years, 66.1% women) and 610 controls (72.7±8.1 years, 62.4% women) by using a multiplex amplicon quantification (MAQ) dosage assay (Brouwers et al., 2012). This found that having the *CR1*-B increased AD risk by 30% in *CR1*-B heterozygotes. The current study used a larger LOAD cohort with 1263 cases and 1005 controls (in total 2268) and the *CR1*-B allele genotyped using the PRT approach to try and replicate the previous association of *CR1*-B with LOAD. 83 case samples were excluded from the LOAD cohort because of uncertainties in their disease status, leaving 2185 samples: 1180 cases and 1005 controls, for the association study.

5.3.2 The Power of the LOAD Study

Using the effect size estimated by the previous study (Brouwers et al., 2012), the power of this study was calculated by Dr Edward Hollox. The three approaches are explained later in this chapter. The number of the case samples used in LOAD study for the first and second approach is 1180 whereas control samples are 1005. In total the sample size is 2185 (as 83 cases are excluded from whole LOAD cohort). The power to detect a significant result ($p<0.05$), of a cohort size similar to Brouwers et al. (2012) was maximum 0.839 and minimum 0.797.

In the third approach, the sample size is smaller. The number of the case samples used in LOAD study for the third approach is 929 (as 83 cases are excluded from whole LOAD cohort) whereas control samples are 940. In total the sample size is 1869 (316 of samples are removed). The power to detect significant result ($p<0.05$) was maximum 0.765 and minimum 0.741.

5.3.3 Characteristics of the Cohort

Three different PRT assays (PRT1, PRT2 and PRT3) are used to deduce the copy number of *CR1* (LCR). In the LOAD cohort, the DNA quality was variable as the DNA was collected from different locations (Mixed [Manchester, Nottingham, Oxford, Bristol and Southampton], Bristol, Oxford and Southampton) and also the DNA amount was limited (20 ng). In comparison, the LOAD cohort data are not as good as the EOAD cohort data. The LOAD cohort has shown a notable failure rate so that some of the samples (316 samples, ~14% of whole LOAD cohort), failed to yield results from all three PRTs. Therefore, PRT1 or PRT1 and PRT3 had to be repeated for these samples that these samples had only one or two (PRT1 or PRT1 and PRT3) copy number data from three PRTs. These samples are still included in the analysis. For example, a sample which gave copy number data from the PRT1 assay or the PRT1 and PRT3 assays was included. The total number of samples used in the analysis is 2268. 1263 of these samples are cases and 1005 of these samples are controls. For the LOAD association study, 2185 samples were used (1180 cases and 1005 controls) and details about the samples are shown below (Table 5.44).

Table 5.44: The *APOE*-ε4 allele count for cases and controls in the LOAD cohort. The sex (factor) differences between cases and controls are also shown. The data were used for the association study as a covariate.

LOAD	Number of samples	Female	Male	*APOE*-ε4 allele (0)	*APOE*-ε4 allele (1)	*APOE*-ε4 allele (2)	Age Onset Average
Cases	1180	710	470	495	546	139	75.061
Controls	1005	568	437	736	244	25	73.167
Total	2185	1278	907	1231	790	164	74.19

5.3.4 Estimation of CR1 LCR copy number

Three approaches were used in order to assess the quality of data in this case and control study. In order to understand which data were the best to do the association analysis, Q values were calculated and compared for each approach. Three approaches were used for this study. They were 'all' (all the LOAD data included cases and controls), 'sites' (the data is fitted according to the location of samples: Bristol, Oxford, mixed and Southampton) and 'PRT123' (the 316 samples which failed to give copy number data for all PRT assays, were removed). The Q values are measured to understand which of the fitting methods is best (Table 5.45). In the first approach, 2268 of the samples were used. In order to obtain the integer copy number, the diploid copy number data of *CR1* LCR were separately plotted (fitted) according to 'all'. In order to obtain integer copy number using histogram, the data can be plotted according to all or case/control. All means that the copy number data to obtain integer copy number is determined by plotting it without any discrimination (no consideration for case/control) whereas cases/controls means the integer copy number is determined by plotting the data set separately into two as cases and controls. In this study, 'cases/controls' were used for fitting to obtain diploid integer copy number of *CR1* LCR because the Q value was 4.05 for cases/controls (degree of clustering is higher) and it was bigger than 4.03 (all).

Table 5.45: The summary of the Q values obtained from the plotting**.** The Q values were measured for both EOAD and LOAD data in order to determine which fitting method is the best and also to check the quality of the data. **(1)** The EOAD data were better than the LOAD data because they give a higher Q value meaning more reliable copy number data. The EOAD is fitted according to 'all' data set and gives a Q value of 5.56. **(2)** In the first approach of LOAD copy number analysis, 2268 of the samples are used and the data is fitted according to 'all' and also cases/controls. The q value is 4.03 for fitting according to all and 4.05 for fitting according to cases/controls. **(3)** In the second approach, 2268 of the samples are used but the data are fitted according to samples collected locations 'sites'. The Q value for the sites is 3.20. **(4)** In the third approach, 1952 of the samples is used. 1012 of these samples are cases and 940 of these samples are controls and data are fitted according to all and cases/controls. The Q value is 4.13 for fitting according to all and 4.05 for fitting according to cases/controls. An R package ('CNVtools') is used to calculate the Q values (Appendix 10).* Not chosen because of the sample size (316 samples removed): The power of the study would be less than 0.8.

Cohort	**Fitting method**	**Q value**	**Chosen/not for association study**
(1) EOAD	All	5.56	**Chosen**
(2) LOAD First Approach	Cases and controls	4.05	**Chosen**
	All	4.03	Not chosen
(3) LOAD (SITES) Second approach	Sites	3.20	Not chosen
(4) LOAD (PRT123) Third approach	Cases and controls	4.05	Not chosen
	All	4.13	Not chosen*

The histogram of PRT ratio data indicates five clusters. The number of clusters was used to measure integer copy number of *CR1* (LCR) using CNVtools. The clusters were counted as ≤3, 4, 5, 6 and 7≤ and fitted as all (Figure 5.43).

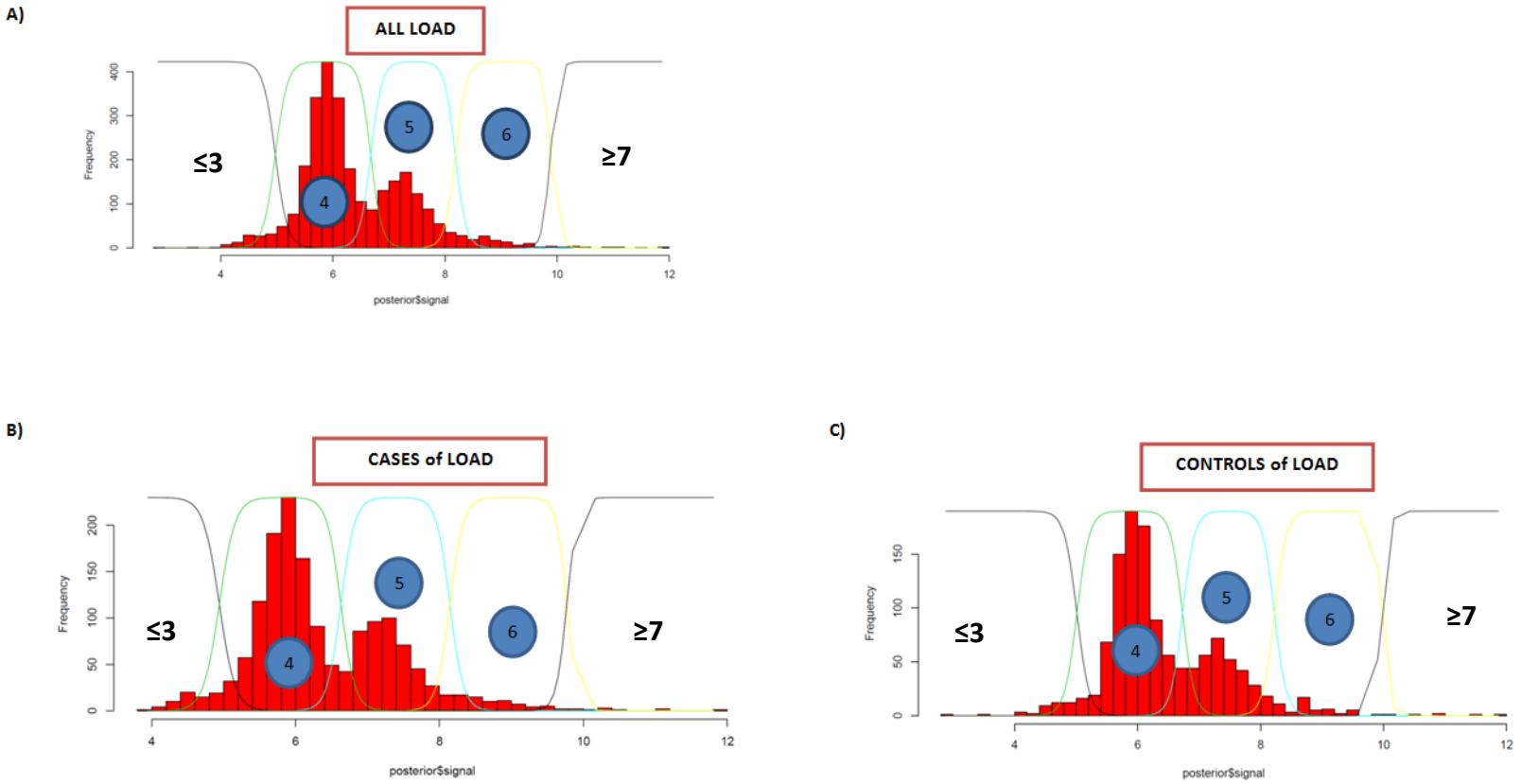


Figure 5.43: Distribution of LCR *CR1* copy number in LOAD samples for first approach. **A)** All samples of LOAD cohort distribution of LCR *CR1* copy number in LOAD samples **B)** Cases of LOAD cohort distribution of LCR *CR1* copy number in LOAD samples **C)** Control of LOAD cohort distribution of LCR *CR1* copy number in LOAD samples. The coloured lines show the Gaussian distributions for each of the five copy number classes (copy number = ≤3, 4, 5, 6 and ≥7). The x-axis indicates LCR *CR1* diploid copy number data after division by the standard deviation of the entire dataset.

In the second approach, 2268 of the samples were analysed. In order to obtain the integer copy number the data are fitted with sample collection city as a covariate. The Q value for the sites is 3.20 (Table 5.45). It is not better than the first approach. The data was used with the analysis in order to see how the data will affect the significance of the results for *CR1*-B.

The histogram of PRT ratio data indicates five clusters. The number of clusters was used to measure integer copy number of *CR1* (LCR) using CNVtools. The clusters were counted as 3, 4, 5 and 6 and fitted as all (Figure 5.44).

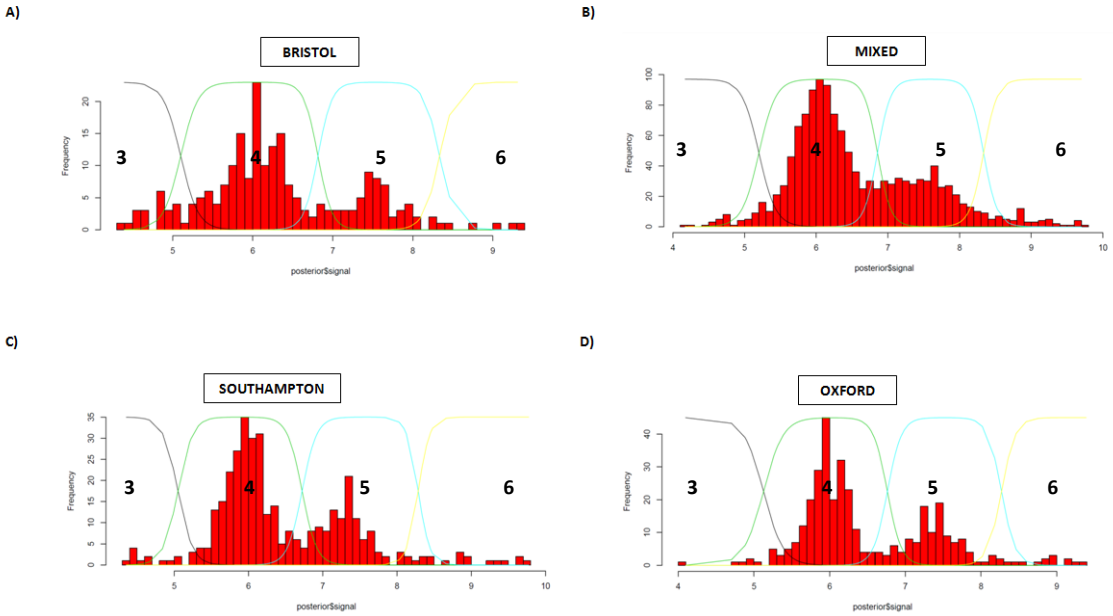


Figure 5.44: Distribution of LCR *CR1* copy number in LOAD samples for second approach. **A)** Bristol samples of LOAD cohort distribution of LCR *CR1* copy number in LOAD samples **B)** Mixed samples of LOAD cohort distribution of LCR *CR1* copy number in LOAD samples **C)** Southampton samples of LOAD cohort distribution of LCR *CR1* copy number in LOAD samples **D)** Oxford samples of LOAD cohort distribution of LCR *CR1* copy number in LOAD samples. The coloured lines show the Gaussian distributions for each of the four copy number classes (copy number = 3, 4, 5 and 6). The x-axis indicates normalized PRT ratio of LCR *CR1*. The higher copy number of more than six copies (7, 8 and 9) and copy number of two were rare and have not been put in the histogram but for further analysis they were added manually after. The x-axis indicates LCR *CR1* diploid copy number data after division by the standard deviation of the entire dataset.

In the third approach, the samples which are not giving copy number results for all PRTs (PRT1, PRT2 and PRT3) are discarded. Therefore, 1952 of the samples remain and are analysed. 1012 of these samples are cases and 940 of these samples are controls. The sample number is smaller for this approach. In order to obtain the integer copy

number the data are fitted according to all and case/controls. The Q value is measured to understand which of them was better. The Q value is 4.13 for fitting according to 'all' and 4.05 for fitting according to cases/controls. 4.13 is bigger than 4.05 so the 'all' fitting method was chosen for the analysis to fit the data (Table 5.45).

The histogram of PRT ratio data indicates five clusters. The number of clusters was used to measure integer copy number of *CR1* (LCR) using CNVtools. The clusters were counted as 3, 4, 5 and 6 and fitted as all and also cases/controls (Figure 5.45).

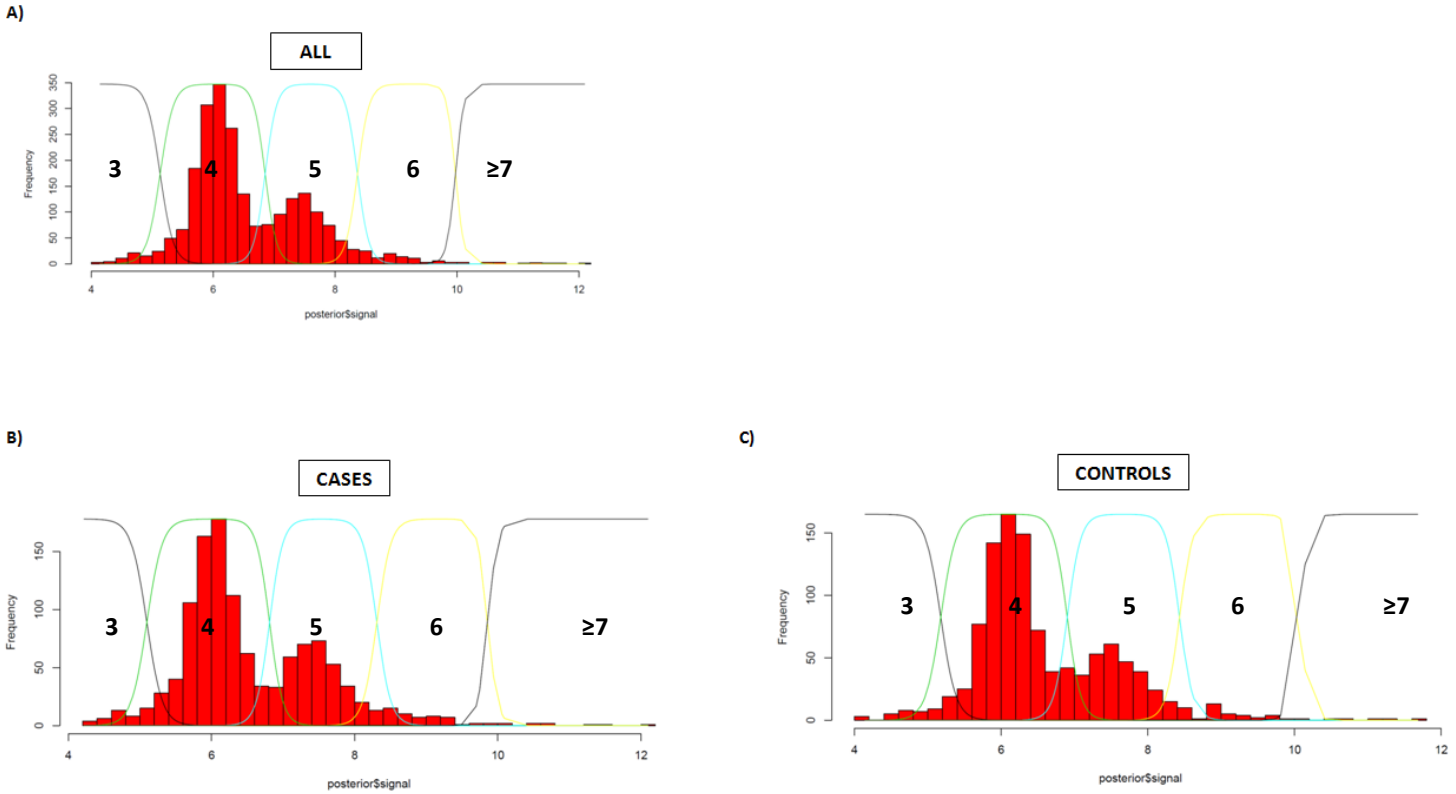
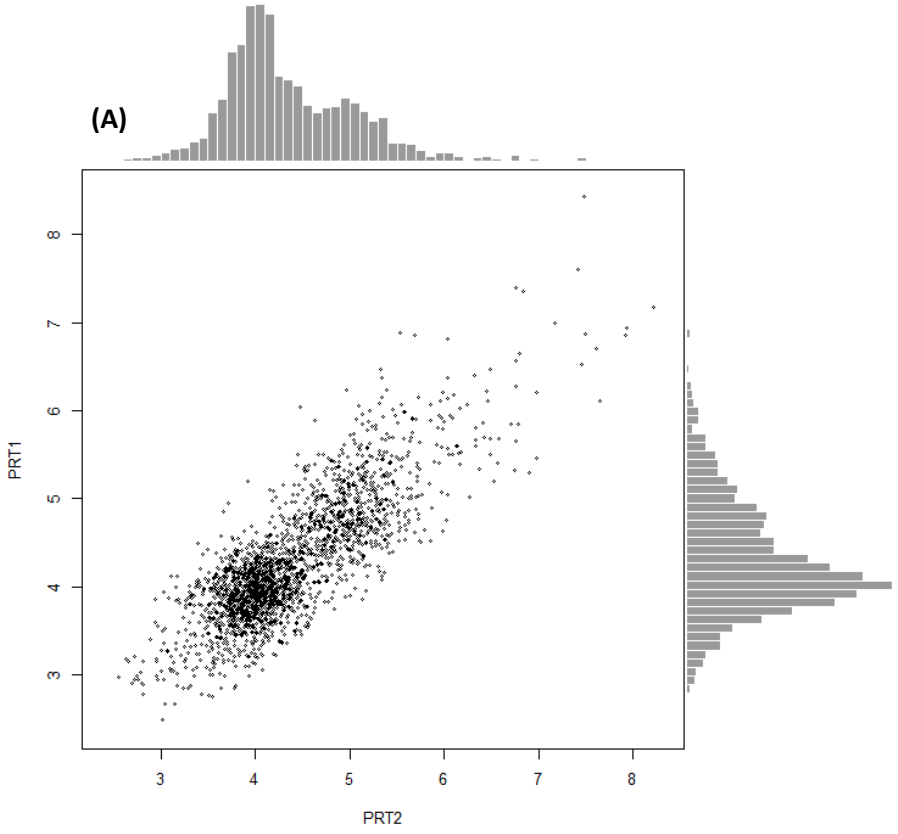
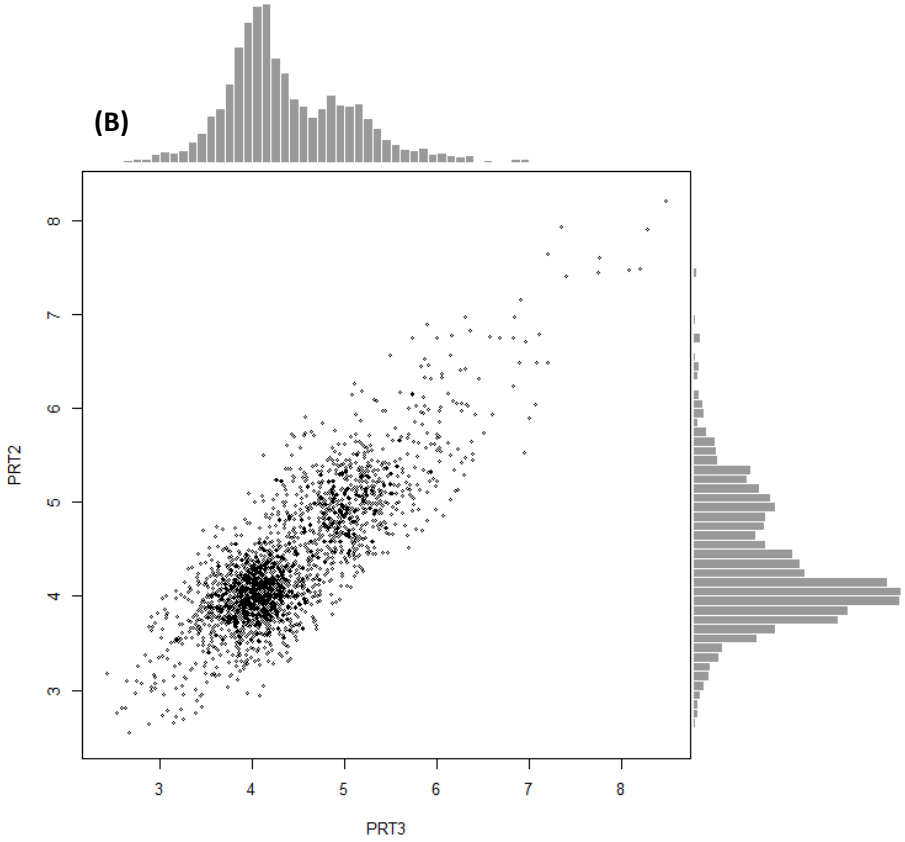


Figure 5.45: Distribution of LCR *CR1* copy number in LOAD samples for third approach. **A)** All samples of LOAD cohort distribution of LCR *CR1* copy number in LOAD samples (PRT1-3) **B)** Cases of LOAD cohort distribution of LCR *CR1* copy number in LOAD samples **C)** Control of LOAD cohort distribution of LCR *CR1* copy number in LOAD samples. The coloured lines show the Gaussian distributions for each of the five copy number classes (copy number =3, 4, 5, 6 and ≥7). The x-axis indicates LCR *CR1* diploid copy number data after division by the standard deviation of the entire dataset.

In order to understand the quality of the data and how each assay works with the data, PRT assays were compared with each other with unrounded copy numbers to see distinct clusters which give copy number of LCR (*CR1*) (Figure 5.46).

ALL





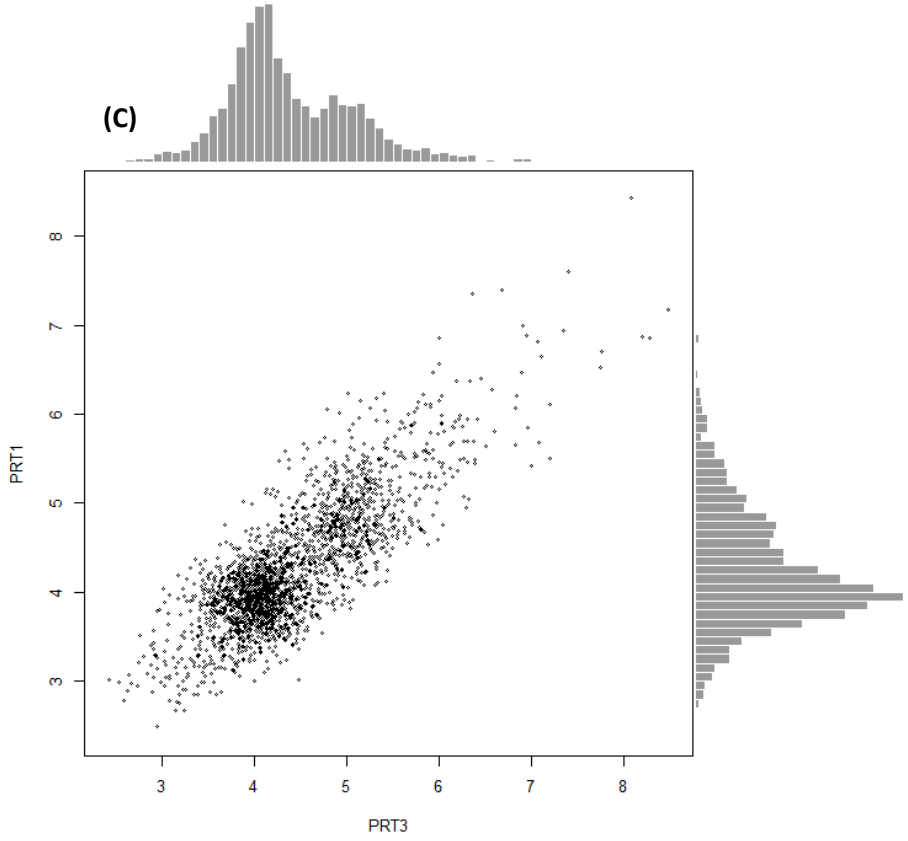
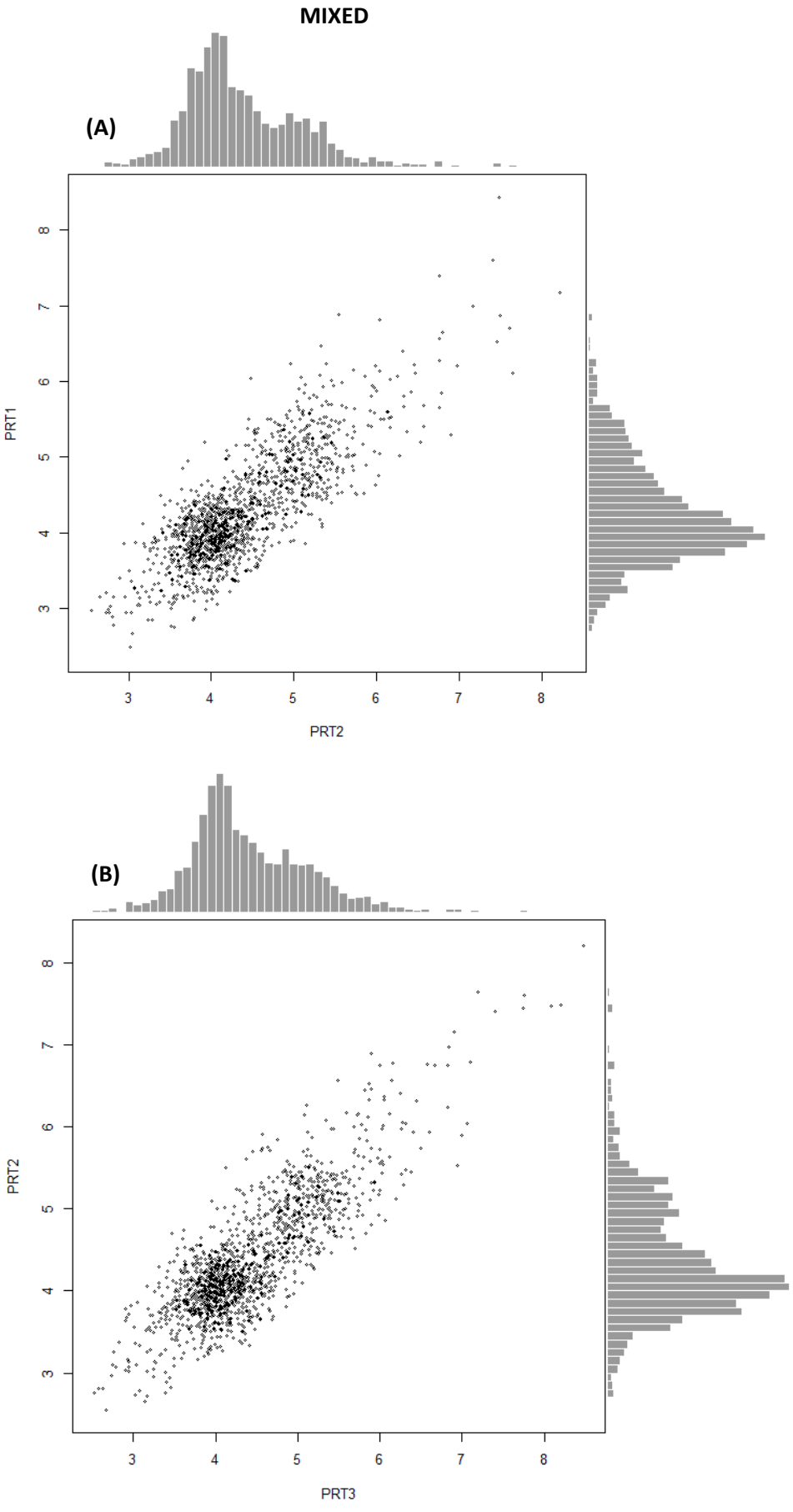


Figure 5.46: The comparison of raw PRT data for all. **(A)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT2 data in the late-onset Alzheimer's disease cohort (LOAD). **(B)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT2 data and raw data normalized to known copy number of PRT3 data in the LOAD cohort. **(C)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT3 data in the LOAD cohort.

This comparison of each assay provides a better understanding of the data quality and also of which assay or assays are good to use for copy number calling either individually or together for a unique cohort. According to the results (Figure 5.46), PRT1 gives noisier data than PRT2 and PRT3 meaning that PRT1 does not give discrete histograms and good clustering for copy number calling in the whole LOAD cohort. Therefore, PRT2 and PRT3 are better to assess the copy number for LCR1 in the LOAD cohort. In general, PRT1-3 gives good clustering for copy number calling in the whole LOAD cohort.


MIXED
(A)
PRT1
PRT2
3
4
5
6
7
8
(B)
PRT2
PRT3

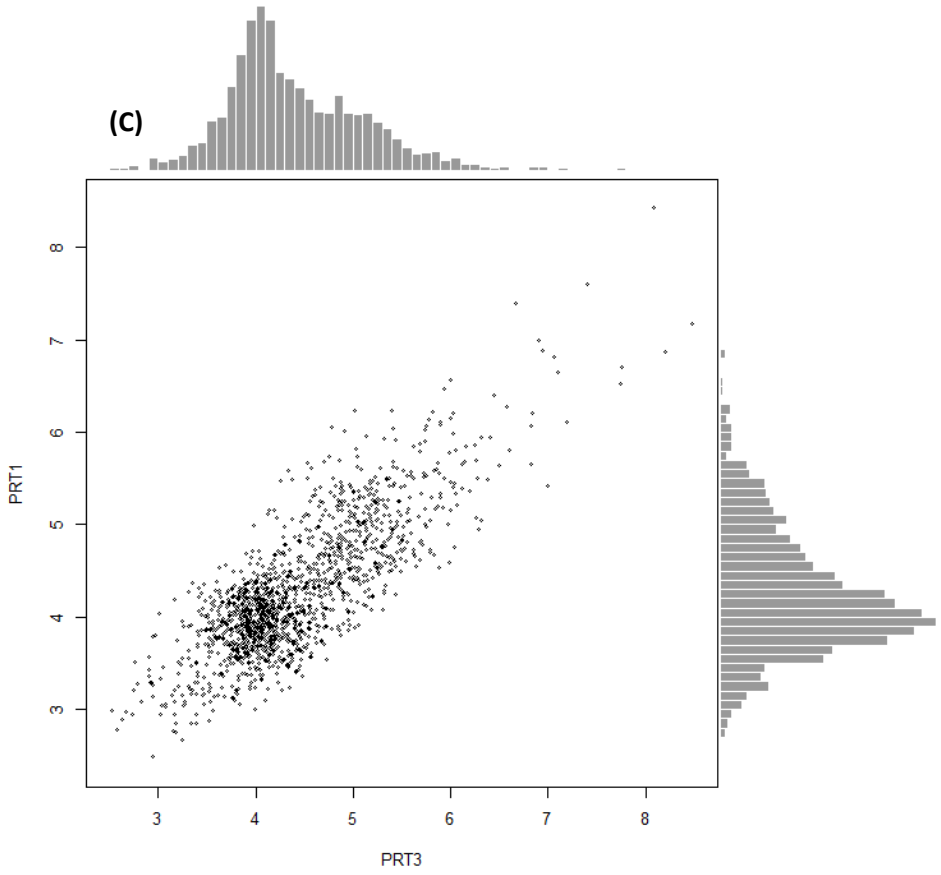
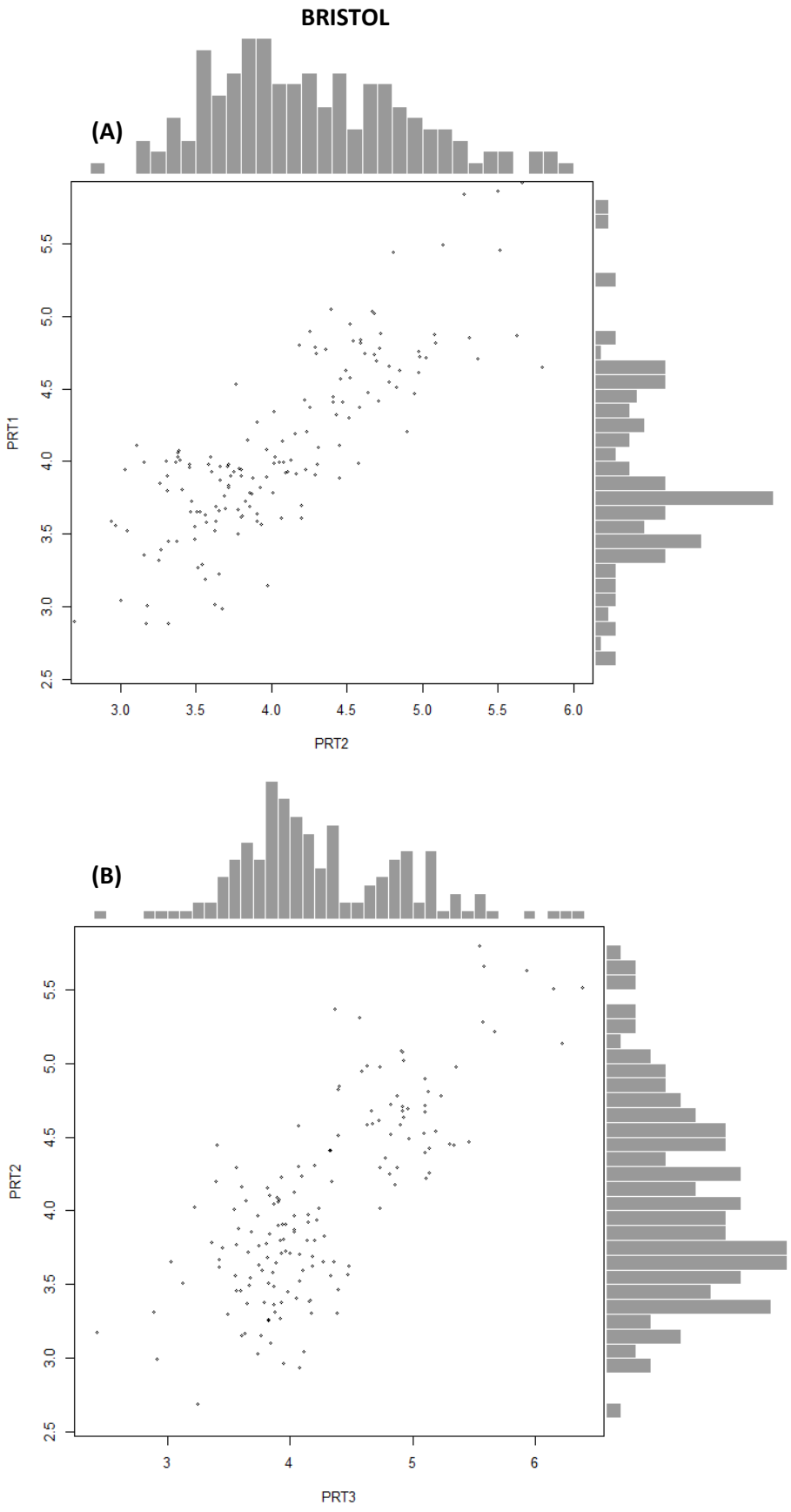


Figure 5.47: The comparison of raw PRT data for mixed. **(A)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT2 data in the late-onset Alzheimer's disease cohort (LOAD) for mixed samples. **(B)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT2 data and raw data normalized to known copy number of PRT3 data in the LOAD cohort for mixed samples. **(C)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT3 data in the LOAD cohort for mixed samples.

According to the results (Figure 5.47), PRT1 gives still noisier data than PRT2 and PRT3 meaning that PRT1 does not give discrete histograms and good clustering for copy number calling in the mixed samples of the LOAD cohort. Therefore, PRT2 and PRT3 are still better to assess the copy number for LCR1 in the LOAD cohort. For mixed samples, PRT1-3 gives good clustering for copy number calling.


BRISTOL
(A)
PRT1
PRT2
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
(B)
PRT2
PRT3
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6
5
4
3

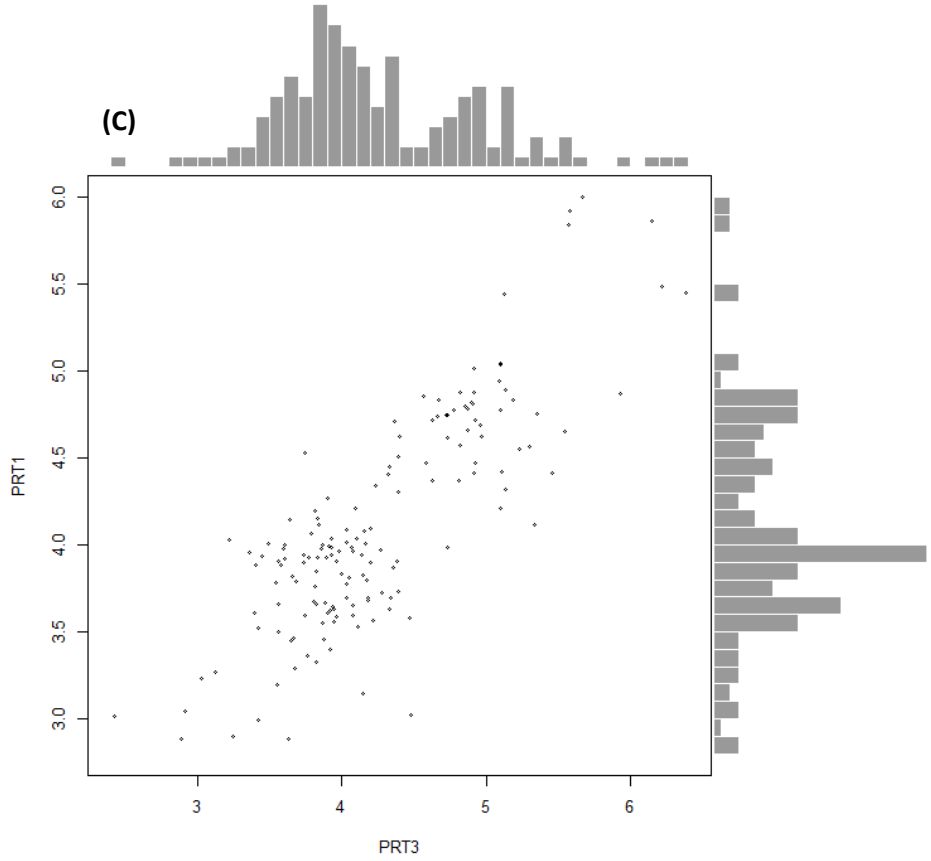
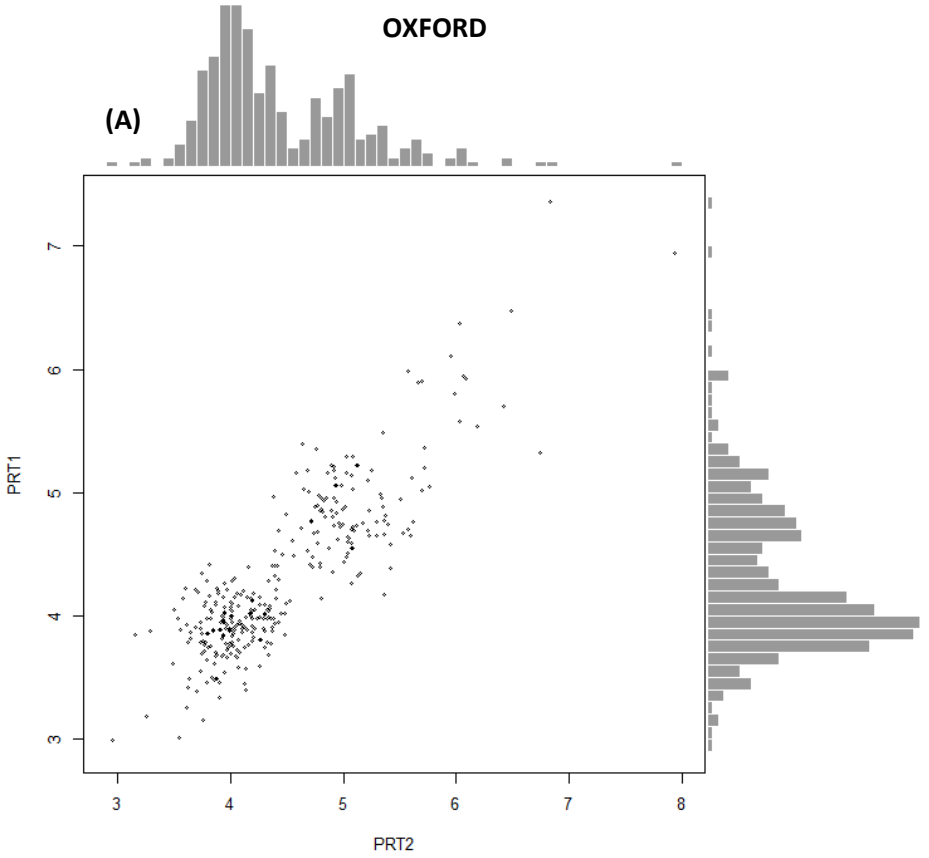
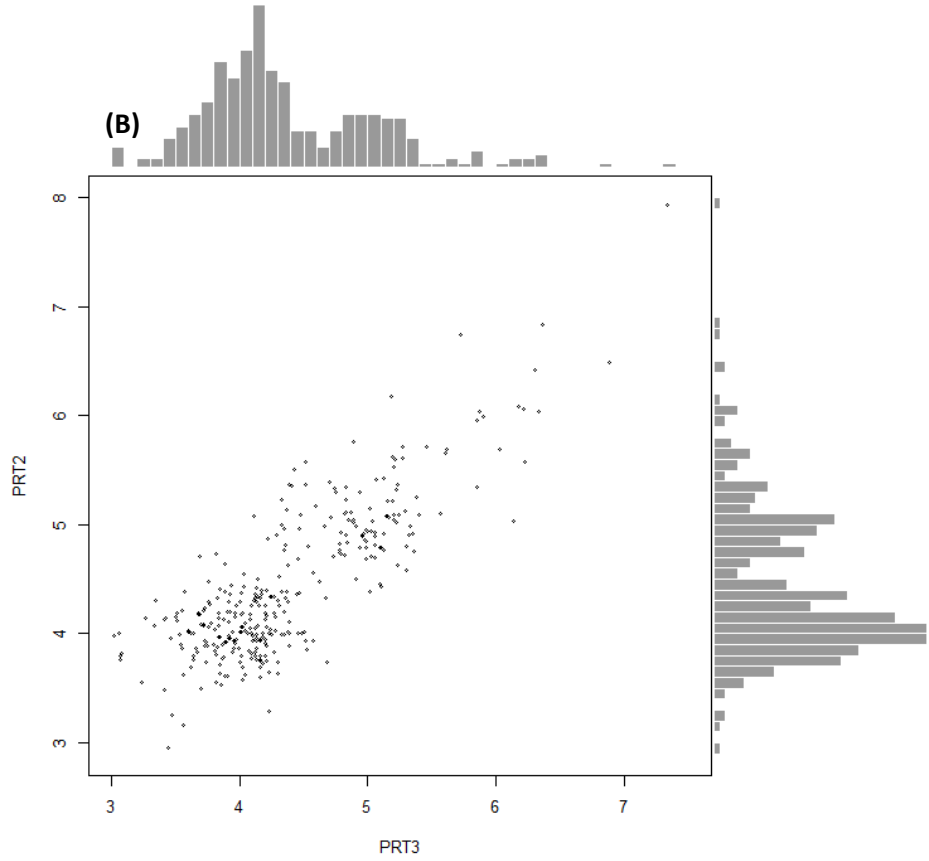


Figure 5.48: The comparison of raw PRT data for Bristol. **(A)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT2 data in the late-onset Alzheimer's disease cohort (LOAD) for Bristol samples. **(B)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT2 data and raw data normalized to known copy number of PRT3 data in the LOAD cohort for Bristol samples. **(C)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT3 data in the LOAD cohort for Bristol samples.

According to the results (Figure 5.48), PRT2 gives noisier data than PRT1 and PRT3 meaning that PRT2 does not give discrete histograms and good clustering for copy number calling in Bristol samples of the LOAD cohort. Therefore, PRT1 and PRT3 are better to assess the copy number for LCR1 in the LOAD cohort. For Bristol samples, PRT1-3 gives good clustering for copy number calling.


OXFORD
(A)
PRT1
PRT2
3
4
5
6
7
8


(B)
PRT2
PRT3
3
4
5
6
7
8

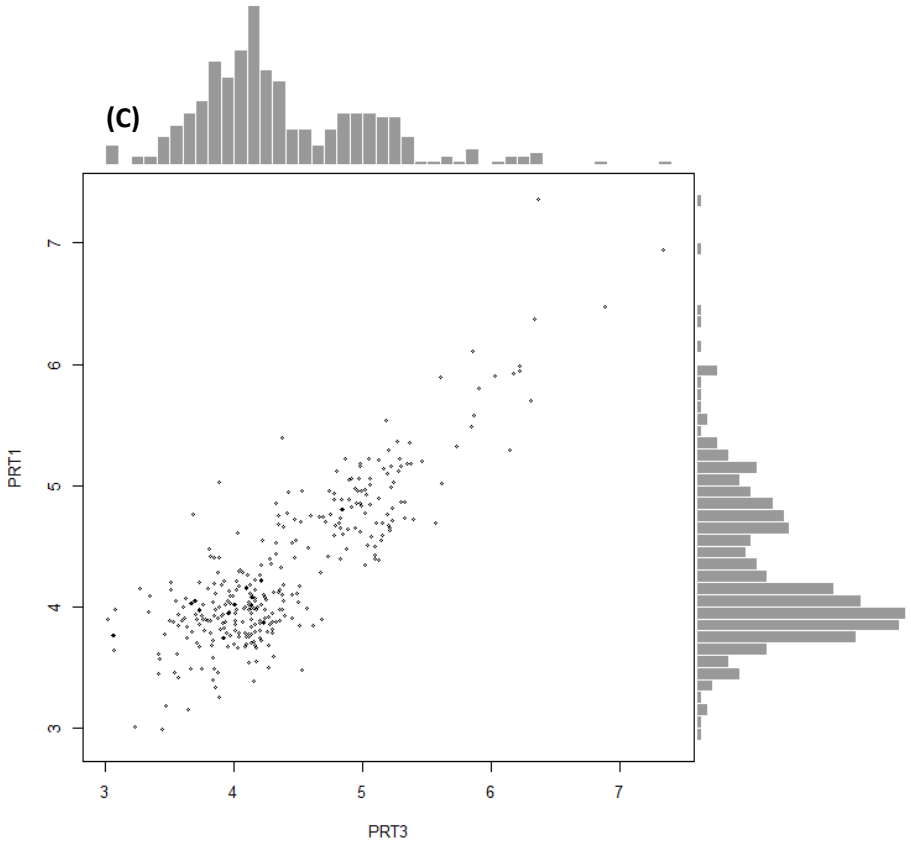
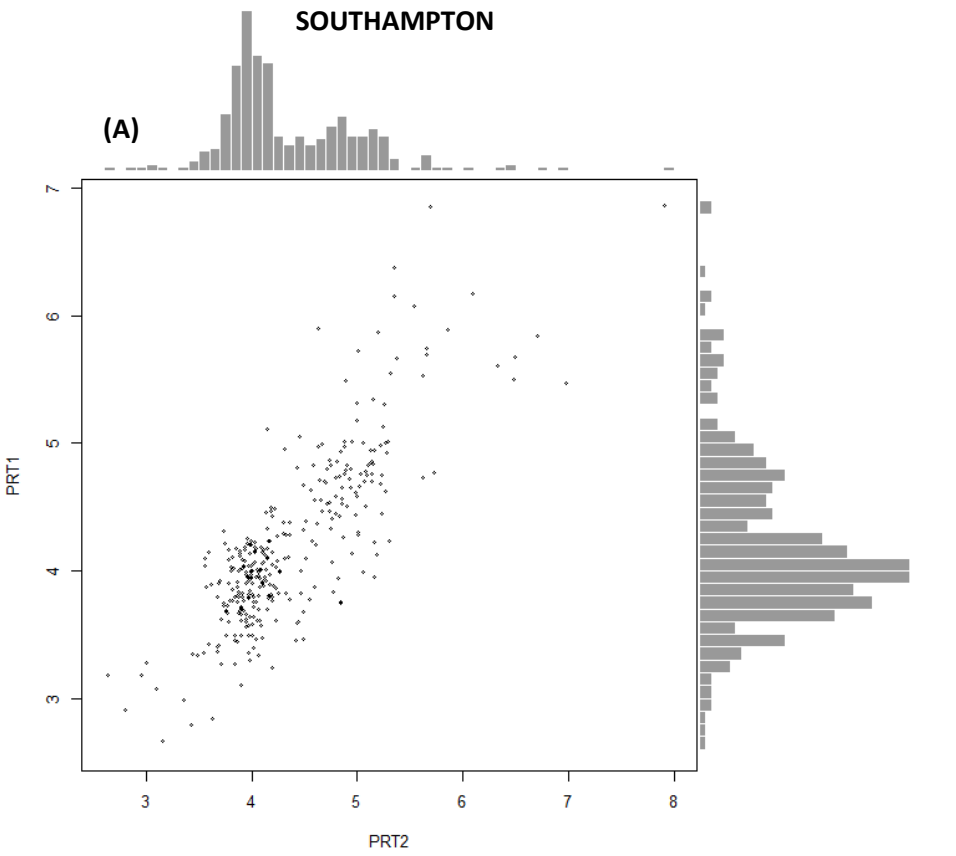
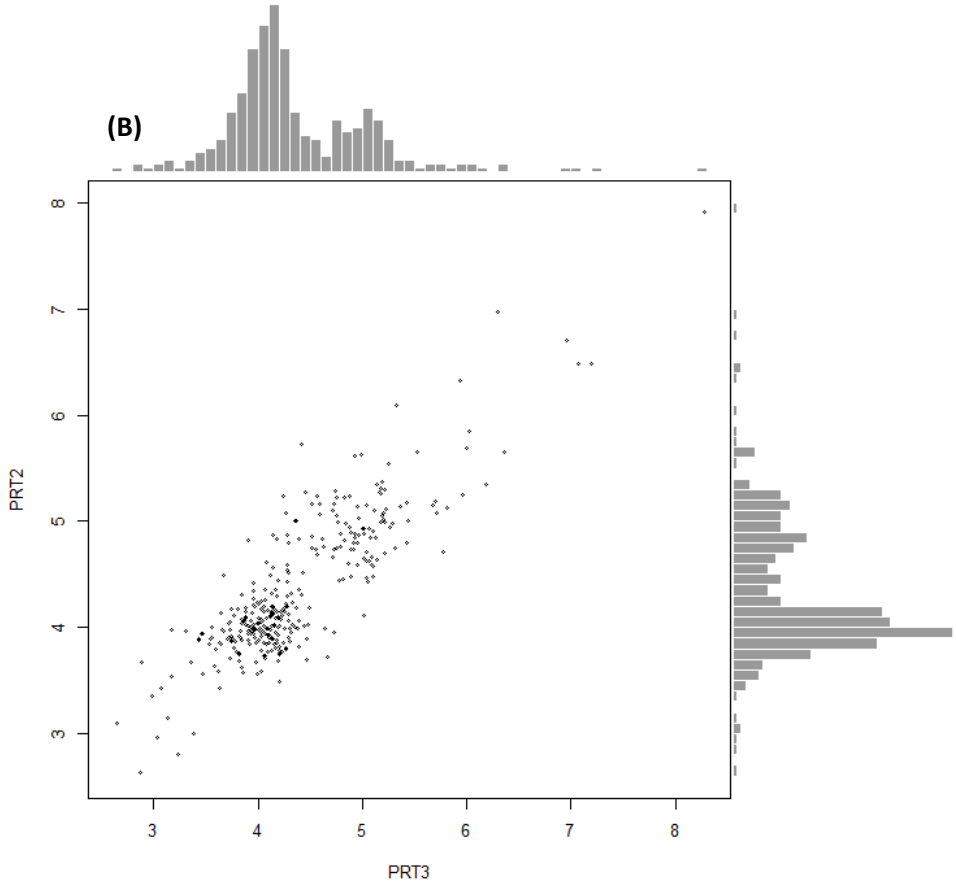


Figure 5.49: The comparison of raw PRT data for Oxford. **(A)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT2 data in the LOAD cohort for Oxford samples. **(B)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT2 data and raw data normalized to known copy number of PRT3 data in the LOAD cohort for Bristol samples. **(C)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT3 data in the LOAD cohort for Bristol samples.

According to the results (Figure 5.49), all PRTs give discrete histograms and good clustering for copy number calling in Oxford samples of the LOAD cohort.


SOUTHAMPTON
(A)
PRT1
PRT2
3
4
5
6
7
8


(B)
PRT2
PRT3
3
4
5
6
7
8

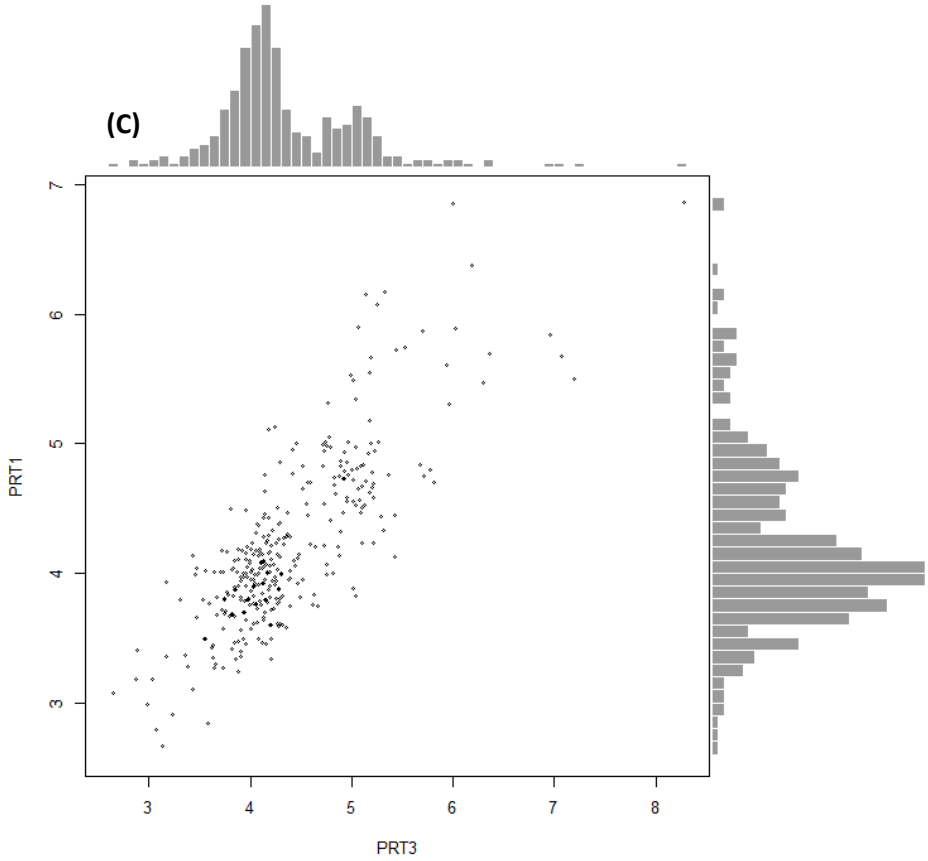


Figure 5.50: The comparison of raw PRT data for Southampton. **(A)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT2 data in the LOAD cohort for Southampton samples. **(B)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT2 data and raw data normalized to known copy number of PRT3 data in the LOAD cohort for Southampton samples. **(C)** The scatter plot shows the comparison between raw data normalized to known copy number of PRT1 data and raw data normalized to known copy number of PRT3 data in the LOAD cohort for Southampton samples.

According to the results (Figure 5.50), all PRTs give discrete histograms and good clustering for copy number calling in Southampton samples of the LOAD cohort.

Therefore, three different PRTs can be used to decide the copy number of *CR1* (LCR) because the DNA quality is variable between locally different samples; PRTs show differences in the copy number assessment.

The association study is done according to the first approach (all). The data obtained from 2185 individuals of the LOAD cohort were used and 1180 of these samples are cases and 1005 of them are controls. This approach was chosen because of the high samples size and giving high Q value.

5.3.5 Association of Copy Number Variation of CR1 in LOAD Cohort for All Samples

The aim in this section is to investigate the association between *CR1*-B and LOAD in this cohort. In order to understand the contribution of *CR1*-B to LOAD, SPSS was used to implement a generalized linear model with logistic regression. The details about the *CR1*-B allele count for cases and controls are shown below in Table 5.46.

Table 5.46: The *CR1*-B allele count for cases and controls was assessed according to the diploid copy number of LCR *CR1* obtained from PRT1-3 assays in the LOAD cohort. The data were used for the association study between *CR1*-B allele and LOAD.

Diploid Copy Number (ALL)	**Genotypes**	**Copies Duplication**	**Cases**	**Controls**
>6	*CR1*-B/B and higher	2	69	45
5	*CR1*-B/C	1	359	280
<4	*CR1*-A/A and lower	0	752	680

Table 5.47: The results from the logistic regression association study of *CR1*-B in the LOAD samples show an association between *CR1*-B and LOAD.

Parameter Estimates							
Parameter	**B**	**Standard Error**	**95% Wald Confidence Interval**		**Hypothesis Test**		
			Lower	**Upper**	**Wald Chi-Square**	**df**	**Significance**
Male (0)	-0.149	0.0927	-0.331	0.033	2.578	1	0.108
Female (1)	0[a]	.	.	.	.	.	.
***ApoE*ε4_allele**	1.180	0.0812	1.020	1.339	211.111	1	$>2x10^{-16}$
Age-onset (LOAD)	0.024	0.0046	0.015	0.033	27.246	1	$1.79x10^{-7}$
***CR1*-B**	0.155	0.0787	0.001	0.310	3.905	1	0.048

Dependent Variable: Cohort (Case/Control), **Factor:** Sex and **Covariates:** Sex, ApoEε4_allele (count), age-onset and CR1-B

The results above suggested (Table 5.47) that there was an association between *CR1*-B and LOAD (p=0.048). In addition, APOE-ε4 allele and age of onset are also associated with LOAD. The result was highly significant for APOE-ε4 allele ($p=>2x10^{-16}$) and age of onset ($p=1.79x10^{-}7$) and it is already known that APOE-ε4 allele is a risk allele for LOAD as shown in previous studies (Brouwers et al., 2012; Farrer et al. 1997). APOE-ε4 allele and age of onset were used as covariates for the rest of the analysis to find any probable risk factor for LOAD in order to make the study stronger. Sex was included in the model as covariate but does not show any association with LOAD.

5.3.6 Association of Copy Number Variation of CR1 in LOAD Cohort for All Samples with supplementary LNA results

The association study was repeated according to the second approach (PRT and junction fragment PCR assay), using the same data (PRT) obtained from 2185 individuals of the LOAD cohort, of which 1180 are cases and 1005 are controls. The main aim was to investigate the association between *CR1*-B and LOAD in this cohort. In

order to understand the contribution of *CR1*-B to LOAD, SPSS was used to implement a generalized linear model with logistic regression. Differing from the first approach, the LOAD samples were typed with using a junction fragment PCR assay (3.5 Breakpoint Analysis for *CR1* (LCR) Duplication).The DNA samples which were used for the PRT assays were limited (20 ng of DNA), so 316 of the samples were not be able to repeated for all PRT assays; and also there was not enough DNA to retype some of the samples which gave the diploid copy number of ~4.5. In order to a make more accurate copy number calling for the LOAD samples, a junction fragment PCR assay was generated, with which the LOAD cohort was typed to ensure an accurate discrimination especially between *CR1*-B and *CR1*-A alleles.

Therefore, the junction fragment PCR assay was used in addition to PRT in order to find the existence of *CR1*-B allele and both of the results were combined. PRT data which did not accurately complement the LNA data were adjusted accordingly to do so (Table 5.48). For example, for a sample that gives a diploid copy number of 5 but fails to show the existence of *CR1*-B allele using a junction fragment PCR assay, the copy number is taken as 4 (A/A) instead of 5, whereas for a sample giving a diploid copy number of 4 but showing the existence of the *CR1*-B allele using a junction fragment PCR assay the copy number is taken as 5 (B/A) instead of 4.

Table 5.48: The *CR1*-B allele count for cases and controls was assessed according to the diploid copy number of LCR *CR1* obtained from PRT1-3 assays and combined with the results from the junction fragment PCR assay in the LOAD cohort. The data were used for the association study between *CR1*-B allele and LOAD.

Diploid Copy Number (LNA)	**Genotypes**	**Copies Duplication**	**Cases**	**Controls**
>6	*CR1*-B/B and higher	2	69	45
5	*CR1*-B/C	1	342	255
<4	*CR1*-A/A and lower	0	769	705

Table 5.49: The results from the logistic regression association study of *CR1*-B in LOAD samples show a stronger association between *CR1*-B and LOAD when the junction fragment PCR assay is used to improve *CR1*-B calling.

Parameter Estimates							
Parameter	**B**	**Standard Error**	**95% Wald Confidence Interval**		**Hypothesis Test**		
			Lower	**Upper**	**Wald Chi-Square**	**df**	**Significance**
Male (0)	-0.149	0.0928	-0.331	0.033	2.571	1	0.109
Female (1)	0[a]	.	.	.	.	.	.
***ApoE*ε4_allele**	1.180	0.0812	1.021	1.339	211.019	1	$>2x10^{-16}$
Age-Onset (LOAD)	0.024	0.0046	0.015	0.033	27.615	1	$1.48x10^{-7}$
***CR1*-B**	0.193	0.0794	0.037	0.349	5.905	1	0.015

The results above (Table 5.49) suggest that there is a stronger association between *CR1*-B and LOAD than the first approach. According to the results the p-value was significant (p=0.015) for *CR1*-B and LOAD. In addition, results are showing robust effects of age of onset (p=1.48x10-7), APOE-ε4 allele (p=>$2x10^{-16}$) and *CR1*-B on LOAD. The size effect for the *CR1*-B allele was 0.193 which leads to a hazard ratio of 1.213. Therefore, the presence of the *CR1*-B allele increases the risk of LOAD by 1.213 compared to its previous value.

The size effect for the APOE-ε4 allele was 1.180 which leads to a hazard ratio of 3.254. Therefore, the presence of the APOE-ε4 allele increases the risk of LOAD by 3.254 compared to its previous value which is approximately a 2.7 times higher risk factor than the *CR1*-B allele for LOAD.

5.3.7. Linkage Disequilibrium Analysis for CR1 gene variants

Previously, GWAS studies showed that *CR1* gene variants are associated with Alzheimer's disease (section 1.7.3 GWAS Results implicating Polymorphisms Surrounding CR1 and Erythrocyte Sedimentation Rate (ESR). In this study, these variants were shown to be in LD with *CR1*-B and *CR1*-C in various populations but especially in LOAD. Our recent discovery of an association between *CR1*-B and LOAD is particularly

important when understanding the potential risk loci for LOAD and their relation to *CR1*-B. The results showed that these SNPs are in LD with *CR1*-B but in low LD with *CR1*-C in CEU and LOAD, whereas they were in low LD with both *CR1*-B and *CR1*-C in JPT/CHB (except rs6656401 with *CR1*-B) and in YRI (except rs3818361 and rs6701713 with *CR1*-B). Therefore, studying these alleles to find the potential risk loci for AD is possibly better in European populations (higher LD). As we did not have all the genotyping data for the LOAD cohort for all SNPs (SNP data were available for only for 400 samples for two SNPs: rs3818361 and rs6701713) we were unable to conduct a risk allele association study for SNPs with LOAD (Table 5.50).

Table 5.50: Linkage disequilibrium analysis in CEU, JPT/CHP, YRI (HapMap Phase 1) and LOAD. The GWAS studies related to the SNPs: **(1)** Lambert et al., 2009; **(2)** Hollingworth et al., 2011; **(3)** Naj et al., 2011; -=no data available and N=Number of samples. Haplo=Haplotype and Freq=Frequency; the samples which were homozygotes for the *CR1*-B allele are coded as AA (variant could be T/A) whereas homozygotes for the *CR1*-C allele are coded as GG (variant could be C/G) during LD analysis.

Variant	**Disease**	**Risk Allele**	**LD with *CR1*-B / *CR1*-C**	**CEU N=60 (unrelated)**		**JPT/CHB N=89**		**YRI N=60 (unrelated)**		**LOAD (this study) N=400**	
				R^2	**D'**	**R^2**	**D'**	**R^2**	**D'**	**R^2**	**D'**
rs3818361	Alzheimer's Disease (1)	A	*CR1*-B (A) A/T	**0.663**	0.814	**0.053**	1.000	**0.317**	0.917	**0.514**	0.792
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**
				AA	0.223	**AA**	0.023	**AA**	0.218	**AA**	0.160
				TA	0.036	**TA**	0.284	**TA**	0.222	**TA**	0.063
				AG	0.036	**AG**	≤0.0001	**AG**	0.011	**AG**	0.030
				TG	0.706	**TG**	0.693	**TG**	0.549	**TG**	0.745
			CR1-C (G) (G/C)	0.016	1.000	0.002	0.667	0.081	1.000	$8.21x10^{-4}$	0.385
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**
				GA	≤0.0001	GA	0.001	GA	≤0.0001	GA	0.00259
				CA	0.259	CA	0.306	CA	0.441	CA	0.221
				GG	0.043	GG	0.010	GG	0.093	GG	0.016
				CG	0.6698	CG	0.683	CG	0.466	CG	0.760
rs6656401	Alzheimer's Disease (1)(2)	A	*CR1*-B(A) A/T	**0.658**	0.850	**0.795**	1.000	**0.021**	0.376	-	-
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**		
				AA	0.214	**AA**	0.023	**AA**	0.022		
				TA	0.027	**TA**	0.006	**TA**	0.020		
				AG	0.044	**AG**	≤0.0001	**AG**	0.207		
				TG	0.714	**TG**	0.972	**TG**	0.751		

			CR1-C (G) (G/C)	0.014	1.000	≤0.0001	1.000	0.032	0.273	-	-
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**		
				GA	≤0.0001	GA	≤0.0001	GA	0.014		
				CA	0.241	CA	0.028	CA	0.028		
				GG	0.043	GG	0.011	GG	0.079		
				CG	0.716	CG	0.960	CG	0.879		
rs6701713	Alzheimer's Disease (3)	A	*CR1*-B(A) A/T	**0.663**	0.814	**0.053**	1.000	**0.321**	0.917	**0.515**	0.792
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**
				AA	0.223	**AA**	0.023	**AA**	0.214	**AT**	0.159
				TA	0.036	**TA**	0.284	**TA**	0.217	**TT**	0.063
				AG	0.036	**AG**	≤0.0001	**AG**	0.011	**AC**	0.0306
				TG	0.706	**TG**	0.693	**TG**	0.558	**TC**	0.747
			CR1-C (G) (G/C)	0.016	1.000	0.002	0.667	0.079	1.000	1.14×10^{-3}	0.442
				Haplo	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**	**Haplo**	**Freq**
				GA	≤0.0001	GA	0.001	GA	≤0.0001	GT	0.00248
				CA	0.259	CA	0.306	CA	0.431	CT	0.22
				GG	0.043	GG	0.010	GG	0.095	GC	0.0175
				CG	0.698	CG	0.683	CG	0.474	CC	0.76

5.4 Discussion

A recent study of Brouwers et al. (2012) showed that the presence of *CR1*-B increases the number of C3b/C4b and cofactor activity sites and Alzheimer's disease (AD) risk with 30% in *CR1*-B carriers in the French LOAD cohort. This study suggested that *CR1*-B may inhibit the complement activation more efficiently than *CR1*-A due to extra C3b/C4b binding sites. However, this mechanism is conflicting with the fact that AD is associated with increased complement activation (Rogers et al., 1992; Jones et al., 2010). *CR1*-B is also linked to increased complement activation due to the differing protein expression levels between *CR1*-B and *CR1*-A isoforms (Hazrati et al., 2012). The studies related to *CR1*-B are explained in more details below.

The potential role of *CR1*-B in LOAD based on the previous study (Brouwers et al., 2012) was revisited in this research with a larger LOAD case-control cohort (2185 individuals: 1180 cases and 1005 controls). Previous findings on *CR1*-B association (Brouwers et al., 2012) with LOAD were confirmed in this research. An association between *CR1*-B allele and LOAD was found (p=0.048). In the presence of an extra assay (junction fragment PCR assay, section 2.3.3 Touchdown PCR for *CR1* Breakpoint Analysis) the results become more significant (p=0.015). The entire LOAD cohort was typed using this assay to test for the *CR1*-B allele. PRT can be used on its own to call the copy number, but extra typing with LNA primers gives a more accurate copy number call. For example, a sample which gave a raw copy number of 4.5 with PRT, after an addition of typing with LNA primers can be understood to contain 4 copies (C/A - less likely because of the *CR1*-C allele frequency, or A/B) or 5 copies (C/D - less likely because of the C allele frequency, or A/B) in order to detect the presence of the *CR1*-B allele.

Similar results which implied a putative role of *CR1*-B in LOAD were observed in previous studies. Fonseca et al. (2016), have recently tested anti-CR1 reactivity in the brain (23 AD cases) and also the CR1 ligands (C1q and C3b) binding to CR1, which was isolated from the erythrocytes of control and AD patients, with or without GWAS-reported AD risk loci rs6656401 and rs4844609. The main reason for the investigation was to determine if these loci or diagnosis would have an effect on CR1 distribution. They found that CR1 distribution in the brain is not correlated with the AD risk loci

rs4844609 and rs6656401, and also that CR1 ligand binding to CR1 is moderately modified by rs4844609 or *CR1*-B (rs6656401) or diagnosis.

In addition, Fonseca et al., (2016) have observed a small but significant increase in soluble CR1 concentration in plasma that correlated with the presence of either one or both AD loci compared to controls. Also, there was a slight difference in the cleavage efficiency of CR1 between CR1-A and CR1-B alleles by the presence of AD risk loci. The cohort however was relatively small, possibly compromising the effectiveness to detect a link between plasma-soluble CR1 levels and dementia (Fonseca et al., 2016).

In another study, CR1 length variants were assessed at protein and gene levels revealing that the CR1 density was significantly lower in LOAD patients expressing the CR1-B isoforms compared with the controls ($p=0.001$) indicating lower expression of *CR1* in CR1-B carriers (Mahmoudi et al., 2015). Here, rs6656401 and rs3818361 were significantly associated with CR1 length variants ($p<0.0001$) and CR1-B associated with AD susceptibility (lower density) suggesting that AD may result from insufficient clearance of plaque deposits instead of increased inflammation (Mahmoudi et al., 2015). Caucasian samples of 100 LOAD cases and 35 controls were used for this study.

In a separate study by Hazrati et al. (2012), individuals with the A/B (*CR1*) genotype were shown to have 1.8 times increased risk for LOAD compared to individuals with the A/A (*CR1*) genotype ($p=0.003$) while rs4844610 (*CR1* variant) was slightly significant ($p=0.0024$). This study used the Canadian (case-control study) cohort consisting of 475 unrelated cases with LOAD and 284 unrelated controls. The analysis of brain samples revealed that CR1-B isoforms are expressed at lower protein level than CR1-A, and therefore likely associated with increased complement activation. According to the neuropathological results, the pattern of CR1 expression in neurons is different between A/A and A/B genotypes. The study also observed different distribution of CR1 in neurons suggesting that CR1-A isoforms are shifting through protein sorting compartments such as the endoplasmic reticulum-Golgi system whereas CR1-B isoforms accumulate in the membrane of cytoplasmic vesicles. In A/A cases CR1 colocalizes in the endoplasmic reticulum whereas in A/B cases CR1 is mostly detected

in lysosomes. This suggests a different trafficking of the secreted form of the CR1-A and CR1-B isoforms in neurons (the length of extra-cellular domain can have an effect on protein release (Hamer et al., 1998; Hazrati et al., 2012). These results show that more research is needed in order to reach a better understanding of CR1-B's role in LOAD for future studies.

No studies have been carried out to seek an association between early-onset Alzheimer's disease and CNV of *CR1*. In this study, our aim was to understand the role of *CR1*-B in early and late onset Alzheimer's disease. Due to the stronger phenotype of the EOAD we were expecting to find an association between CR1 and EOAD more easily than for LOAD and *CR1* association. Our study shows no association between LCR *CR1* copy number variation and EOAD. Our results have shown that AD risk loci (rs3818361, rs6656401 and rs6701713) within the *CR1* gene were not risk factors for EOAD and they were in moderate LD with the *CR1*-B. Our results suggest that these LOAD risk loci may not be even related with early-onset Alzheimer's disease. The reason for these negative results can be the cohort size of our study resulting in the power of the study being too low to detect any significant effect of CR1 on the disease. Also, early and late onsets of Alzheimer's disease have difference in progression of the disease that may have different causes (Rogaeva, 2002).

6 DISCUSSION

My PhD thesis investigated the diversity and characterization of copy number variation of the complement C3b/C4b receptor 1 (*CR1*) gene across various populations from around the globe. My thesis also explored the association of CNV of the *CR1* to malaria and Alzheimer's disease using case-control studies.

The main objective of my thesis was to characterize and measure CNV at the *CR1* gene. In order to do this, three PRT (1-3) assays were designed to measure the diploid copy number of LCR *CR1*. Previously, the four alleles of the *CR1* gene and their frequencies were studied by RFLP and Southern blotting (Hollers et al., 1987; Vik and Wong, 1993; Wong et el., 1991; Wong et el, 1986). However, these methods were not practicable to use in large case-control studies. Therefore, I saw a need to develop a robust method to genotype LCR *CR1* CNV in large cohorts. This led me to develop and implement three PRT (1-3) assays. The HGDP and HapMap phase 1 samples (Utah residents with ancestry from Northern and Western Europe, Yoruba of Ibadan, Japanese from Tokyo, and Han Chinese from Beijing) were typed using PRT assays (1-3) which target the CNV region (LCR) in the *CR1* gene, to study the *CR1* CNV pattern across worldwide populations. A range of LCR *CR1* copy number of 2-9 was found per diploid genome in HGDP and HapMap phase 1 samples. The allelic architecture and copy number genotype for LCR *CR1* CNV was estimated from diploid copy number using the CoNVEM program (Gaunt et al., 2010). The observed and expected allele frequencies were very similar for LCR *CR1* CNV. PRT assays are robust in accurately calling CNVs of *CR1* (LCR) to estimate CNVs of *CR1* in large disease association (case-control) studies.

The comparison of the PRT data which were obtained from the samples of HapMap phase 1, to the aCGH (Conrad Agilent aCGH data) and the sequencing results (sequencing read depth from 1000 Genomes data), showed that PRTs are giving matching CN data with sequencing data while showing differences to the aCGH results. There was a noted difference between PRT copy number data and the aCGH data. The reported copy number estimation based on the aCGH assay showed inaccurate copy number prediction of LCR *CR1* as they have shown the *CR1*-C allele frequency to be the second most common *CR1* allele (more common than *CR1*-B allele) in the population. Other studies in Caucasian, African-American, Chinese, Mexican, Asian Indian and

Indian populations have shown that *CR1*-A is the most common allele whereas *CR1*-B allele is the second commonest allele and *CR1*-C and *CR1*-D alleles are rarer (Dykman et al., 1985; Moulds et al., 1996; Moulds et al., 1998; Katyal et al. 2003; Panchamoorthy et al., 1993). In our PRT results, we have shown that the *CR1*-A (2 copies) allele is the most common allele and the *CR1*-B (3 copies) is the second most common allele (except in the Americans of the HGDP) whereas *CR1*-C and *CR1*-D alleles are rarer in all populations. The PRT assays gave good interpretation and estimation of the *CR1* (LCR) copy number. The results of copy number calling by PRTs are in agreement with the literature as the allele frequencies of *CR1* matched with the previous *CR1* population studies (Dykman et al., 1985; Katyal et al., 2003; Moulds et al., 1996). Therefore, these PRTs can be used in case-control studies in order to find probable association of *CR1* to other diseases such as malaria and Alzheimer's disease.

Differences in *CR1* allele frequencies can arise through genetic drift. The *CR1*-B allele particularly showed lower frequencies in the Asian-Indian, Chinese and Indian populations (Katyal et al. 2003; Moulds et al., 1998; Panchamoorthy et al., 1993) and higher in populations such as Caucasians, African-Americans, Mexicans, Europeans and Africans (Moulds et al., 1996; Dykman et al., 1985; this study). A study by Madi et al. (1991) showed that erythrocytes were more efficient in binding immune complexes when carrying *CR1*-B/D alleles compared to those having *CR1*-C or *CR1*-A structural alleles of *CR1*. Therefore, having the *CR1*-B allele (or *CR1*-D) may be favoured in some of these populations due to selective pressure of the infectious diseases place upon these populations. This theory of selection of the duplication (*CR1-B*) for malaria was tested in HGDP samples via the pathogen diversity index. The results showed that there was no correlation between protozoa (including malaria) diversity, or that of viruses or helminths and duplication frequency of *CR1* but bacterial diversity showed correlation with duplication (*CR1*-B) meaning that populations that experience a high diversity of bacterial pathogens show a high frequency of *CR1*-B. Therefore, bacterial pathogens may be driving *CR1*-B allele frequency change, but not malaria. The association between copy number variation of LCR *CR1* and malaria and also Knops blood group SNPs and malaria were also investigated in this thesis.

In this thesis, the breakpoint for *CR1*-B allele in European populations is confirmed with our LNA primers designed based on the suggested breakpoint position by Vik and Wong (1993). That study showed that the breakpoint for *CR1*-B generation is in exon 6 of the *CR1*-A allele. The primers may not work accurately in African populations as the LNA primers are specifically designed for European samples (SNPs specific to African populations may affect the binding efficiency of primers) and there may be another breakpoint for the *CR1*-B allele generation in African populations. This assay can also be used with the PRT1-3 assays to genotype LCR *CR1* CNV in large disease cohorts in European samples. Therefore, we used PRT1-3 assays to genotype LCR *CR1* CNV in the HGDP panel, HapMap phase 1 samples, HRC1 (only 40 samples), EOAD, LOAD and the Tori-Bossito cohorts.

In this thesis, the PRT1-3 assays are used for accurate genotyping *CR1* LCR CNV in various populations. In a small number of the African samples from the HGDP (1.13%) and the HapMap3 (2.1%), the PRT3 assay has shown different *CR1* LCR copy numbers from the PRT1 and PRT2 assays. This can be a result of mutation or polymorphism in the binding region of PRT3 primers or a small deletion in the *CR1* or duplication in the *CR1L* gene, and therefore further investigations are needed. Alternatively, sequencing could be used around the primer binding regions in both *CR1* and *CR1L* for future studies.

Due to the association of *CR1* with malaria being suggested by several studies (shown below), the role of Knops blood group-determining antigens (SNPs) and LCR *CR1* CNV was investigated in this thesis. Several studies tested SNPs including those involved in the Knops blood group and a SNP affecting *CR1* expression levels on the erythrocyte surface but none of these studies looked at the association between *CR1* CNV and malaria. Therefore, I have tested a possible association between CNV of LCR *CR1* and malaria. I did not find any association between *CR1* CNV and malaria. This study could potentially be conducted with a much larger malaria cohort in order to discover the slighter effect of *CR1* CNV for future studies.

The P. *falciparum* adhesin PfRh4 binds to the CR1 receptor on erythrocytes (inhibits the decay-accelerating activity of CR1) therefore it can be described as a receptor for

parasite invasion (Tham et al., 2011). Rosetting is associated with several diseases by blocking the microvasculature of vital organs such as the brain (Rowe et al., 1997). The key for rosetting lies in the CPPs 15-17 region of CR1 (Tetteh-Quarcoo et al., 2012). Erythrocytes which were exhibiting the Sl:2 phenotype showed a reduced binding to the parasite rosetting ligand PfEMP1. Rowe et al. (1997) have shown that low CR1 expression and the *CR1* polymorphism (Sl(a^-) or Sl2)) are associated with reduced rosetting of erythrocytes (Rowe et al., 1997). Erythrocytes infected with *P. falciparum* form rosettes with infected and un-infected erythrocytes. This then raised the question of the Sl2 allele's relation to malaria. Sl2 presents in ~70% of Africans whereas this proportion is 40-50% for African-Americans and less than 1% in Europeans (Daniels, 2013). Therefore, having the Sl2 allele may confer an advantage in *P. falciparum* endemic areas. According to previous studies the presence of McC^b and Sl2 alleles are associated with resistance to *M. tuberculosis* infections in Mali (Nounsi et al., 2011) and homozygosity for McC^b gives protection against leprosy in Malawi (Fitness et al., 2004) but the relationship between *CR1* and malaria is not simple.

We analysed the distribution of Swain-Langley (rs17047661) across global populations and a cohort of 563 infants from Benin (Tori-Bossito) followed since birth to observe malaria acquisition and treatment. Our study of Tori-Bossito plus studies of Brazil (Fontes et al., 2011) and Kenya (Thathy et al., 2005) revealed an association between malaria and *CR1* alleles encoding Sl2 whereas Gambia (Zimmerman et al., 2003), Ghana (Hansson et al., 2013), India (Gandhi et al., 2009) and our study of Tolimmunpal studies did not show an association between the Sl2 allele and malaria. There are appreciable differences between the studies such as cohort sizes and age differences which may explain conflicting results between them. The malaria study in the Kenyan population showed that children with the Sl2/2 genotype were less likely to have cerebral malaria than those with Sl1/1 (Thathy et al., 2005). Combination alleles and ethnic groups were used as covariates in Kenyan study (Thathy et al., 2005) whereas the association study from Gambia (Zimmerman et al., 2003) used variants as haemoglobin genotypes, sex, ethnic group and location of residence and the association study from Ghana (Hansson et al., 2013) used the variants as parasitaemia, age and sex as covariates. Gambian (n=483) and Ghanaian (n=146) studies used more

covariates than the Kenyan (n=274) study meaning that the results are more reliable. However, the samples' age ranges were significantly different from each other. The Kenyan study used children with ages of 14 months whereas Gambia used older children with an average age of 3.78 and Ghana used children with ages between 0-15 years. According to World Health Organization (WHO) it was reported that newborns and infants less than 12 months of age (WHO, 2016b) are one of the most vulnerable groups affected by malaria and they are at risk of rapid disease progression, severe malaria and death. Severe malaria is specifically common in this age group and after the age of five the risk of malaria decreases. Therefore, the Kenyan study with 14-month old infants provides a better approach to detect genetic factors which give protection against malaria as the risk of having malaria decreases by age especially after the age of 5. Our study of Tori-Bossito is a bigger cohort (563) than all other studies mentioned and the infants were actively monitored clinically, parasitologically and immunologically from birth till the age of 18 months. An allele-specific hybridization assay (ASO) was used to genotype alleles of the Knops blood group system. I found that the Sl2 allele (rs17047661-G) appears to provide protection against early acquisition of malaria and subsequent number of malarial infections. My study used the most diverse covariates compared to the other studies. The covariates used for this study are the mother's age, malaria suspicion during pregnancy, sex, birth term, mosquito net, ethnic group, sickle cell, chloroquine intake during pregnancy and Knops blood group SNPs. Therefore, our Tori-Bossito study is the most reliable among these studies. The Tori-Bossito study was repeated with the Tolimmunpal study in order to confirm the Sl2 allele association with malaria. For this study, the genotyping data of Knops blood group SNPs were provided by David Courtin and André Garcia (not obtained by ASO assay). In this cohort, the Sl2 allele did not show protection against early acquisition of malaria and subsequent number of malarial infections. The Tolimmunpal cohort is smaller and does not have as many covariates as the Tori-Bossito study. The Tolimmunpal study did not have any information about chloroquine intake during mother's pregnancy which could easily affect the association studies. For example, if the mother had placental malaria and had taken chloroquine, it would mask the real effect of the genetic variants. We fail to confirm the protective role of Sl2 allele against malaria in the Tolimmunpal study. The Tori-Bossito study should be

repeated with a similar cohort and covariates in order to decide if Sl2 allele provides protection against malaria or not.

Previous study reported that rs6691117 in the *CR1* gene was associated with ESR (Kullo et al., 2011). The minor allele G of rs6691117 was associated with lower ESR (Kullo et al., 2011) and the other SNPs within or surrounding *CR1* (rs12034383, rs12034598 and rs7527798) were associated with ESR (Kullo et al., 2011; Naitz et al., 2012). We could not find any association between rs6691117 or other Knops blood group SNPs (except rs17047661) and malaria phenotypes in Tori-Bossito. These SNPs were not in high LD with either *CR1*-B or *CR1*-C of *CR1* in the HapMap populations. Therefore, even if the SNPs were associated with ESR it may not always be the case that it is due to malaria. The Sl2 allele of *CR1* was not in high LD with any other SNPs in CR1 and surrounding for the YRI population, which is likely to have the same LD pattern as Tori-Bossito and Tolimmunpal populations. Therefore, Sl2 is likely to be the true protective allele against malaria.

There was no association between the Sl2 allele and malaria prevalence in all African populations including the populations whose data were publically available. The first explanation for lack of association between malaria prevalence and the Sl2 allele may be lack of malaria prevalence data which was obtained from the Malaria Atlas Project providing malaria prevalence data only since 2000. This means that the malaria prevalence data which had driven the evolution towards protective alleles for malaria conferring its survival advantage could not be provided by the Malaria Atlas Project; therefore, it could not be applied to our analysis. The second explanation for this lack of association may be the migration between the African populations. For example, some of the South African Bantu populations (Zulu, Sotho) live in a low malaria environment but have high frequency of Sl2, perhaps because allele frequencies have been maintained at a high level by drift since the Bantu migration from malaria-endemic West Africa (Gurdasani et al., 2015). The last explanation for the lack of association can be a true lack of association between the Sl2 allele frequency and malaria prevalence within the African populations studied in this thesis.

I tested the departure from HWE for all of the Knops blood group-determining SNPs and it is not possible to interpret that these SNPs are selectively favoured because of their survival advantage against malaria. Therefore, we fail to show these SNPs and the Sl2 allele are favoured because of resistance to malaria.

More studies need to be done in order to interpret if the Sl2 allele is protective or not. The research can be replicated with a similar population to Tori-Bossito with the same quality of data.

In this thesis, the association between the *CR1*-B allele and LOAD and also between EOAD and *CR1*-B were investigated separately. The idea for this research is based on the study of Brouwers et al. (2012). Their study showed that that the presence of *CR1*-B increases the number of C3b/C4b and cofactor activity sites and Alzheimer's disease (AD) risk with 30% in *CR1*-B carriers in the French LOAD cohort. Their study suggested that CR1-B may inhibit complement activation more efficiently than CR1-A due to additional C3b/C4b binding sites. This mechanism conflicts with the fact that AD is associated with increased complement activation (Rogers et al., 1992; Jones et al., 2010). CR1-B is also linked with increased complement activation due to the differing protein expression levels between CR1-B and CR1-A isoforms (Hazrati et al., 2012).

The probable association between AD and the CR1-B allele was first tested on the EOAD cohort due to the stronger phenotype of EOAD. Additionally, no studies have been carried out to find the association between EOAD and CNV of *CR1* therefore we wanted to investigate a probable association for the first time. We fail to find any association between *CR1*-B ($p=0.755$) and EOAD but APOE-ε4 ($p=4.25x10^{-13}$) allele showed strong association with EOAD. This was expected as the presence of APOE-ε4 allele is a known risk factor for both EOAD and LOAD (Chartier-Harlin et al., 1994; Houlden et al., 1988). Our results have also shown that AD risk loci (rs3818361, rs6656401 and rs6701713) within the *CR1* gene were not risk factors for EOAD, and they were in moderate LD with *CR1*-B. These results were not unexpected, as early and late onsets of AD have differences in progression of the disease and therefore they may have different causes (Rogaeva, 2002).

The potential role of *CR1*-B in LOAD based on the previous study (Brouwers et al., 2012) was investigated in this research with a larger LOAD case-control cohort (2185 individuals: 1180 cases and 1005 controls). The previous findings on *CR1*-B's association with LOAD were confirmed in this research (p=0.048). I also found a significant association between APOE-ε4 allele (>$2x10^{-16}$) and LOAD and also with the age of onset ($1.79x10^{-7}$). It has been shown that the presence of APOE-ε4 allele lowers the age of onset of LOAD significantly (Sando et al., 2008). Therefore, we were expecting an association between the age of onset and LOAD. Because of having limited amount of DNA for the PRT1-3 assays, we were unable to obtain copy number data from all three PRT assays for 316 samples. For these samples, averages of one or two PRT assays were used to collect LCR *CR1* copy number data. This approach was not accurate. An additional assay was needed to get a better interpretation of the LCR *CR1* copy number. Therefore, we designed and used a new junction fragment PCR assay (section; 2.3.3 Touchdown PCR for *CR1* Breakpoint Analysis) for the LOAD cohort. This basic PCR method gives the information about the presence and absence of *CR1*-B allele in each sample. The entire LOAD cohort was typed with junction fragment PCR assay (with LNA primers) to test the existence of *CR1*-B allele. The junction fragment PCR assay made the association stronger between the *CR1*-B allele (p=0.015) and LOAD. We also found a significant association between the APOE-ε4 allele (p<$2x10^{-16}$) and LOAD and also with age of onset (p=$1.48x10^{-7}$). The PRT assay can be used on its own to call the LCR *CR1* copy number but this additional junction fragment PCR assay with PRT1-3 gave a more accurate and reliable copy number call therefore for a sample which gives a raw copy number of 4.5 (can be 4 copy (*CR1*-A/A) or 5 copy (*CR1*-A/B)) with the PRT, the junction PCR assay can be used to assess the presence of the *CR1*-B allele.

The putative role of CR1-B in LOAD may be explained by other studies. Mahmoudi et al. (2015) suggested that LOAD may result from insufficient clearance of plaque deposits instead of increased inflammation. They discovered that CR1 density was significantly lower in LOAD patients who express the CR1-B isoforms compared to the controls meaning that lower expression of CR1 was found in CR1-B carriers. They also showed that rs6656401 and rs3818361 (AD risk loci) were significantly associated with CR1 length variants (p<0.0001). Similarly, Hazrati et al. (2012) found that individuals

with the A/B (*CR1*) genotype had 1.8 times increased risk for LOAD compared to individuals with the A/A (*CR1*) genotype (p=0.003). In their study, the analysis of brain samples revealed that CR1-B isoforms are expressed at lower protein levels than CR1-A and therefore likely associated with increased complement activation. They also found that the pattern of CR1 expression in neurons is different between A/A and A/B genotypes. They also show different distributions of CR1 in neurons. In A/A cases CR1 colocalizes in the endoplasmic reticulum whereas in A/B cases CR1 is mostly detected in lysosomes. Because of this, they suggested that there should be a different trafficking of the secreted form of the CR1-A and CR1-B isoforms in neurons. The study could not provide strong evidence about the association between CR1-B and LOAD. Fonseca et al. (2016) found that CR1 distribution is not correlated with the AD risk loci rs4844609 and rs6656401 in the brain but they found a slight difference in the cleavage efficiency of CR1 between CR1-A and CR1-B alleles by the presence of AD risk loci. They also found a small significant increase in soluble CR1 concentration in plasma that correlated with the presence of either one or both AD loci compared to donors. Therefore, more studies are needed to interpret the potential role of CR1-B in the brain during AD.

PRT (1-3) assays which were applied successfully in this study, are robust methods that should be sufficiently accurate for measuring CNV (LCR *CR1*) in large case-control association studies in the future. My results show an association between LOAD and *CR1*-B therefore the role of CR1 in the pathogenesis of AD remains to be determined in future studies. The discovery of the protective Sl2 allele against malaria in the Tori-Bossito cohort was showing a potential association between *CR1* and malaria. More research about the protective role of Sl2 allele is needed to be conducted for future studies, as in a similar but smaller cohort we failed to confirm the protective effect of the Sl2 allele.

7 APPENDICES

Appendix 1: Lymphoblastoid Cell Lines used for DNA extraction

Cell Line	Cell Type	Biopsy Source	Sex	Age	Country of Origin and Ethnicity	Transformant	Tissue Type
NA18507/ GM18507	B-Lymphocyte	Peripheral vein	Male	No data	Nigeria/Yoruba	Epstein-Barr Virus	Blood
NA18555/ GM18555	B-Lymphocyte	Peripheral vein	Female	No data	China/Han Chinese	Epstein-Barr Virus	Blood
NA18517/ GM18517	B-Lymphocyte	Peripheral vein	Female	No data	Nigeria/Yoruba	Epstein-Barr Virus	Blood
NA18572/ GM18572	B-Lymphocyte	Peripheral vein	Male	No data	China/Han Chinese	Epstein-Barr Virus	Blood
NA19239/ GM19239	B-Lymphocyte	Peripheral vein	Male	No data	Nigeria/Yoruba	Epstein-Barr Virus	Blood

Coriell Institute for medical Research (https://catalog.coriell.org/)

Appendix 2: Genomic DNA extraction from lymphoblastoid Cells (Phenol/Chloroform)

Solutions Used
1x TE buffer
1xPhosphate buffered saline
10mM Tris-Cl pH 8.0
1mM EDTA pH 8.0
RNase A (10mg/ml) (SIGMA-ALDRICH and catalog number: **R4642**)
Proteinase K (20mg/ml-stock dissolved in 50mM Tris-Cl pH 8.0, SIGMA-ALDRICH and catalog number: **P6556**)
Phenol:chloroform:IAA
70% (v/w) ethanol
3M Sodium acetate pH 5.6
Lysis Buffer : 100mM NaCl, 100mM Tris-Cl pH 8.0, 25mM EDTA pH 8.0 and 0.5% SDS

Appendix 3: 10X Low dNTPs PCR Mix

10X Low dNTPs PCR mix was another buffer containing the final concentrations of 50mM Tris-HCl (pH8.8@25°C), 12.5mM ammonium sulphate, 1.4mM magnesium chloride, 125µg/ml BSA, 7.5mM 2-mercaptoethanol and 200µM of each dNTP (Promega).

Appendix 4: DNA Sequence Used for ASO Assay's Primer Design (420bp)

TCCAACCATATCCAATGGAGACTTCTACAGCAACAATAGAACATCTTTTC
ACAATGGAACGGTGGTAACTTACCAGTGCCACACTGGACCAGATGGAGAA
CAGCTGTTTGAGCTTGTGGGAGAACGGTCAATATATTGCACCAGCAAAGA
TGATCAAGTTGGTGTTTGGAGCAGCCCTCCCCCTCGGTGTATTTCTACTA
ATAAATGCACAGCTCCAGAAGTTGAAAATGCAATTAGAGTACCAGGAAAC
AGGAGTTTCTTTACCCTCACTGAGATCATCAGATTTAGATGTCAGCCCGG
GTTTGTCATGGTAGGGTCCCACACTGTGCAGTGCCAGACCAATGGCAGAT
GGGGGCCCAAGCTGCCACACTGCTCCAGGGGTGAGTGTGACCCATCAAGA
CTTTGCTGGGTGTGAGGGTA

Appendix 5: Oligo Labelling Solution

Final Concentration	**X1**
10x kinase mix*	1µl
H_2O	7.8µl
T4 polynucleotide kinase (10U/µl) -TNK	0.35µl
γ^{32} P ATP	0.12µl

*10x Kinase Mix

Final Concentration	**Mixed**
70mM Tris-HCl, pH 7.5	700µl 1M stock
100mM $MgCl_2$	100µl 1M stock
50mM spermidine trichloride	100µl 0.5 M
20mM dithiothreitol	40µl 0.5M
	60µl dH_2O

*Store at -20^{o}C

Appendix 6: Kinase Stop Solution

Final Concentration	**Mixed**
25mM diNa EDTA	250µl 0.5M
0.1% (w/v) SDS	50µl 10% (w/v)
10µM ATP	0.5µl 100mM
	4.7ml dH_2O

*Store at -20^{o}C

Appendix 7: Denaturing Mix

Final Concentration	**Mixed**
0.5M NaOH	10ml 10M NaOH stock
2M NaCl	80ml 5M NaCl stock
25mM EDTA	10 ml 0.5M EDTA stock
	100ml dH_2O

*Stored at room temperature

Appendix 8: TMAC Hybridisation Solution

Final Concentration	**Mixed**
3M TMAC	120ml 5M TMAC
0.6% (w/v) SDS	12ml 10% (w/v) SDS
1mM diNaEDTA	0.4ml 0.5M pH8.0
10mM Na phosphate pH6.8	20ml 0.1M
5X Denhardt's solution	20ml 50x
4µg/ml yeast RNA	80µl 10mg/ml
	27.6 ml dH_2O

* Store at 4^oC

50x Denhardt's Solution

Mixed
5g Ficoll 400
5g polyvinylpyrrolidone
5g BSA (fraction V, Sigma)
H_2O to 500 ml

* Store at 4^oC

Appendix 9: TMAC Wash Solution

Final Concentration	**Mixed**
3M TMAC	240ml 5M TMAC
0.6% (w/v) SDS	24ml 10% (w/v) SDS
1mM diNaEDTA	0.8ml 0.5M pH8.0
10mM Na phosphate pH6.8	40ml 0.1M *
	96ml H_2O

* Store at 4^oC

***Sodium phosphate pH 6.8 made using Maniatis instructions:**

1M stock at pH6.8 (46.3ml 1M Na_2HPO_4 + 53.7ml 1M NaH_2PO_4)

Appendix 10: Clustering Quality (Q)

getQualityScore(fit.mean$posterior.H0)

Appendix 11: Linkage Disequilibrium Commands on Plink

cnv1 refers to *CR1*-B (duplication)

cnv2 refers to CR1-C (deletion)

SNP and SNP1/ SNP2: This part should contain the SNP (such as rs3818361) that you are aiming to find the LD relation to a SNP (such as rs6701713) or CNV (such as cnv1/cnv2).

file: The file's name you saved your data (the file should contain one ped and one map file with same name)

my data: The name which the ped and map files were saved.

In System tools (terminal)

1) module load plink
2) cd **file**

- plink --file **my data** --ld cnv1 SNP
- Or plink --file **my data** --ld cnv2 SNP
- Or plink --file **my data** --ld SNP1 SNP2

3) In order to obtain all cnv1 and cnv2 LD relation to SNPs:

- plink --file **my data** --proxy-assoc SNP
- or plink --file **my data** --proxy-assoc cnv1
- or plink --file **my data** --proxy-assoc cnv2

4) In order to obtain Hary-Weinberg Equilibrium results:

- plink --file **my data** --hardy

5) In order to obtain Mendel Errors:

- plink --file **my data** --mendel

Appendix 12: The Power Calculations for LOAD on R

```
sampleDist = function(n,p,q) { sample(x = c(0,1), n,
replace = T, prob = c(p,q)) }

case_hom_A=793

case_het=354

case_hom_B=74

case_B_freq=(case_hom_B*2+case_het)/(case_hom_A*2+case_het*
2+case_hom_B*2)

control_hom_A=406

control_het=142

control_hom_B=28

control_B_freq=(control_hom_B*2+control_het)/(control_hom_A
*2+control_het*2+control_hom_B*2)

# sample sizes for replication study

model_cases=1180

model_controls=1005

nsim=1000

cat("difference in allele frequency",control_B_freq-
case_B_freq)

# I assinged the pval, it was missing

pval <- c()

for (i in 1:nsim)
```

```
  {
  # create vectors of case/control status and allele status


  allele_case<-sampleDist(model_cases*2,1-
case_B_freq,case_B_freq)
  case<-rep(1,model_cases*2)
  allele_control<-sampleDist(model_controls*2,1-
control_B_freq,control_B_freq)
  control<-rep(0,model_controls*2)
  allele<-c(allele_case,allele_control)
  cohort<-c(case,control)
  df<-data.frame(cohort,allele)
  # do logistic regression
  model<-glm(cohort~allele,data=df,family=binomial)
  # store pvalue
  pval[i]=coef(summary(model))["allele","Pr(>|z|)"]
}
cat("power to detect significant result (p<0.05)",
sum(pval<0.05)/nsim)
```

Appendix 13: The Sample Information for the samples used in PRT assays

Sample ID	Plate	Population	Geographic Origin	Region	Sex	Plate No	Plate Position	Age
HGDP00921	HDDP	Yoruba	Nigeria	Subsaharan Africa	Female	13	B2	
HGDP00923	HGDP	Yoruba	Nigeria	Subsaharan Africa	Male	13	D2	
NA19152	HapMap 3	Yoruba	Nigeria	Subsaharan Africa	Female	3	A5	
NA19094	HapMap 3	Yoruba	Nigeria	Subsaharan Africa	Female	3	F9	
NA18507	HapMap	Yoruba	Nigeria	Ibadan	Male			
NA18555	HapMap	Han Chinese and Japanese	Beijing and Tokyo	China and Japan	Female			
NA18517	HapMap	Yoruba	Nigeria	Ibadan	Female			
C0140			United Kingdom		Female			40
C0182			United Kingdom		Female			39
NA19239	HapMap	Yoruba	Nigeria	Ibadan	Male			
NA18572		Han Chinese and Japanese	Beijing and Tokyo	China and Japan	Male			

Appendix 14: Samples showing heterogeneity for PRT3 assay

Sample ID	Population	Geographic Origin	Region	Sex	Plate No	Plate Position
HGDP00460	Biaka_Pygmy	Central African Republic	Subsaharan Africa	Male	1	A2
HGDP00449	Mbuti_Pygmy	Democratic Republic of Congo	Subsaharan Africa	Male	1	B4
HGDP00456	Mbuti_Pygmy	Democratic Republic of Congo	Subsaharan Africa	Male	1	D4
HGDP00913	Mandenka	Senegal	Subsaharan Africa	Male	1	H12
HGDP00915	Mandenka	Senegal	Subsaharan Africa	Female	2	B1
HGDP00175	Sindhi	Pakistan	Asia	Male	5	G7
HGDP00701	Bedouin	Israel (Negev)	Middle East	Female	9	A12
HGDP00739	Palestinian	Israel (Central)	Middle East	Female	9	D5
HGDP00921	Yoruba	Nigeria	Subsaharan Africa	Female	13	B2
HGDP00923	Yoruba	Nigeria	Subsaharan Africa	Male	13	D2
HGDP00468	Mbuti_Pygmy	Democratic Republic of Congo	Subsaharan Africa	Male	13	F1
NA19152	Yoruba	Nigeria	Subsaharan Africa	Female	3	A5
NA19094	Yoruba	Nigeria	Subsaharan Africa	Female	3	F9

8 BIBLIOGRAPHY

Abbott NJ, Patabendige AA, Dolman DE, Yusof SR, Begley DJ. 2010. Structure and function of the blood-brain barrier. Neurobiol Dis 37(1):13-25.

Adu-Gyasi D, Adams M, Amoako S, Mahama E, Nsoh M, Amenga-Etego S, Baiden F, Asante KP, Newton S, Owusu-Agyei S. 2012. Estimating malaria parasite density: assumed white blood cell count of 10,000/mul of blood is appropriate measure in Central Ghana. Malar J 11:238.

Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E and others. 2006. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439(7078):851-5.

Alzheimer's association (2017), 'Alzheimer's disease facts and figures: prevalence', [online] Available from: http://www.alz.org/facts/ (Accessed 23 April 2017)

Alzheimer's Society (2014-2015), 'Dementia in UK', [online] Available from: https://www.alzheimers.org.uk/info/20025/policy_and_influencing/251/dementia_uk (Accessed 23 April 2017)

Andrew SE, Goldberg YP, Kremer B, Telenius H, Theilmann J, Adam S, Starr E, Squitieri F, Lin BY, Kalchman MA and others. 1993. The Relationship between Trinucleotide (Cag) Repeat Length and Clinical-Features of Huntingtons-Disease. Nature Genetics 4(4):398-403.

Antunez C, Boada M, Lopez-Arrieta J, Moreno-Rey C, Hernandez I, Marin J, Gayan J, Alzheimer's Disease Neuroimaging I, Gonzalez-Perez A, Real LM and others. 2011. Genetic association of complement receptor 1 polymorphism rs3818361 in Alzheimer's disease. Alzheimers Dement 7(4):e124-9.

Armour JA, Palla R, Zeeuwen PL, den Heijer M, Schalkwijk J, Hollox EJ. 2007. Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. Nucleic Acids Res 35(3):e19.

Arning L, Monte D, Hansen W, Wieczorek S, Jagiello P, Akkad DA, Andrich J, Kraus PH, Saft C, Epplen JT. 2008. ASK1 and MAP2K6 as modifiers of age at onset in Huntington's disease. J Mol Med (Berl) 86(4):485-90.

Arning L, Saft C, Wieczorek S, Andrich J, Kraus PH, Epplen JT. 2007. NR2A and NR2B receptor gene variations modify age at onset in Huntington disease in a sex-specific manner. Hum Genet 122(2):175-82.

Arora M, Arora R, Tiwari SC, Das N, Srivastava LM. 2000. Expression of complement regulatory proteins in diffuse proliferative glomerulonephritis. Lupus 9(2):127-31.

Arora V, Mondal AM, Grover R, Kumar A, Chattopadhyay P, Das N. 2007. Modulation of CR1 transcript in systemic lupus erythematosus (SLE) by IFN-gamma and immune complex. Mol Immunol 44(7):1722-8.

Arora V, Verma J, Dutta R, Marwah V, Kumar A, Das N. 2004. Reduced complement receptor 1 (CR1, CD35) transcription in systemic lupus erythematosus. Mol Immunol 41(4):449-56.

Ashford JW, Mortimer JA. 2002. Non-familial Alzheimer's disease is mainly due to genetic factors. J Alzheimers Dis 4(3):169-77.

Atkins ER, Bulsara MK, Panegyres PK. 2012. Cerebrovascular risk factors in early-onset dementia. J Neurol Neurosurg Psychiatry 83(6):666-7.

Atkins ER, Panegyres PK. 2011. The clinical utility of gene testing for Alzheimer's disease. Neurol Int 3(1):e1.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7(7):552-64.

Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME. 2008. A robust statistical method for case-control association testing with copy number variation. Nat Genet 40(10):1245-52.

Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. 2007. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet 39(1):17-23.

Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, Battle KE, Moyes CL, Henry A, Eckhoff PA and others. 2015. The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. Nature 526(7572):207-11.

Boampong JN, Acquah S, Sam-Awortwi EN, Ofori MF. 2010. A preliminary study of association of erythrocyte sedimentation rate with malaria specific

immunoglobulin G and malaria-induced anemia. Journal of the Ghana Science Association 12(1):89-98.

Bolliger MF, Martinelli DC, Sudhof TC. 2011. The cell-adhesion G protein-coupled receptor BAI3 is a high-affinity receptor for C1q-like proteins. Proc Natl Acad Sci U S A 108(6):2534-9.

Brinkman RR, Mezei MM, Theilmann J, Almqvist E, Hayden MR. 1997. The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size. American Journal of Human Genetics 60(5):1202-1210.

Broadwell RD, Sofroniew MV. 1993. Serum proteins bypass the blood-brain fluid barriers for extracellular entry to the central nervous system. Exp Neurol 120(2):245-63.

Brouwers N, Van Cauwenberghe C, Engelborghs S, Lambert JC, Bettens K, Le Bastard N, Pasquier F, Montoya AG, Peeters K, Mattheijssens M and others. 2012. Alzheimer risk associated with a copy number variation in the complement receptor 1 increasing C3b/C4b binding sites. Molecular Psychiatry 17(2):223-233.

Campbell CD, Sampas N, Tsalenko A, Sudmant PH, Kidd JM, Malig M, Vu TH, Vives L, Tsang P, Bruhn L and others. 2011. Population-Genetic Properties of Differentiated Human Copy-Number Polymorphisms. American Journal of Human Genetics 88(3):317-332.

Campion D, Dumanchin C, Hannequin D, Dubois B, Belliard S, Puel M, Thomas-Anterion C, Michon A, Martin C, Charbonnier F and others. 1999. Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. Am J Hum Genet 65(3):664-70.

Cantsilieris S, Baird PN, White SJ. 2013. Molecular methods for genotyping complex copy number polymorphisms. Genomics 101(2):86-93.

Carel JC, Frazier B, Ley TJ, Holers VM. 1989. Analysis of epitope expression and the functional repertoire of recombinant complement receptor 2 (CR2/CD21) in mouse and human cells. J Immunol 143(3):923-30.

Carlson J, Helmby H, Hill AV, Brewster D, Greenwood BM, Wahlgren M. 1990. Human cerebral malaria: association with erythrocyte rosetting and lack of anti-rosetting antibodies. Lancet 336(8729):1457-60.

Chartier-Harlin MC, Parfitt M, Legrain S, Perez-Tur J, Brousseau T, Evans A, Berr C, Vidal O, Roques P, Gourlet V and others. 1994. Apolipoprotein E, epsilon 4 allele as a major risk factor for sporadic early and late-onset forms of Alzheimer's disease: analysis of the 19q13.2 chromosomal region. Hum Mol Genet 3(4):569-74.

Chattopadhyay B, Baksi K, Mukhopadhyay S, Bhattacharyya NP. 2005. Modulation of age at onset of Huntington disease patients by variations in TP53 and human caspase activated DNase (hCAD) genes. Neurosci Lett 374(2):81-6.

Chattopadhyay B, Ghosh S, Gangopadhyay PK, Das SK, Roy T, Sinha KK, Jha DK, Mukherjee SC, Chakraborty A, Singhal BS and others. 2003. Modulation of age at onset in Huntington's disease and spinocerebellar ataxia type 2 patients originated from eastern India. Neurosci Lett 345(2):93-6.

Chen CB, Wallis R. 2004. Two mechanisms for mannose-binding protein modulation of the activity of its associated serine proteases. Journal of Biological Chemistry 279(25):26058-26065.

Chu Y, Jin X, Parada I, Pesic A, Stevens B, Barres B, Prince DA. 2010. Enhanced synaptic connectivity and epilepsy in C1q knockout mice. Proc Natl Acad Sci U S A 107(17):7975-80.

Cockburn IA, Mackinnon MJ, O'Donnell A, Allen SJ, Moulds JM, Baisor M, Bockarie M, Reeder JC, Rowe JA. 2004. A human complement receptor 1 polymorphism that reduces Plasmodium falciparum rosetting confers protection against severe malaria. Proceedings of the National Academy of Sciences of the United States of America 101(1):272-277.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P and others. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464(7289):704-12.

Cooling L. 2015. Blood Groups in Infection and Host Susceptibility. Clin Microbiol Rev 28(3):801-70.

Cooper NR, Nemerow GR. 1989. Complement and infectious agents: a tale of disguise and deception. Complement Inflamm 6(4):249-58.

Covas DT, de Oliveira FS, Rodrigues ES, Abe-Sandes K, Silva WA, Jr., Fontes AM. 2007. Knops blood group haplotypes among distinct Brazilian populations. Transfusion 47(1):147-53.

Crehan H, Hardy J, Pocock J. 2012a. Microglia, Alzheimer's disease, and complement. Int J Alzheimers Dis 2012:983640.

Crehan H, Hardy J, Pocock J. 2013. Blockage of CR1 prevents activation of rodent microglia. Neurobiol Dis 54:139-49.

Crehan H, Holton P, Wray S, Pocock J, Guerreiro R, Hardy J. 2012b. Complement receptor 1 (CR1) and Alzheimer's disease. Immunobiology 217(2):244-50.

Daniels G. 2013. Human Blood Groups: Wiley-Blackwell:443-445.

Du Pasquier L. LG. 2000. Origin and Evolution of the Vertebrate Immune System. Springer Edition.

Dunkelberger JR, Song WC. 2010. Complement and its role in innate and adaptive immune responses. Cell Res 20(1):34-50.

Dykman TR, Hatch JA, Aqua MS, Atkinson JP. 1985. Polymorphism of the C3b/C4b receptor (CR1): characterization of a fourth allele. J Immunol 134(3):1787-9.

Eichler EE. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet 17(11):661-9.

Eikelenboom P, Stam FC. 1982. Immunoglobulins and complement factors in senile plaques. An immunoperoxidase study. Acta Neuropathol 57(2-3):239-42.

Eikelenboom P, Veerhuis R, Scheper W, Rozemuller AJ, van Gool WA, Hoozemans JJ. 2006. The significance of neuroinflammation in understanding Alzheimer's disease. J Neural Transm (Vienna) 113(11):1685-95.

Faik I, de Carvalho EG, Kun JF. 2009. Parasite-host interaction in malaria: genetic clues and copy number variation. Genome Med 1(9):82.

Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, van Duijn CM. 1997. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. JAMA 278(16):1349-56.

Fitness J, Floyd S, Warndorff DK, Sichali L, Mwaungulu L, Crampin AC, Fine PE, Hill AV. 2004. Large-scale candidate gene study of leprosy susceptibility in the Karonga district of northern Malawi. Am J Trop Med Hyg 71(3):330-40.

Flores M, Morales L, Gonzaga-Jauregui C, Dominguez-Vidana R, Zepeda C, Yanez O, Gutierrez M, Lemus T, Valle D, Avila MC and others. 2007. Recurrent DNA

inversion rearrangements in the human genome. Proc Natl Acad Sci U S A 104(15):6099-106.

Folch J, Petrov D, Ettcheto M, Abad S, Sanchez-Lopez E, Garcia ML, Olloquequi J, Beas-Zarate C, Auladell C, Camins A. 2016. Current Research Therapeutic Strategies for Alzheimer's Disease Treatment. Neural Plast 2016:8501693.

Fonseca MI, Chu S, Pierce AL, Brubaker WD, Hauhart RE, Mastroeni D, Clarke EV, Rogers J, Atkinson JP, Tenner AJ. 2016. Analysis of the Putative Role of CR1 in Alzheimer's Disease: Genetic Association, Expression and Function. PLoS One 11(2):e0149792.

Fontes AM, Kashima S, Bonfim-Silva R, Azevedo R, Abraham KJ, Albuquerque SR, Bordin JO, Junior DM, Covas DT. 2011. Association between Knops blood group polymorphisms and susceptibility to malaria in an endemic area of the Brazilian Amazon. Genet Mol Biol 34(4):539-45.

Frank I, Stoiber H, Godar S, Stockinger H, Steindl F, Katinger HW, Dierich MP. 1996. Acquisition of host cell-surface-derived molecules by HIV-1. AIDS 10(14):1611-20.

Fraser DA, Pisalyaput K, Tenner AJ. 2010. C1q enhances microglial clearance of apoptotic neurons and neuronal blebs, and modulates subsequent inflammatory cytokine production. J Neurochem 112(3):733-43.

Gandhi M, Singh A, Dev V, Adak T, Dashd AP, Joshi H. 2009. Role of CR1 Knops polymorphism in the pathophysiology of malaria: Indian scenario. J Vector Borne Dis 46(4):288-94.

Gandy S, Haroutunian V, DeKosky ST, Sano M, Schadt EE. 2013. CR1 and the "vanishing amyloid" hypothesis of Alzheimer's disease. Biol Psychiatry 73(5):393-5.

Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, Pedersen NL. 2006. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry 63(2):168-74.

Gaunt TR, Rodriguez S, Guthrie PA, Day IN. 2010. An expectation-maximization program for determining allelic spectrum from CNV data (CoNVEM): insights into population allelic architecture and its mutational history. Hum Mutat 31(4):414-20.

Girirajan S, Campbell CD, Eichler EE. 2011. Human copy number variation and complex genetic disease. Annu Rev Genet 45:203-26.

Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, Haynes A, Irving N, James L and others. 1991. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature 349(6311):704-6.

Griffin JT, Ferguson NM, Ghani AC. 2014. Estimates of the changing age-burden of Plasmodium falciparum malaria disease in sub-Saharan Africa. Nat Commun 5:3136.

Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. Pathogenetics 1(1):4.

Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A and others. 2015. The African Genome Variation Project shapes medical genetics in Africa. Nature 517(7534):327-32.

Gutekunst CA, Li SH, Yi H, Ferrante RJ, Li XJ, Hersch SM. 1998. The cellular and subcellular localization of huntingtin-associated protein 1 (HAP1): Comparison with huntingtin in rat and human. Journal of Neuroscience 18(19):7674-7686.

Hamer I, Paccaud JP, Belin D, Maeder C, Carpentier JL. 1998. Soluble form of complement C3b/C4b receptor (CR1) results from a proteolytic cleavage in the C-terminal region of CR1 transmembrane domain. Biochem J 329 (Pt 1):183-90.

Hansson HH, Kurtzhals JA, Goka BQ, Rodriques OP, Nkrumah FN, Theander TG, Bygbjerg IC, Alifrangis M. 2013. Human genetic polymorphisms in the Knops blood group are not associated with a protective advantage against Plasmodium falciparum malaria in Southern Ghana. Malar J 12:400.

Hardwick RJ, Amogne W, Mugusi S, Yimer G, Ngaimisi E, Habtewold A, Minzi O, Makonnen E, Janabi M, Machado LR and others. 2012. beta-defensin genomic copy number is associated with HIV load and immune reconstitution in sub-saharan Africans. J Infect Dis 206(7):1012-9.

Hardwick RJ, Machado LR, Zuccherato LW, Antolinos S, Xue Y, Shawa N, Gilman RH, Cabrera L, Berg DE, Tyler-Smith C and others. 2011. A worldwide analysis of

beta-defensin copy number variation suggests recent selection of a high-expressing DEFB103 gene copy in East Asia. Hum Mutat 32(7):743-50.

Hardy J, Selkoe DJ. 2002. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science 297(5580):353-6.

Haridan US, Mokhtar U, Machado LR, Abdul Aziz AT, Shueb RH, Zaid M, Sim B, Mustafa M, Nik Yusof NK, Lee CK and others. 2015. A comparison of assays for accurate copy number measurement of the low-affinity Fc gamma receptor genes FCGR3A and FCGR3B. PLoS One 10(1):e0116791.

Hauptmann G, Tappeiner G, Schifferli JA. 1988. Inherited deficiency of the fourth component of human complement. Immunodefic Rev 1(1):3-22.

Hazrati LN, Van Cauwenberghe C, Brooks PL, Brouwers N, Ghani M, Sato C, Cruts M, Sleegers K, St George-Hyslop P, Van Broeckhoven C and others. 2012. Genetic association of CR1 with Alzheimer's disease: a tentative disease mechanism. Neurobiol Aging 33(12):2949 e5-2949 e12.

Holers VM, Chaplin DD, Leykam JF, Gruner BA, Kumar V, Atkinson JP. 1987. Human-Complement C3b/C4b Receptor (Cr-1) Messenger-Rna Polymorphism That Correlates with the Cr-1 Allelic Molecular-Weight Polymorphism. Proceedings of the National Academy of Sciences of the United States of America 84(8):2459-2463.

Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo MM, Abraham R, Hamshere ML, Pahwa JS, Moskvina V and others. 2011. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet 43(5):429-35.

Hollox EJ. 2012. The challenges of studying complex and dynamic regions of the human genome. Methods Mol Biol 838:187-207.

Hollox EJ, Detering JC, Dehnugara T. 2009. An integrated approach for measuring copy number variation at the FCGR3 (CD16) locus. Hum Mutat 30(3):477-84.

Hosokawa M, Klegeris A, Maguire J, McGeer PL. 2003. Expression of complement messenger RNAs and proteins by human oligodendroglial cells. Glia 42(4):417-23.

Houlden H, Crook R, Backhovens H, Prihar G, Baker M, Hutton M, Rossor M, Martin JJ, Van Broeckhoven C, Hardy J. 1998. ApoE genotype is a risk factor in

nonpresenilin early-onset Alzheimer's disease families. Am J Med Genet 81(1):117-21.

Hourcade D, Holers VM, Atkinson JP. 1989. The regulators of complement activation (RCA) gene cluster. Adv Immunol 45:381-416.

Ifeanyichukwu M O-EE, Okeke C,, F ABaI. 2015. Haemorrheological Changes in Malaria Infected Pregnant Subject Attending Antenatal Clinic at Eku Baptist Government Hospital. 2(2).

Jiang H, Burdick D, Glabe CG, Cotman CW, Tenner AJ. 1994. beta-Amyloid activates complement by binding to a specific region of the collagen-like domain of the C1q A chain. J Immunol 152(10):5050-9.

Jiang T, Yu JT, Tan L. 2012. Novel disease-modifying therapies for Alzheimer's disease. J Alzheimers Dis 31(3):475-92.

Joachim CL, Mori H, Selkoe DJ. 1989. Amyloid beta-protein deposition in tissues other than brain in Alzheimer's disease. Nature 341(6239):226-30.

Jobling M., Hollox E., Hurles M., Kivisild T., Tyler-Smith C. 2014. Human Evolutionary Genetics (2nd Ed.). New York and London: Garland Science:44-690.

Johansson L, Rytkonen A, Bergman P, Albiger B, Kallstrom H, Hokfelt T, Agerberth B, Cattaneo R, Jonsson AB. 2003. CD46 in meningococcal disease. Science 301(5631):373-5.

Jones L, Holmans PA, Hamshere ML, Harold D, Moskvina V, Ivanov D, Pocklington A, Abraham R, Hollingworth P, Sims R and others. 2010. Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer's disease. PLoS One 5(11):e13950.

Josling GA, Llinas M. 2015. Sexual development in Plasmodium parasites: knowing when it's time to commit. Nat Rev Microbiol 13(9):573-87.

Kammer GM, Perl A, Richardson BC, Tsokos GC. 2002. Abnormal T cell signal transduction in systemic lupus erythematosus. Arthritis Rheum 46(5):1139-54.

Katyal M, Sivasankar B, Ayub S, Das N. 2003. Genetic and structural polymorphism of complement receptor 1 in normal Indian subjects. Immunol Lett 89(2-3):93-8.

Katyal M, Tiwari SC, Kumar A, Dinda AK, Arora V, Kumar R, Das N. 2004. Association of complement receptor 1 (CR1, CD35, C3b/C4b receptor) density polymorphism with glomerulonephritis in Indian subjects. Mol Immunol 40(18):1325-32.

Kaul DK, Roth EF, Nagel RL, Howard RJ, Handunnetti SM. 1991. Rosetting of Plasmodium-Falciparum-Infected Red-Blood-Cells with Uninfected Red-Blood-Cells Enhances Microvascular Obstruction under Flow Conditions. Blood 78(3):812-819.

Kehoe P, Krawczak M, Harper PS, Owen MJ, Jones AL. 1999. Age of onset in Huntington disease: sex specific influence of apolipoprotein E genotype and normal CAG repeat length. J Med Genet 36(2):108-11.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Research 12(6):996-1006.

Kindt, T. J., Goldsby, R. A., Osborne, B. A., & Kuby, J. 2007. *Kuby immunology* (6th ed.). New York: W.H. Freeman:174-181.

Klickstein LB, Wong WW, Smith JA, Weis JH, Wilson JG, Fearon DT. 1987. Human C3b/C4b receptor (CR1). Demonstration of long homologous repeating domains that are composed of the short consensus repeats characteristics of C3/C4 binding proteins. J Exp Med 165(4):1095-112.

Kong PL, Odegard JM, Bouzahzah F, Choi JY, Eardley LD, Zielinski CE, Craft JE. 2003. Intrinsic T cell defects in systemic autoimmunity. Ann N Y Acad Sci 987:60-7.

Kooner JS, Saleheen D, Sim X, Sehmi J, Zhang W, Frossard P, Been LF, Chia KS, Dimas AS, Hassanali N and others. 2011. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. Nat Genet 43(10):984-9.

Kullo IJ, Ding KY, Shameer K, McCarty CA, Jarvik GR, Denny JC, Ritchie MD, Ye Z, Crosslin DR, Chisholm RL and others. 2011. Complement Receptor 1 Gene Variants Are Associated with Erythrocyte Sedimentation Rate. American Journal of Human Genetics 89(1):131-138.

Kumps C, Van Roy N, Heyrman L, Goossens D, Del-Favero J, Noguera R, Vandesompele J, Speleman F, De Preter K. 2010. Multiplex Amplicon Quantification (MAQ), a fast and efficient method for the simultaneous detection of copy number alterations in neuroblastoma. BMC Genomics 11:298.

Kun JF, Schmidt-Ott RJ, Lehman LG, Lell B, Luckner D, Greve B, Matousek P, Kremsner P.. 1998. Merozoite surface antigen 1 and 2 genotypes and rosetting of Plasmodium falciparum in severe and mild malaria in Lambarene, Gabon. Trans R Soc Trop Med Hyg 92(1):110-4.

Lam KW, Jeffreys AJ. 2006. Processes of copy-number change in human DNA: the dynamics of {alpha}-globin gene deletion. Proc Natl Acad Sci U S A 103(24):8921-7.

Lam KW, Jeffreys AJ. 2007. Processes of de novo duplication of human alpha-globin genes. Proc Natl Acad Sci U S A 104(26):10950-5.

Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, Combarros O, Zelenika D, Bullido MJ, Tavernier B and others. 2009. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet 41(10):1094-9.

Latorra D, Campbell K, Wolter A, Hurley JM. 2003. Enhanced allele-specific PCR discrimination in SNP genotyping using 3' locked nucleic acid (LNA) primers. Hum Mutat 22(1):79-85.

Le Port A, Cottrell G, Martin-Prevel Y, Migot-Nabias F, Cot M, 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):e000342.

Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131(7):1235-47.

Liu S, Yao L, Ding D, Zhu H. 2010. CCL3L1 copy number variation and susceptibility to HIV-1 infection: a meta-analysis. PLoS One 5(12):e15778.

Lucotte G, Turpin JC, Riess O, Epplen JT, Siedlaczk I, Loirat F, Hazout S. 1995. Confidence-Intervals for Predicted Age-of-Onset, Given the Size of (Gag)(N) Repeat, in Huntingtons-Disease. Human Genetics 95(2):231-232.

Lupski JR. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14(10):417-22.

Lupski JR. 2007. Genomic rearrangements and sporadic disease. Nat Genet 39(7 Suppl):S43-7.

Lupski JR, Deocaluna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ, Saucedocardenas O, Barker DF, Killian JM, Garcia CA and others. 1991. DNA Duplication Associated with Charcot-Marie-Tooth Disease Type-1a. Cell 66(2):219-232.

Lupski JR, Stankiewicz P. 2005. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet 1(6):e49.

Ma H, Difazio S. 2008. An efficient method for purification of PCR products for sequencing. Biotechniques 44(7):921-3.

MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 42(Database issue):D986-92.

Macdonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, Barnes G, Taylor SA, James M, Groot N and others. 1993. A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntingtons-Disease Chromosomes. Cell 72(6):971-983.

Madi N, Paccaud JP, Steiger G, Schifferli JA. 1991. Immune complex binding efficiency of erythrocyte complement receptor 1 (CR1). Clin Exp Immunol 84(1):9-15.

Mahmoudi R, Kisserli A, Novella JL, Donvito B, Drame M, Reveil B, Duret V, Jolly D, Pham BN, Cohen JH. 2015. Alzheimer's disease is associated with low density of the long CR1 isoform. Neurobiol Aging 36(4):1766 e5-12.

Malaviya AN, Singh RR, Singh YN, Kapoor SK, Kumar A. 1993. Prevalence of systemic lupus erythematosus in India. Lupus 2(2):115-8.

Martin JB, Gusella JF. 1986. Huntington's disease. Pathogenesis and management. N Engl J Med 315(20):1267-76.

Mattias Möller MJ, Jill R. Storry and Martin L. Olsson. 2016. Erythrogene: a database for in-depth analysis of the extensive variation in 36 blood group systems in the 1000 Genomes Project. The American Society of Hematology 1:1:240-249.

Metzger S, Rong J, Nguyen HP, Cape A, Tomiuk J, Soehn AS, Propping P, Freudenberg-Hua Y, Freudenberg J, Tong L and others. 2008. Huntingtin-associated protein-1

is a modifier of the age-at-onset of Huntington's disease. Human Molecular Genetics 17(8):1137-1146.

Miller DJ, Hemmrich G, Ball EE, Hayward DC, Khalturin K, Funayama N, Agata K, Bosch TC. 2007. The innate immune repertoire in cnidaria--ancestral complexity and stochastic gene loss. Genome Biol 8(4):R59.

Morgan BP, Gasque P. 1996. Expression of complement in the brain: role in health and disease. Immunol Today 17(10):461-6.

Moulds JM. 2002. Understanding the Knops blood group and its role in malaria. Vox Sang 83 Suppl 1:185-8.

Moulds JM. 2010. The Knops blood-group system: a review. Immunohematology 26(1):2-7.

Moulds JM, Brai M, Cohen J, Cortelazzo A, Cuccia M, Lin M, Sadallah S, Schifferli J, Bala Subramanian V, Truedsson L and others. 1998. Reference typing report for complement receptor 1 (CR1). Exp Clin Immunogenet 15(4):291-4.

Moulds JM, Kassambara L, Middleton JJ, Baby M, Sagara I, Guindo A, Coulibaly S, Yalcouye D, Diallo DA, Miller L and others. 2000. Identification of complement receptor one (CR1) polymorphisms in west Africa. Genes Immun 1(5):325-9.

Moulds JM, Pierce S, Peck KB, Tulley ML, Doumbo O, Moulds JJ. 2005. KAM: A new allele in the Knops blood group system. Transfusion 45(3):27a-27a.

Moulds JM, Reveille JD, Arnett FC. 1996. Structural polymorphisms of complement receptor 1 (CR1) in systemic lupus erythematosus (SLE) patients and normal controls of three ethnic groups. Clin Exp Immunol 105(2):302-5.

Moulds JM, Thomas BJ, Doumbo O, Diallo DA, Lyke KE, Plowe CV, Rowe JA, Birmingham DJ. 2004. Identification of the Kna/Knb polymorphism and a method for Knops genotyping. Transfusion 44(2):164-9.

Moulds JM, Zimmerman PA, Doumbo OK, Diallo DA, Atkinson JP, Krych-Goldberg M, Hourcade DE, Moulds JJ. 2002. Expansion of the Knops blood group system and subdivision of Sl(a). Transfusion 42(2):251-6.

Moulds JM, Zimmerman PA, Doumbo OK, Kassambara L, Sagara I, Diallo DA, Atkinson JP, Krych-Goldberg M, Hauhart RE, Hourcade DE and others. 2001. Molecular identification of Knops blood group polymorphisms found in long homologous region D of complement receptor 1. Blood 97(9):2879-85.

Multiplicom (2015), MAQ Control for WG and ROI MAQ, [online] Available from:

http://www.multiplicom.com/sites/default/files/ifu511_user_guide_maq_control_v151105.pdf (Accessed 14 September 2017)

Murdock, George P. *Ethnographic Atlas.* Pittsburgh: The University of Pittsburgh Press. 1967

Murray CJ, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, Haring D, Fullman N, Naghavi M, Lozano R, Lopez AD. 2012. Global malaria mortality between 1980 and 2010: a systematic analysis. Lancet 379(9814):413-31.

Naitza S, Porcu E, Steri M, Taub DD, Mulas A, Xiao X, Strait J, Dei M, Lai S, Busonero F and others. 2012. A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. PLoS Genet 8(1):e1002480.

Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, Gallins PJ, Buxbaum JD, Jarvik GP, Crane PK and others. 2011. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet 43(5):436-41.

Nash GB, Cooke BM, Carlson J, Wahlgren M. 1992. Rheological Properties of Rosettes Formed by Red-Blood-Cells Parasitized by Plasmodium-Falciparum. British Journal of Haematology 82(4):757-763.

Naze P, Vuillaume I, Destee A, Pasquier F, Sablonniere B. 2002. Mutation analysis and association studies of the ubiquitin carboxy-terminal hydrolase L1 gene in Huntington's disease. Neurosci Lett 328(1):1-4.

Ng YC, Schifferli JA, Walport MJ. 1988. Immune complexes and erythrocyte CR1 (complement receptor type 1): effect of CR1 numbers on binding and release reactions. Clin Exp Immunol 71(3):481-5.

NHS (2017), 'Blood tests: Erythrocyte sedimentation rate (ESR)',[online] Available from:

http://www.nhs.uk/Conditions/Blood-tests/Pages/What-it-is-used-for.aspx#ESR (Accessed 21 September2017)

Nochlin D, van Belle G, Bird TD, Sumi SM. 1993. Comparison of the severity of neuropathologic changes in familial and sporadic Alzheimer's disease. Alzheimer Dis Assoc Disord 7(4):212-22.

Noumsi GT, Tounkara A, Diallo H, Billingsley K, Moulds JJ, Moulds JM. 2011. Knops blood group polymorphism and susceptibility to Mycobacterium tuberculosis infection. Transfusion 51(11):2462-9.

Panchamoorthy G, Tiwari SC, Srivastava LM. 1993. Inherited structural and quantitative polymorphisms of C3b receptor (CR1) in normals and patients with glomerular diseases. Asian Pac J Allergy Immunol 11(2):123-9.

Panegyres PK, Chen HY. 2013. Differences between early and late onset Alzheimer's disease. Am J Neurodegener Dis 2(4):300-6.

Pascual M, Danielsson C, Steiger G, Schifferli JA. 1994. Proteolytic cleavage of CR1 on human erythrocytes in vivo: evidence for enhanced cleavage in AIDS. Eur J Immunol 24(3):702-8.

Patin E, Siddle KJ, Laval G, Quach H, Harmant C, Becker N, Froment A, Regnault B, Lemee L, Gravel S and others. 2014. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat Commun 5:3163.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R and others. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet 39(10):1256-60.

Petty AC, Green CA, Poole J, Daniels GL. 1997. Analysis of Knops blood group antigens on CR1 (CD35) by the MAIEA test and by immunoblotting. Transfus Med 7(1):55-62.

Pisalyaput K, Tenner AJ. 2008. Complement component C1q inhibits beta-amyloid- and serum amyloid P-induced neurotoxicity via caspase- and calpain-independent mechanisms. J Neurochem 104(3):696-707.

Raeymaekers P, Timmerman V, Nelis E, De Jonghe P, Hoogendijk JE, Baas F, Barker DF, Martin JJ, De Visser M, Bolhuis PA and others. 1991. Duplication in chromosome 17p11.2 in Charcot-Marie-Tooth neuropathy type 1a (CMT 1a). The HMSN Collaborative Research Group. Neuromuscul Disord 1(2):93-7.

Rao N, Ferguson DJ, Lee SF, Telen MJ. 1991. Identification of human erythrocyte blood group antigens on the C3b/C4b receptor. J Immunol 146(10):3502-7.

Reiling L, Richards JS, Fowkes FJ, Wilson DW, Chokejindachai W, Barry AE, Tham WH, Stubbs J, Langer C, Donelson J and others. 2012. The Plasmodium falciparum erythrocyte invasion ligand Pfrh4 as a target of functional and protective human antibodies against malaria. PLoS One 7(9):e45253.

Ricklin D, Hajishengallis G, Yang K, Lambris JD. 2010. Complement: a key system for immune surveillance and homeostasis. Nat Immunol 11(9):785-97.

Rogaeva E. 2002. The solved and unsolved mysteries of the genetics of early-onset Alzheimer's disease. Neuromolecular Med 2(1):1-10.

Rogers J, Cooper NR, Webster S, Schultz J, McGeer PL, Styren SD, Civin WH, Brachova L, Bradt B, Ward P and others. 1992. Complement activation by beta-amyloid in Alzheimer disease. Proc Natl Acad Sci U S A 89(21):10016-20.

Rowe A, Obeiro J, Newbold CI, Marsh K. 1995. Plasmodium falciparum rosetting is associated with malaria severity in Kenya. Infect Immun 63(6):2323-6.

Rowe JA, Moulds JM, Newbold CI, Miller LH. 1997. P. falciparum rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. Nature 388(6639):292-5.

Rowe JA, Raza A, Diallo DA, Baby M, Poudiougo B, Coulibaly D, Cockburn IA, Middleton J, Lyke KE, Plowe CV, Doumbo OK, Moulds JM: Erythrocyte CR1 expression level does not correlate with a HindIII restriction fragment length polymorphism in Africans; implications for studies on malaria susceptibility. 2002. Genes Immun. 3: 497-500.

Rowe JA, Rogerson SJ, Raza A, Moulds JM, Kazatchkine MD, Marsh K, Newbold CI, Atkinson JP, Miller LH. 2000. Mapping of the region of complement receptor (CR) 1 required for Plasmodium falciparum rosetting and demonstration of the importance of CR1 in rosetting in field isolates. J Immunol 165(11):6341-6.

Rubinsztein DC, Leggo J, Chiano M, Dodge A, Norbury G, Rosser E, Craufurd D. 1997. Genotypes at the GluR6 kainate receptor locus are associated with variation in the age of onset of Huntington disease. Proc Natl Acad Sci U S A 94(8):3872-6.

Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, Jones CG, Zaitlen NA, Varilo T, Kaakinen M and others. 2009. Genome-wide association analysis

of metabolic traits in a birth cohort from a founder population. Nat Genet 41(1):35-46.

Sando SB, Melquist S, Cannon A, Hutton ML, Sletvold O, Saltvedt I, White LR, Lydersen S, Aasly JO. 2008. APOE epsilon 4 lowers age at onset and is a high risk factor for Alzheimer's disease; a case control study from central Norway. BMC Neurol 8:9.

Sarma JV, Ward PA. 2011. The complement system. Cell and Tissue Research 343(1):227-235.

Seltzer B, Sherwin I. 1983. A comparison of clinical features in early- and late-onset primary degenerative dementia. One entity or two? Arch Neurol 40(3):143-6.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R and others. 2005. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77(1):78-88.

Shastry BS. 2009. Copy number variation and susceptibility to human disorders (Review). Mol Med Rep 2(2):143-7.

Shaw CJ, Lupski JR. 2004. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet 13 Spec No 1:R57-64.

Shen Y, Lue L, Yang L, Roher A, Kuo Y, Strohmeyer R, Goux WJ, Lee V, Johnson GV, Webster SD and others. 2001. Complement activation by neurofibrillary tangles in Alzheimer's disease. Neurosci Lett 305(3):165-8.

Small SA, Duff K: Linking Abeta and tau in late-onset Alzheimer’s disease: a dual pathway hypothesis. 2008. Neuron 60 (4): 534-542.

Speth C, Dierich MP, Gasque P. 2002. Neuroinvasion by pathogens: a key role of the complement system. Mol Immunol 38(9):669-79.

Stankiewicz P, Lupski JR. 2002. Genome architecture, rearrangements and genomic disorders. Trends Genet 18(2):74-82.

Stevens B, Allen NJ, Vazquez LE, Howell GR, Christopherson KS, Nouri N, Micheva KD, Mehalow AK, Huberman AD, Stafford B and others. 2007. The classical complement cascade mediates CNS synapse elimination. Cell 131(6):1164-78.

Stoiber H, Ebenbichler C, Schneider R, Janatova J, Dierich MP. 1995. Interaction of several complement proteins with gp120 and gp41, the two envelope glycoproteins of HIV-1. AIDS 9(1):19-26.

Stoute JA. 2011. Complement receptor 1 and malaria. Cell Microbiol 13(10):1441-50.

Stoute, J.A., Odindo, A.O., Owuor, B.O., Mibei, E.K., Opollo, M.O., and Waitumbi, J.N. 2003. Loss of red blood cellcomplement regulatory proteins and increased levels of circulating immune complexes are associated with severe malarial anemia. J Infect Dis 187(3): 522–525.

Stuart PE, Huffmeier U, Nair RP, Palla R, Tejasvi T, Schalkwijk J, Elder JT, Reis A, Armour JAL. 2012. Association of beta-Defensin Copy Number and Psoriasis in Three Cohorts of European Origin. Journal of Investigative Dermatology 132(10):2407-2413.

Taherzadeh-Fard E, Saft C, Andrich J, Wieczorek S, Arning L. 2009. P.C-1alpha as modifier of onset age in Huntington disease. Mol Neurodegener 4:10.

Tako EA, Zhou A, Lohoue J, Leke R, Taylor DW, Leke RF. 2005. Risk factors for placental malaria and its effect on pregnancy outcome in Yaounde, Cameroon. Am J Trop Med Hyg 72(3):236-42.

Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution. 30(12): 2725–2729.

Tetteh-Quarcoo PB, Schmidt CQ, Tham WH, Hauhart R, Mertens HD, Rowe A, Atkinson JP, Cowman AF, Rowe JA, Barlow PN. 2012. Lack of evidence from studies of soluble protein fragments that Knops blood group polymorphisms in complement receptor-type 1 are driven by malaria. PLoS One 7(4):e34820.

Tham WH, Schmidt CQ, Hauhart RE, Guariento M, Tetteh-Quarcoo PB, Lopaticki S, Atkinson JP, Barlow PN, Cowman AF. 2011. Plasmodium falciparum uses a key functional site in complement receptor type-1 for invasion of human erythrocytes. Blood 118(7):1923-33.

Tham WH, Wilson DW, Lopaticki S, Schmidt CQ, Tetteh-Quarcoo PB, Barlow PN, Richard D, Corbin JE, Beeson JG, Cowman AF. 2010. Complement receptor 1 is the host erythrocyte receptor for Plasmodium falciparum PfRh4 invasion

ligand. Proceedings of the National Academy of Sciences of the United States of America 107(40):17327-17332.

Thambisetty M, An Y, Nalls M, Sojkova J, Swaminathan S, Zhou Y, Singleton AB, Wong DF, Ferrucci L, Saykin AJ and others. 2013. Effect of complement CR1 on brain amyloid burden during aging and its modification by APOE genotype. Biol Psychiatry 73(5):422-8.

Thathy V, Moulds JM, Guyah B, Otieno W, Stoute JA. 2005. Complement receptor 1 polymorphisms associated with resistance to severe malaria in Kenya. Malar J 4:54.

Treutiger CJ, Hedlund I, Helmby H, Carlson J, Jepson A, Twumasi P, Kwiatkowski D, Greenwood BM, Wahlgren M. 1992. Rosette formation in Plasmodium falciparum isolates and anti-rosette activity of sera from Gambians with cerebral or uncomplicated malaria. Am J Trop Med Hyg 46(5):503-10.

Trouw LA, Nielsen HM, Minthon L, Londos E, Landberg G, Veerhuis R, Janciauskiene S, Blom AM. 2008. C4b-binding protein in Alzheimer's disease: binding to Abeta1-42 and to dead cells. Mol Immunol 45(13):3649-60.

Turner DJ, Miretti M, Rajan D, Fiegler H, Carter NP, Blayney ML, Beck S, Hurles ME. 2008. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. Nat Genet 40(1):90-5.

Tyler-Smith C, Xue Y. 2012. Sibling rivalry among paralogs promotes evolution of the human brain. Cell 149(4):737-9.

Uneke CJ. 2007. Impact of placental Plasmodium falciparum malaria on pregnancy and perinatal outcome in sub-Saharan Africa: I: introduction to placental malaria. Yale J Biol Med 80(2):39-50.

Veal CD, Xu H, Reekie K, Free R, Hardwick RJ, McVey D, Brookes AJ, Hollox EJ, Talbot CJ. 2013. Automated design of paralogue ratio test assays for the accurate and rapid typing of copy number variation. Bioinformatics 29(16):1997-2003.

Veerhuis R, Nielsen HM, Tenner AJ. 2011. Complement in the brain. Mol Immunol 48(14):1592-603.

Veldhuisen B, Ligthart PC, Vidarsson G, Roels I, Folman CC, van der Schoot CE, de Haas M. 2011. Molecular analysis of the York antigen of the Knops blood group system. Transfusion 51(7):1389-1396.

Vik DP, Wong WW. 1993. Structure of the gene for the F allele of complement receptor type 1 and sequence of the coding region unique to the S allele. J Immunol 151(11):6214-24.

Waitumbi, J.N., Opollo, M.O., Muga, R.O., Misore, A.O., and Stoute, J.A. 2000. Red cell surface changes and erythrophagocytosis in children with severe Plasmodium falciparum anemia. Blood 95(4): 1481–1486.

Waitumbi, J.N., Donvito, B., Kisserli, A., Cohen, J.H., and Stoute, J.A. 2004. Age-related changes in red blood cell complement regulatory proteins and susceptibility to severe malaria. J Infect Dis 190(6): 1183–1191.

Walport MJ. 2001. Complement. First of two parts. N Engl J Med 344(14):1058-66.

Weis JH, Morton CC, Bruns GA, Weis JJ, Klickstein LB, Wong WW, Fearon DT. 1987. A complement receptor locus: genes encoding C3b/C4b receptor and C3d/Epstein-Barr virus receptor map to 1q32. J Immunol 138(1):312-5.

Weterman MA, van Ruissen F, de Wissel M, Bordewijk L, Samijn JP, van der Pol WL, Meggouh F, Baas F. 2010. Copy number variation upstream of PMP22 in Charcot-Marie-Tooth disease. Eur J Hum Genet 18(4):421-8.

Wexler NS, Res UVC. 2004. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. Proceedings of the National Academy of Sciences of the United States of America 101(10):3498-3503.

Weydt P, Soyal SM, Gellera C, Didonato S, Weidinger C, Oberkofler H, Landwehrmeyer GB, Patsch W. 2009. The gene coding for P.C-1alpha modifies age at onset in Huntington's Disease. Mol Neurodegener 4:3.

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)

Wilson JG, Murphy EE, Wong WW, Klickstein LB, Weis JH, Fearon DT. 1986. Identification of a Restriction-Fragment-Length-Polymorphism by a Cr-1 Cdna That Correlates with the Number of Cr-1 on Erythrocytes. Journal of Experimental Medicine 164(1):50-59.

Wolfe F, Pincus T. 2001. The level of inflammation in rheumatoid arthritis is determined early and remains stable over the longterm course of the illness. J Rheumatol 28(8):1817-24.

Wong WW, Cahill JM, Rosen MD, Kennedy CA, Bonaccio ET, Morris MJ, Wilson JG, Klickstein LB, Fearon DT. 1989. Structure of the human CR1 gene. Molecular basis of the structural and quantitative polymorphisms and identification of a new CR1-like allele. J Exp Med 169(3):847-63.

Wong WW, Farrell SA. 1991. Proposed structure of the F' allotype of human CR1. Loss of a C3b binding site may be associated with altered function. J Immunol 146(2):656-62.

Wong WW, Kennedy CA, Bonaccio ET, Wilson JG, Klickstein LB, Weis JH, Fearon DT. 1986. Analysis of multiple restriction fragment length polymorphisms of the gene for the human complement receptor type I. Duplication of genomic sequences occurs in association with a high molecular mass receptor allotype. J Exp Med 164(5):1531-46.

Wong WW, Wilson JG, Fearon DT. 1983. Genetic regulation of a structural polymorphism of human C3b receptor. J Clin Invest 72(2):685-93.

Wooster DG, Maruvada R, Blom AM, Prasadarao NV. 2006. Logarithmic phase Escherichia coli K1 efficiently avoids serum killing by promoting C4bp-mediated C3b and C4b degradation. Immunology 117(4):482-93.

Wyss-Coray T, Yan F, Lin AH, Lambris JD, Alexander JJ, Quigg RJ, Masliah E. 2002. Prominent neurodegeneration and increased plaque formation in complement-inhibited Alzheimer's mice. Proc Natl Acad Sci U S A 99(16):10837-42.

Yoon SH, Fearon DT. 1985. Characterization of a soluble form of the C3b/C4b receptor (CR1) in human plasma. J Immunol 134(5):3332-8.

Zawia NH, Lahiri DK, Cardozo-Pelaez F. 2009. Epigenetics, oxidative stress, and Alzheimer disease. Free Radic Biol Med 46(9):1241-9.

Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10:451-81.

Zhou J, Fonseca MI, Pisalyaput K, Tenner AJ. 2008. Complement C3 and C4 expression in C1q sufficient and deficient mouse models of Alzheimer's disease. J Neurochem 106(5):2080-92.

Zhu XC, Yu JT, Jiang T, Wang P, Cao L, Tan L. 2015. CR1 in Alzheimer's disease. Mol Neurobiol 51(2):753-65.

Zhu Y, Thangamani S, Ho B, Ding JL. 2005. The ancient origin of the complement system. EMBO J 24(2):382-94.

Zimmerman PA, Fitness J, Moulds JM, McNamara DT, Kasehagen LJ, Rowe JA, Hill AV. 2003. CR1 Knops blood group alleles are not associated with severe malaria in the Gambia. Genes Immun 4(5):368-73.

Zotova E, Nicoll JA, Kalaria R, Holmes C, Boche D. 2010. Inflammation in Alzheimer's disease: relevance to pathogenesis and therapy. Alzheimers Res Ther 2(1):1.

Zuhlke C, Riess O, Schroder K, Siedlaczck I, Epplen JT, Engel W, Thies U. 1993. Expansion of the (Cag)(N) Repeat Causing Huntingtons-Disease in 352 Patients of German Origin. Human Molecular Genetics 2(9):1467-1469.