SEQUENCE VARIATION OF COPY NUMBER VARIABLE REGIONS IN THE HUMAN GENOME

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

by

Hasret Ozturk Pala MSc

Department of Genetics and Genome Biology
College of Medicine, Biological Sciences and Psychology
University of Leicester

2020

ABSTRACT

SEQUENCE VARIATION OF COPY NUMBER VARIABLE REGIONS IN THE HUMAN GENOME

Accurate genotyping of sequence variation in repeated and copy number variable regions of genomes remains challenging, because of the problems inherent in mapping short sequence reads to a reference genome. A computational pipeline was designed to attempt to resolve the short-read mapping ambiguity for duplicated DNA regions mapping short-reads to reference sequence comprising of a single copy of a region repeated in the reference genome. The *RHCE/RHD*, the beta-defensin and the low-affinity FC gamma receptor repeat regions were chosen as initial analyses. The reliability of mapping to a reduced reference was assessed by comparing sequence read depth and known copy number across a subset of samples from the 1000 Genomes Project and a three-generation pedigree from Illumina's Platinum Genomes Project.

The accuracy of variant calling was analysed by comparing variant calls at the inhibitory low-affinity Fc gamma receptor gene *FCGR2B* with 1000 Genomes variant calls and the variant calls generated by paralogue-specific long PCR and Ion Torrent sequencing.

Both the reduced reference read approach and the 1000 genomes variant calls did not call all variants found by the Ion Torrent sequencing variant calls, with the 1000 Genomes variant calls significantly underestimating and mis-genotyping samples. Several variants in *FCGR2B* were found to be in strong LD with variants previously associated with complex traits by genome wide association studies (GWAS). However, these GWAS variants were in weak linkage disequilibrium with a gene conversion variant upstream of *FCGR2B*.

Given that a coding variant of *FCGR2B* (rs1050501) has been previously associated with protection against severe malaria and susceptibility to systemic lupus erythematosus, the variation data was interrogated for signature of selection across global populations, and the genetic diversity of this locus revealed high haplotype diversity with 52 haplotypes. However, the population genetic statistics showed no evidence of natural selection at *FCGR2B*.

Acknowledgements

I gratefully acknowledge my PhD sponsor the Turkish Ministry of National Education for the financial support during my PhD.

First of all, I would like to express my great appreciation, deepest gratitude to my adviser Dr Edward J Hollox, I thank him for his valuable and excellent guidance, patience for doing my research work. It has been an honour to be his PhD student. Also, I would like to thank my PGR committee Dr Sandra Beleza and Dr Sinead Drea for going carefully through my reports over the years and giving useful suggestions regarding my project.

I would like to thank Faisal Almalki, Adel Alharbi, Walid Algady, 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 like to thank, Rachel Madison and Gurdeep M. Lall for technical support in the lab, Chiara Batini for helping me bioinformatics analyses, Rita Neumann for teaching me Ion Torrent sequencing. I thank Dr Richard Badge, Dr Ezio Rosato, and Prof. Raymond Dalgleish for choosing me to be a demonstrator in the undergraduate practical courses, it has been great experience.

I would like to thank my husband Jean-Paul B. Pala for his love, patience and all the support he provided over the years. I thank my family for believing in me and their extraordinary support, my brother Umut Kader Ozturk, my parents Sehriban and Dogan Ozturk, and my beloved grandmother Zekiye Pur I know she always prayed for me. I thank my cousin M. Inan Celik and Laman Maharramova for being my family in UK. I thank precious friend Ezgi Kucukkilic for her never-ending support and friendship.

Lastly, I offer my regards to all of those who supported me in any respect during the completion of the research. This thesis would have been impossible to complete without you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AF	Allele frequency
AFS	Allele frequency spectrum
Array-CGH	Array comparative genomic hybridization
AS	*De novo* Assembly
BAM	Binary Alignment Map
BCR	B cell antigen receptors
BLAT	Basic Local Alignment Search Tool
BWA	Burrow-Wheeler algorithm
bp	base pair
CEPH	Centre d'Étude du Polymorphisme Humain
CNV	Copy number variation
CN	Copy number
Cy3	Cyanine 3
Cy5	Cyanine 5
DEFB	Human Beta-Defensin
DNA	Deoxyribonucleic acid
DSBs	DNA double-strand break
dNTPs	Deoxy Nucleotide Triphosphates
FAM	Fluorescein Amidite
FoSTes	Fork Stalling and Template Switching
FCGRs	FC Gamma Receptors
gnomAD	Genome Aggregation Database
GWAS	Genome Wide Association Study
GPI	Glycosylphosphatidylinositol
HEX	Hexachloro-Fluorescein
HTS	High throughout sequencing
ITAMs	Tyrosine-based activation motifs
ITIMs	Tyrosine-based inhibition motifs
kb	kilobase
LCRs	Low Copy Repeats
LINE-1	Long interspersed element 1
LD	Linkage Disequilibrium
LNA	Locked Nucleic Acid
mAb	Monoclonal antibodies
Mb	megabase
MJ	Median joining
MLPA	Multiplex ligation-dependent probe amplification
mya	Million years ago
NAHR	Non-Allelic Homologous Recombination
NGS	Next Generation Sequencing
NHEJ	Non-Homologous End Joining
NHLBI-ESP	NHLBI GO Exome Sequencing Project
NI	Neutrality index
PRT	Paralogue Ratio Test
PCR	Polymerase Chain Reaction
qPCR	Quantitative PCR
Q	Phred quality score
PolyPhen	Polymorphism phenotyping

RP	Read pair
RD	Read depth
SAM	Sequence Alignment Map
SD	Segmental duplication
SLE	Systemic Lupus Erythematosus
SIFT	Sorting intolerant from tolerant
SNP	Single Nucleotide Polymorphism
SMRT	Single molecule real time
SR	Split read
SV	Structural Variation
TAE	Tris-Acetate-EDTA buffer
TBE	Tris-Borate-EDTA buffer
TVC	Torrent Variant Caller
USCS	University of California, Santa Cruz
UV	Ultraviolet Light
VEP	Variant Effect Predictor
ZMWs	zero-mode waveguides
µl	Microliter

CHAPTER 1 INTRODUCTION

Segmental duplications (SDs) are one of the most important components that contribute to human disease directly or through leading to other forms of structural variation such as Copy number Variations (CNVs). They also contribute to human evolution and adaptation. In addition, the diversity of SDs in human populations may be used as maskers for the population genetic studies (Numanagić et al., 2018). SDs can be difficult to characterize due to the genetic variation and high sequence similarity within the region. Therefore, they may be underrepresented or wrongly assembled in genome assembly databases and the ambiguities in read mapping may lead to potentially incorrect assumptions. According to Ebbert el., 2019, there are "dark" genes/regions in the human genomes. These regions are defined as regions where short read sequencing technologies cannot sufficiently be assembled or aligned. Therefore, the identification of the variation in these regions that are relevant to disease cannot be detected. Many dark regions are claimed to be arisen from duplicated genomic regions where alignment of short reads to unique location is not possible, these regions are termed as "camouflaged". Regardless of whether the duplication is transcriptionally and translationally active or inactive, however, any genomic region that has been nearly identically duplicated and is large enough to prevent sequencing reads from aligning unambiguously will be "dark" due to the capability of the aligner programs.

This study primarily aims to design a computational pipeline in an attempt to resolve the read mapping ambiguity in short read sequencing and to investigate the sequence variation in the duplicated DNA regions of the human genome by using a reduced reference.

1.1 Segmental duplications

SDs (also called "low-copy repeats" (LCRs)) are blocks of DNA sequence larger than 1kb, which can be found at least twice with more than 90% sequence similarity in the genome (Sharp et al., 2006). Locating and characterizing segmentally duplicated regions in the human genome is important and a great interest. Firstly, SDs have played an important part for shaping the evolution of the human genome. Some studies have shown that human genome has undergone several segmental duplications during the past 35 Myr

(Bailey et al., 2002; Samonte and Eichler, 2002; Cheung et al., 2003). In addition, these recent genomic changes might have contributed to the species divergence between human and the apes or Old-World monkeys in a significant amount (Stankiewicz et al., 2001). Secondly, these genomic rearrangements have been found to be the associated with many genetic diseases in humans (Stankiewicz and Lupski, 2002; Zhang et al., 2005).

Approximately 5.2% of the human genome is thought to consists of segmental duplications and duplicons (Bailey et al., 2002; Cheung et al. 2003). SDs are either intrachromosomal (on the same chromosome, 3.9%), or inter-chromosomal (on different chromosomes, 2.3%). Most of the duplicons can be found in the pericentromeric regions. The intrachromosomal duplications play a significant role in evolution and they are risk factors for genomic rearrangements that cause human disorders such as cases of velocardiofacial syndrome on chromosome 22q (Guo et al., 2011), Williams-Beuren syndrome on chromo-some 7q (Ebert et al., 2014), and Smith-Magenis syndrome on chromosome 17p (Park et al., 2002).

SDs are assumed to be the result of copy number variations (CNVs) reaching fixation. In addition, it has been suggested that CNV formation is partly mediated by SDs (Freeman et al. 2006; Sharp et al. 2005, 2006; Kim et al., 2008). This would suggest that SD formation tend to occur in regions where previously existing SDs. An SD-rich region can generate more CNVs than other regions, some of which become fixed as SDs. Then, there would be an imbalance in the distribution of SDs because some regions in the genome would be very rich in SDs while other regions would very poorly contain SDs. This would result in a highly skewed distribution of SDs. Therefore, it is suggested that segmental duplications follow a power-law (“the rich get richer”, Albert and Barabasi, 2002) pattern in the human genome (Kim et al., 2008).

It has been suggested that non-allelic homologous recombination (NAHR) during meiosis can contribute to the formation of larger deletion and duplication. Recombination mechanism such as NAHR are claimed to be mediated by pre-existing repeats such as Alu elements have previously involved in the formation of SDs (Bailey et al., 2003) which is consistent with NAHR-based formation. Similarly, SDs have been proposed as mediating CNV formation (Freeman et al., 2006; Sharp et al., 2006). However, not all duplications are thought to arise because of NAHR-based mechanisms. Non-

homologous end-joining (NHEJ) has been suggested for SD formation in the subtelomere regions and human subtelomeres are hot spots of inter-chromosomal recombination and segmental duplication. (Linardopoulou et al., 2005). Not all CNVs are associated with segmental duplications. Maximally 28% of CNVs were found to be formed by an SD-mediated mechanism (Kim et al., 2008). Other mechanisms also facilitate for the formation of CNVs (see Chapter 1.2.1).

1.2 Copy number variation

CNVss a type of genomic structural variation that can include different numbers of segmental duplications or deletions of a DNA fragment. CNVs have been shown to be very common in a variety of organisms, including humans (Sebat et al. 2004; Conrad et al. 2006; McCarroll et al. 2006; Redon et al. 2006), mice (Graubert et al. 2007; She et al. 2008), chimpanzees (Perry et al. 2006), rhesus macaques (Lee et al. 2008), cows (Liu et al. 2010), dogs (Chen et al. 2009; Nicholas et al. 2009), chickens (Griffin et al. 2008), maize (Springer et al. 2009), *Arabidopsis thaliana* (the thale cress) (Ossowski et al. 2008), fruit flies (Dopman & Hartl 2007; Emerson et al. 2008), *Caenorhabditis elegans* (a free-living transparent nematode) (Maydan et al. 2010) and *Saccharomyces cerevisiae* (a species of *yeast*) (Carreto et al. 2008). Though it is often harder to experimentally identify and genotype CNVs relative to SNPs and indels, many are big enough to encompass whole genes and are therefore more likely to affect organismal fitness (Schrider and Hahn, 2010).

The CNV size was originally defined as the amplification or deletion of DNA in the size of >1 kb, and it was later widened to include much smaller sizes (>50 bp) due to methodology development (Girirajan et al., 2011). Because CNVs range in size from several dozens of bases (> 50 bp) (MacDonald et al. 2014; Zarrei et al. 2015) to megabases within a single human genome, they can give rise to 1.2% difference from the reference human genome (Pang et al. 2010).

The simplest type of copy number variation is the presence or absence of a gene (Figure 1.2A). A diploid genome consists of two copies of a gene. A simple duplication of a genomic segment could result in diploid copy numbers of two, three, or four (Figure 1.2B). Diallelic CNVs have two alleles and could produce three different genotypes in

both deletion and duplication events. and successive rounds of duplication could produce a wide range of diploid copy numbers, known as multiallelic copy number variants (Figure 1.2C). These deletions and duplications result in variants that range from a few hundred to several million bp and could contain an entire gene, part of a gene, a region outside of a gene, or several genes in case of larger variants (Wain et al., 2009).

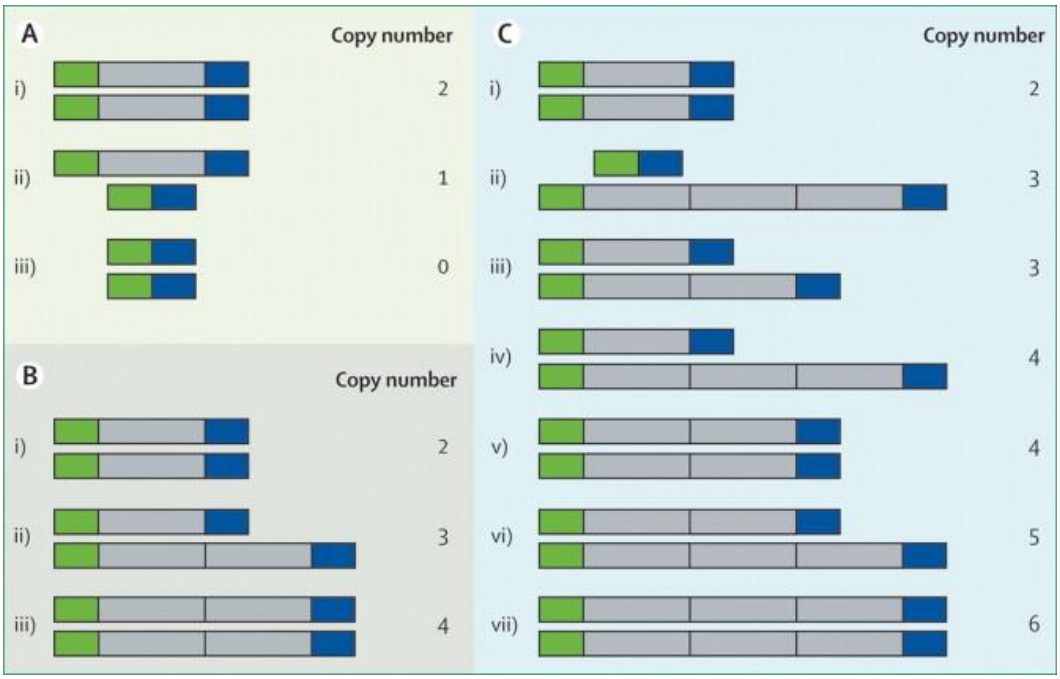


Figure 1. 1 Diallelic and multiallelic copy number variation. Diallelic locus (grey) and flanking loci (green and blue) with copy number variation caused by (A) deletion and (B) duplication, each showing the locus with (i) normal diploid copy number, (ii) heterozygous state, and (iii) homozygous state. (C) Multiallelic locus showing (i) normal copy number, (ii) multiple rounds of duplication on one chromosome and a deletion on the homologous chromosome, (iii) duplication on one chromosome and no deletion on the homologous chromosome, (iv) multiple rounds of duplication on one chromosome and no deletion on the homologous chromosome, (v) one round of duplication on each chromosome, (vi) one round of duplication on one chromosome and multiple rounds of duplication on the homologous chromosome, and (vii) multiple rounds of duplication on both chromosomes. Multiallelic assays measure total diploid copy number but cannot describe genotype status of (ii) and (iii), or (iv) and (v). Obtained from Wain et al., 2009.

CNVs can be critical in susceptibility or resistance to human diseases, such as Alzheimer disease (Rovelet-Lecrux, et al, 2006), autism (Sebat et al., 2007) and psoriasis (de Cid et al., 2009). Numerous CNVs have been associated with complex human diseases (Table 1.1). For example, complement component 4 (*C4*) gene mutations have been investigated to see the association with systemic lupus erythematosus (SLE), an autoimmune disease. The copy number of the *C4* varies from 2 to 6. Higher copy of the *C4* region is associated with lower risk of SLE, and low copy of C4 is associated with higher risk of SLE (Yang et al., 2007). Additionally, the *C4* gene has different alleles; C4A-short, C4B-short, C4A-long, and C4B-long. It is stated that these alleles are associated with schizophrenia, especially associated with the highest C4A expression (Sekar et al., 2016). As another example, the *IDS* gene is located on the X chromosome. People with abnormal copy number of the *IDS* gene, due to several deletions, may result in the development of Hunter syndrome (mucopolysaccharidosis type II, MPS II) which primarily occurs in male individuals, as an X- linked recessive genetic disorder (Zhang et al., 2011a). The alpha-synuclein (*SNCA*) gene encodes the SNCA protein. This protein is mainly found in neurons in the brain. It is predicted to be functional as vesicle turnover at the presynaptic terminals and trigger voluntary and involuntary movements by regulating the neurotransmitter release. Whole gene duplication and triplication of *SNCA* are responsible for producing more protein in brain tissue as a result of gene dosage effect, and are therefore thought to cause Parkinson disease, a degenerative disorder of the central nervous system (Miller et al., 2004; Fuchs et al., 2007).

Table 1. 1 Some copy number polymorphisms associated with complex diseases.

Gene	Disease/Trait	Variant type	Associated allele	Reference
DEFB4, DEFB103, DEFB104	Psoriasis	Amplification[a,b]	High copy number	Hollox et al., 2008b.
DEFB4, DEFB103, DEFB104	Crohn's disease	Amplification[a,b]	Low copy number	Aldhous et al., 2010. Fellermann et al., 2006.
CCL3L1	HIV/AIDS	Amplification[a,b]	Low copy number	Field et al., 2009. Gonzalez et al., 2005. Urban et al., 2009.
C4	Lupus	Amplification[a,b]	Low copy number	Yang et al., 2007.
FCGR3B	Glomerulonephritis in Lupus patients	Amplification[a,b]	Low copy number	Aitman et al.,2006. Fanciulli et al., 2007.
FCGR3B	Lupus	Amplification[a,b]	Low copy number	Fanciulli et al., 2007.
RHD	Rh-negative blood group	Deletion[a]	Deletion	Colin et al., 1991.
NBPF23	Neuroblastoma	Deletion[a,b]	Deletion	Diskin et al., 2009.
UGT2B17	Osteoporosis	Deletion[b]	No deletion	Yang et al., 2008.
HLA	Crohn's disease, rheumatoid arthritis, type 1 diabetes	Multiple CNVs[a]	Various	Consortium TWTCC. 2010.

[a]Multicopy CNV, more than three diploid copy numbers observed in the population.
[b]CNV is in a segmental duplication.
Reproduced from Girirajan et al.,2011.

In 2006, the first CNV map of the human genome was constructed through the investigation of 270 individuals from four populations with ancestry in Europe, Africa, or Asia (Redon et al. 2006). 1,447 CNV regions (including genes, disease loci, functional elements and SDs) covering 360 megabases (12% of the genome) were identified in these populations. Using the Whole Genome Tile Path array (WGTP), which comprised of 26,574 large insert clones representing 93.7% of the euchromatic portion of the human genome (Fiegler et al. 2006), the average number of CNVs detected per genome was 70 and the mean size was 341 kb (Redon et al. 2006). A higher resolution map of CNVs based on 55 studies was developed later (Zarrei et al. 2015). It was estimated that up to 9.5% of the genome contributes to CNV. It was also found that approximately 100 genes can be homozygously deleted without producing apparent phenotypic

consequences. This map is a great and important contribution to the understanding of new CNV findings, for any research and clinical applications (Zarrei et al., 2015; Nowakowska, 2017).

CNVs can occur at different frequencies in the population. If the frequency is lower than 1%, then the CNV is considered as rare in contrast to common or polymorphic CNVs, which occur in the population with frequency higher than 1% (Valsesia et al. 2013). Both types of CNVs can occur in a normal population as well as in patients with abnormal phenotypes (Redon et al. 2006; Valsesia et al. 2012). However, only a very few common variants have been associated with diseases (Wellcome Trust Case Control Consortium and Craddock, 2010). In rare CNVs, association studies are much more difficult and require a large cohort to gain statistical power. Therefore, the analyses of association with diseases can be particularly challenging for rare CNVs. On the other hand, rare CNVs are particularly enriched in individuals with complex neurodevelopmental phenotypes (Coe et al. 2012; Iyer and Girirajan, 2015; Nowakowska, 2017). CNVs are a large source of both normal and pathogenic variants. Many CNVs are considered benign, while others are pathogenic. Between these two types, a wide range of variants can be identified (Nowakowska, 2017).

1.2.1 Mechanisms of CNV Formation

Non-allelic homologous recombination (NAHR), non-homologous end-joining (NHEJ), fork stalling and template switching (FoSTeS), and (long interspersed element-1, LINE-1) L1-mediated retro transposition are some of the mechanisms known for the formation of copy number variation. These mechanisms generate rearrangements in the genome and possibly account for the majority of CNV formation (Gu et al., 2008; Hastings et al., 2009).

1.2.1.1 Non-allelic homologous recombination (NAHR)

Most recurrent genomic rearrangements are caused by NAHR between two LCRs aka SDs. The non-allelic copies of LCRs, instead of the copies at the usual allelic position, can be aligned in meiosis and mitosis because there is high degree of sequence identity. This misalignment and following crossover between them can results in genomic rearrangements in progeny cells (Figure 1.2). These non-allelic copies are responsible

for the observed breakpoint clustering. If the two LCRs are located on the same chromosome and in direct orientation, NAHR between them causes duplications and deletions. If they are on the same chromosome but in opposite orientation, NAHR results in inversion of the fragment flanked by them. NAHR between repeats on different chromosomes can lead to chromosomal translocation (Gu et al., 2008).

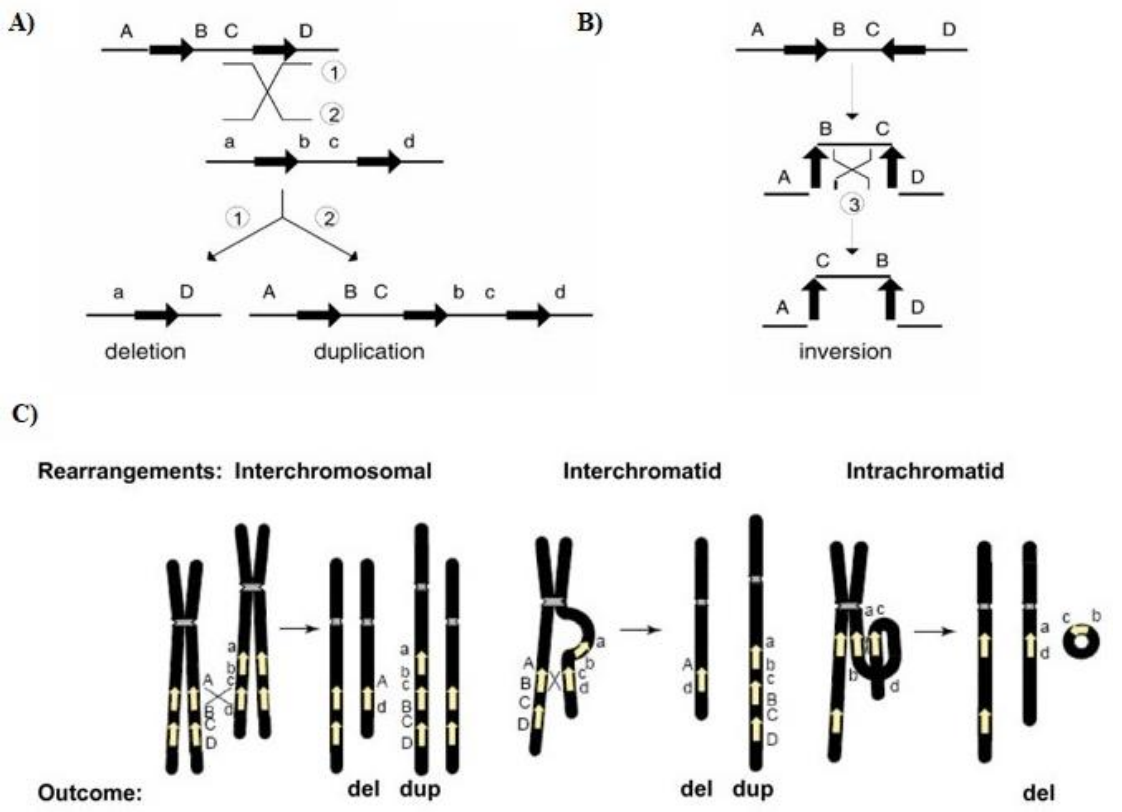


Figure 1. 2 NAHR as the mechanism for recurrent genomic rearrangements. A and B are the genomic rearrangements resulting from recombination between low copy repeats (LCRs). LCRs are marked as black arrows with the orientation indicated by the direction of the arrowhead. Capital letters above the thin horizontal lines refer to the flanking unique sequences (for example, A). Homologues on the other strand (can be another chromatid or the homologous chromosome) are also shown (for example, a). Thin diagonal lines refer to a recombination event with the results shown by numbers 1, 2 and 3. A shows the recombination between direct repeats results in deletion and/or duplication. B shows ther recombination between inverted repeats results in an inversion. C Schematic representation of reciprocal duplications and deletions mediated by interchromosomal (left), interchromatid (middle) and intrachromatid (right) non-allelic homologous recombination (NAHR) using LCR pairs in direct orientation. Chromosomes are shown in black, with the centromere depicted by hashed lines. Yellow arrows depict LCRs. Letters adjacent to the chromatids refer to the flanking unique sequence (for example, A, a). Interchromosomal and interchromatid NAHR between LCRs in direct orientation result in reciprocal duplication and deletion, whereas intrachromatid NAHR only creates deletion. Obtained from Gu et al., 2008.

1.2.1.2 Non-homologous end-joining (NHEJ)

As a non-replicative pathway, NHEJ is one of the major mechanisms used by eukaryotic cells and t has been described in organisms from bacteria to mammals. It aims to repair a DNA double-strand break (DSB) which is a critical form of DNA damage. This pathway uses several proteins; it recognizes the DNA lesion, stabilizes the protein complex (NHEJ complex) at the broken ends, makes a bridge between the broken ends and ligates them (Lieber, 2008). It is the most prominent DNA repair mechanism because it can happen any phase of the cell cycle (Chen at al., 2010a). NHEJ does not require a homologous reference during the repair of the broken ends, however, the presence of terminal microhomologies (typically 1–4 bp) facilitates classical NHEJ but this is not absolutely necessary. It can leave an information scar in the form of loss or addition of several nucleotides at the junction point (Lieber, 2008). NHEJ of ends from simultaneous DSBs has the potential to account for a diverse range of genomic rearrangements, with some possible outcomes illustrated in Figure 1.3.

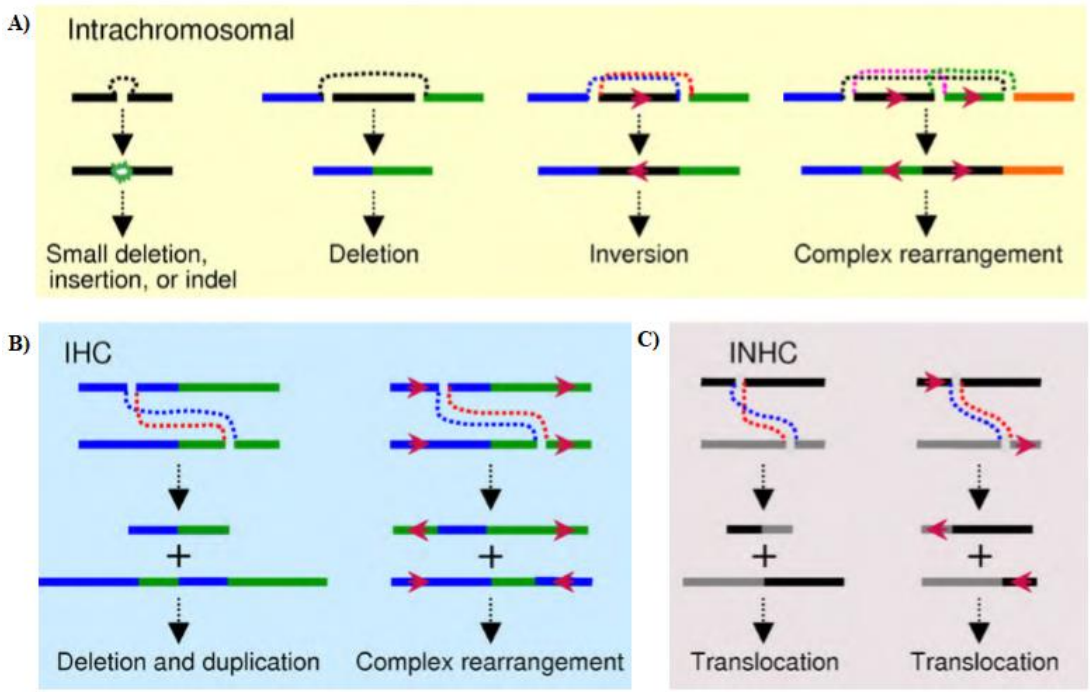


Figure 1. 3 Consequences of genomic rearrangements by NHEJ mechanisms. In A, the examples of genomic rearrangements resulting from non-homologous end joining (NHEJ) are shown. Ends ligated are indicated by dotted lines. In B and C, the final outcome, unlike non-allelic homologous recombination (NAHR), is not necessarily reciprocal. In theory, the flexibility of NHEJ implies an unlimited number of different types of genomic rearrangement. IHC, inter-homologous chromosomes. INHC, inter-nonhomologous chromosomes. Obtain from Chen at al., 2010a.

1.2.1.3 The Fork Stalling and Template Switching (FosTes)

FosTes is a DNA replication–based model and predicted to occur during miosis. During DNA replication, the DNA replication fork stalls at one position, the lagging strand disengages from the original template, transfers and then anneals, by virtue of microhomology at the 3' end, to another replication fork in physical proximity and DNA synthesis is restarted (Lee et al., 2007). FoSTeS is a unique mechanism compared with NAHR and NHEJ because it is a replication-based rearrangement pathway, the rearrangement is induced by errors in the replication procedure and does not depend on the preformation of DSBs (Gu et al., 2008). The invasion and annealing rely on the microhomology between the invaded site and the original site. Upon annealing, the transferred strand primes its own template-driven extension at the transferred fork. Switching to another fork positioned downstream (forward invasion) would give rise to a deletion, while switching to a fork positioned upstream (backward invasion) cause a duplication. Depending on whether the lagging or leading strand in the new fork was invaded and copied, and the direction of the fork progression, the erroneously incorporated fragment from the new replication fork would be in direct or inverted orientation to its original position (Figure 1.4).

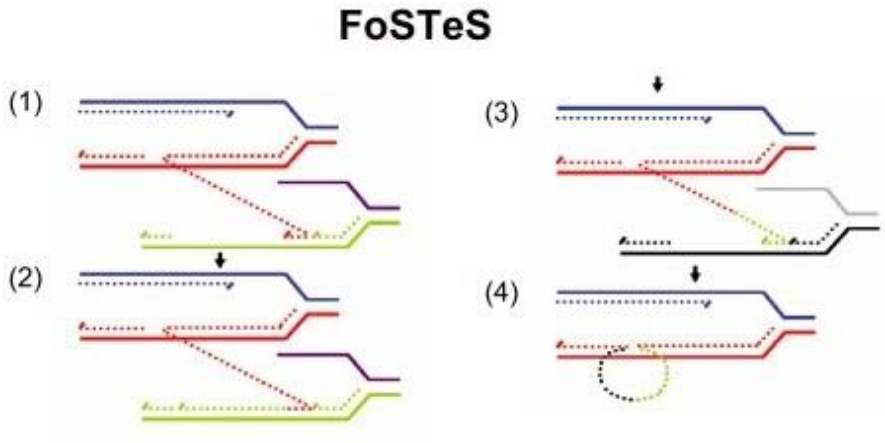


Figure 1. 4 Genomic rearrangements by FoSTeS mechanisms. After the original stalling of the replication fork (dark blue and red, solid lines), the lagging strand (red, dotted line) disengages and anneals to a second fork (purple and green, solid lines) via microhomology (1), followed by (2) extension of the now 'primed' second fork and DNA synthesis (green, dotted line). After the fork disengages (3), the tethered original fork (dark blue and red, solid lines) with its lagging strand (red and green, dotted lines) could invade a third fork (gray and black, solid lines). Dotted lines represent newly synthesized DNA. Serial replication fork disengaging and lagging strand invasion could occur several times (e.g. FoSTeS x 2, FoSTeS x 3, ... etc.) before (4) resumption of replication on the original template. Obtained from Gu et al., 2008.

1.2.1.4 Long interspersed element-1 (L1) retro-transposition

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 occurs via an RNA intermediate that is probably transcribed by RNA polymerase II. The reverse transcription and integration are thought to occur in a coupled process called target primed reverse transcription (TPRT). The resultant insertion is bounded by duplicated target sites (TSD), characteristic of TPRT (Zhang et al., 2009). Besides simple self-insertion, L1 elements can mobilize their 5'- and 3'-flanking DNA sequences in cis and non-autonomous sequences in trans (e.g. Alu sequences) to new genomic locations. Moreover, L1 retro-transposition can also give rise to large genomic deletions (Chen et al., 2010a).

1.2.2 Classes of CNVs

CNVs are divided into two categories based on mutational origin and mechanism type: non-recurrent and recurrent CNVs. Both categories can have different mutational rates as well as different evolutionary directions (Hollox and Hoh, 2014). Recurrent CNVs exist in regions containing large segmental duplications. NAHR is thought to be the major mechanism to create recurrent CNVs in human genome. 20-40 % of the polymorphic CNVs are thought to be recurrent as result of NAHR (Conrad et al. 2010). Recurrent CNVs occur in many places in the genome, however, the main hotspots are subtelomeric and pericentromeric regions (Redon et al., 2006; Hollox and Hoh, 2014). Non-recurrent CNVs involve large genomic regions, can be in any size and can be originated by NHEJ or FosTes mechanisms (Zhang et al., 2009). Such CNVs can have very strong negative effects on phenotypes because they are frequently large and consequently impact on genes (Arlt et al., 2012). It is also claimed that many non-recurrent CNVs are rare because negative selection acts to quickly to remove the deletion from populations (Hollox and Hoh, 2014).

1.2.3 Phenotypic consequences of CNVs

CNVs are very common in human populations and can significantly affect phenotypes in several ways (Figure 1.5). First, it may create gene dosage effects. The alteration of the number of copies of the full gene produces more gene products, it increases the amount of mRNA and protein in cells. For example, the risk of psoriasis, a common inflammatory skin condition, increases because of the increased copy number of β-defensins block in

human (Hollox et al., 2008a). When CNV occurs between a regulatory element (an enhancer or repressor), and the gene as a consequence of position effect, it can change the amount mRNA coded by the gene and the amount of protein produced. *HoxD* gene cluster affects the skeletal pattern in mouse (Montavon et al., 2012). CNVs may create novel genes which are result of unequal crossing over of the copies of DNA sequences. In this case, non-identical copies may create novel genes. The Butyrophilin-like (*BTNL*) proteins have an important role in inflammation and immune response. *BTNL3* gene is expressed in neutrophils and *BTNL8* in eosinophils. These genes have high sequence similarity as well as similar expression patterns. The new *BTNL8*3* deletion is a result of a common 56-kilobase deletion in a segmental duplication region. The presence of the novel BTNL8*3 fusion-protein product affects the immune system and the expression of the *BTNL9* gene and shows different allele frequency among human populations (Aigner et al. 2013). Additionally, extra coding domains may be created when the number of tandem repeats of coding exons is changed. This may affect the size of amino acid sequences, and consequently the protein since it changes the number functional protein domains. The complement receptor 1 (*CR1*) is large type-I transmembrane glycoprotein (Wong et al. 1989). The presence of long homologous consensus repeats (LHRs) results in different isoforms of CR1 and creates additional C3b/C4b binding sites; therefore, the complement activation is affected by the isoforms which have different affinities (Brouwers et al., 2012).

The microscopically visible CNVs are mostly associated with phenotypic consequences. The bigger the CNV region, the more genomic variants are of clear clinical effect (Buysse et al., 2009). Systematic assessment of the population frequency of CNVs at different size ranges shows important increases in large CNVs in the affected cohort compared to the control group (Cooper et al., 2011). While 8% of the general population carries a CNV larger than 500kb, almost 25% of patients with intellectual disability have larger CNVs. (Itsara et al., 2009; Cooper et al., 2011; Coe at al., 2012). It is possible that very large CNVs can be benign in nature (Barber 2005; Filges et al., 2009; Itsara et al., 2009; Bateman et al.,2010) and very small CVs can be clinically important (Nowakowska et al., 2010; Nowakowska, 2017).

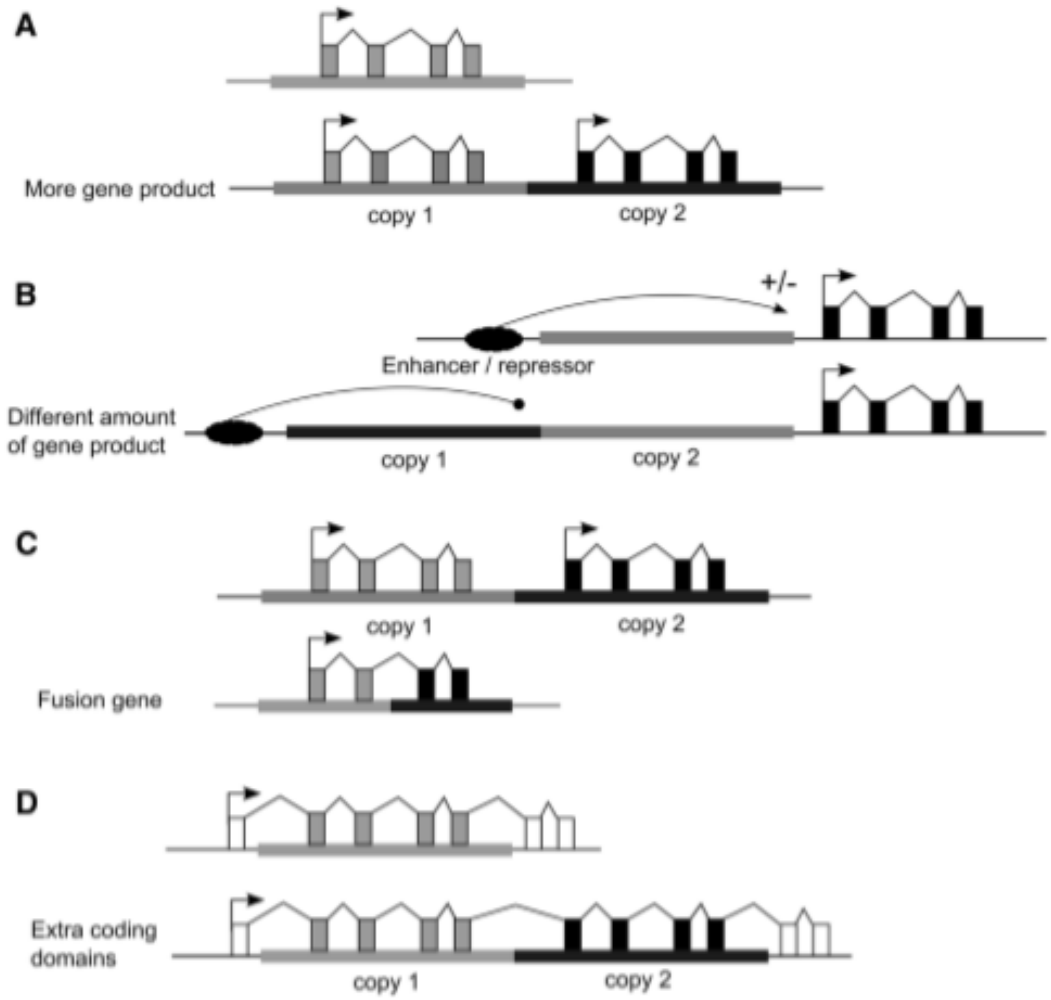


Figure 1. 5 The states that CNV can affect phenotype. A Gene dosage effect. The altered number of copies of the full gene affect the total amount of mRNA and protein encoded by the gene. B Position effect. The CNV affects the distance between the gene and a regulatory element that can be an enhancer or a repressor. C Fusion gene. For this figure, a deletion because of an unequal crossing over between the two copies of DNA sequence results in a fusion gene. D Extra protein coding domains. CNV could alter the number of tandem repeats of coding exons within a gene so that alters the number of functional protein domains, and final size of the protein would be altered. Obtained from Hollox and Hoh, 2014.

1.3 CNV Detection methods

Several techniques have been described to detect and measure CNVs in the human genome. However, there is no single existing methodology has the scope for accurately genotyping all CNV classes.

1.3.1 PCR techniques

1.3.1.1 Quantitative PCR (qPCR)

qPCR is a high throughput technique to measure gene copy number. The measurement of PCR amplicon accumulation in real time is the main principle of this method. The fractional cycle number (Ct) indicates the amount of starting template, when PCR amplification reaches a defined threshold during the exponential phase of the reaction. qPCR primarily involves using fluorescent techniques, either DNA intercalating dyes such as SYBR green or probe-based methodologies such as TaqMan, Scorpion and molecular beacons where the Ct between the target gene and a reference gene is compared. The generated ΔCt values are then used to determine the copy number. qPCR is normally used as a justification technique for computationally identified loci (Li and Olivier, 2013).

For large scale association studies, qPCR has been a good method of choice regarding to the advantages of simple workflow procedure, high throughput capabilities and cost effectiveness. It requires relatively small amounts of DNA compared to many other methodologies for CNV detection. However, genotyping multi-allelic CNV regions is significantly a limitation for the qPCR application. The studies comparing Paralogue Ratio Test (PRT) and qPCR have resulted that qPCR can be affected by differential sample preparation, storage conditions and DNA degradation Additionally, qPCR can generate results where there is extensive overlap between copy number integer classes (Aldhous et al., 2010; Fode et al., 2011). When qPCR was compared to Multi-plex Ligation-dependent Probe Amplification (MLPA) for analysis of beta-defensin, it is yielded markedly different results using qPCR, often mis-scoring CN by several copies (Perne et al., 2018). The comparison between qPCR and other established techniques have important implications for large scale association studies. The inability of qPCR to multiplex more than one locus for CNV analysis means that multiple measurements cannot be performed simultaneously. Therefore, a consensus result can be obtained by performing many replicates (Cantsilieris et al., 2013).

1.3.1.2 Multiplex ligation-dependent probe amplification (MLPA)

MLPA is a useful method that relies upon hybridization and ligation of two adjacently located oligonucleotides to a specific genomic DNA sequence. All probes have identical 5'sequences that allows the amplification with a single primer pair. Typically, capillary sequencer can be used to separate the products based on their size. Then, the resultant fluorescence intensities are transferred for further analysis. The powerful side of his method is the sensitivity of the ligation step for analysing sequences of high identity, by designing probes with mismatches at the ligation site. It is effective for studying SDs because a probe can distinguish two sequences differing only by a single mismatch at the ligation site. However, it could be limited if sequence polymorphisms exists at and around the ligation site which can disturb ligation sufficiently to give the appearance of an apparent deletion event.

Multi-allelic CNPs can be calculated by this assay. MLPA analysis was performed on the NSF gene, and the discrete copy number classes ranging from 2 to 7 copies with distinct differences between each of the subgroups were identified (White et al., 2007). This methodology has also been shown to be effective for *theFCGR3B* region (Marques at al., 2010). The "nearest neighbour approach" was also utilized to analyse beta-defensins region, where each test probe was normalized against the sum of five nearest neighbour reference peaks (CN of 2). An average of 10 beta-defensin test probes plotted on a scatter plot demonstrated discrete steps corresponding to a stepwise increase in copy number integers from 2 to 9 copies (Groth et al., 2008).

The advantages of MLPA technique are the number of loci can be analysed in a single reaction, the specificity of the ligation step, the reliability and accuracy of CNV measurement, and the relative low cost for conducting large scale association studies. However, there are multiple steps required to complete the whole procedure so that possible errors can be introduced at several points (Cantsilieris et al., 2013).

1.3.1.3 Paralogue ratio test (PRT)

PRT is a comparative PCR–based method where single primer pairs are used to amplify a putative CNV target locus (test) relative to a single copy reference locus (Armour et al. 2007). In this case, the primers need to be designed carefully to estimate actual copy number of gene by targeting paralogous sequences (Veal et al., 2013). The PCR is performed under quantitative conditions (30 cycles) and the PCR amplicons of the test and reference locus can be distinguished easily by capillary electrophoresis based on their small size difference (Figure 1.6). Experiments can be performed in duplicate by using two different fluorescent dyes to label the same primer to generate concordant data. Raw copy integer number can be determined by calculating the peak area ratio between the test and the reference (Armour et al., 2007). An estimation of the integer copy number is achieved by combining all raw values for each assay and is calculated using maximum-likelihood approach. This methodology is particularly suited for analysis of complex regions of the genome such as SDs. Inter-chromosomal repeat sequences are preferable as they are unlikely to be linked to the CNV locus. PRT assays have been successfully applied in several multi-allelic CNVs including beta-defensin, CCL3L1, FCGR3B and Complement Component 4 (C4) (Aldhous et al., 2010; Hollox et al., 2009; Carpenter et al., 2011; Fernando et al., 2010).

PRT has been shown to be a better technique compared to qPCR in terms of determining integer copy number and accuracy in copy number measurement >4 (Aldhous et al., 2010; Fode et al., 2011). Furthermore, the workflow is fast as it does not require the long overnight hybridization steps that are components of MLPA. It is a relatively inexpensive for determining copy number in large scale association studies and requires low quantities of genomic DNA (10–20 ng). It has been shown as reliable as other well validated methods such as MLPA (Armour et al., 2007).

One limitation is the assumption that the paralogous reference loci do not itself vary in copy number; the sites of primer binding are free from polymorphisms that can affect primer binding efficiency. As a technical limitation, regions of high sequence identity, such as FCGR3A/FCGR3B andC4A/C4B do not contain the sequence differences required for primer specificity. To avoid this, genes can be amplified and cut by using restriction enzymes (with a digestion time >4 h) (Hollox et al., 2009; Fernando et al., 2010). Lower

multiplexing capacity compared to MLPA is another limitation of PRT (Cantsilieris et al., 2013).

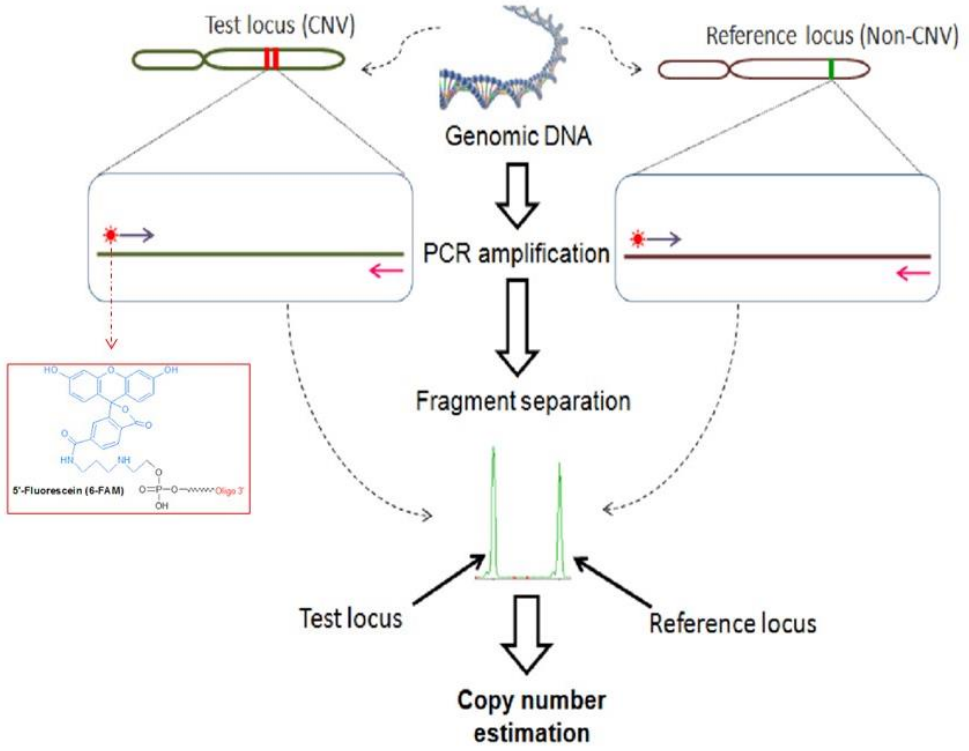


Figure 1. 6 Schematic picture describing different steps of PRT. Both test (CNV) and reference regions are PCR amplified using fluorescent labelled primers (shown as red asterisks). The amount of PCR products is quantified with capillary electrophoresis and copy number is estimated by comparing amount of test products with reference products. Adapted from Hollox and Hoh, 2014.

1.3.2 Array based methods

Two different DNA chip-based methods can be used to detect genome-wide CNVs: CGH-based CNV detection or SNP array-based CNV detection. The differences between these two methods depend on the hybridization method with the type of microarray used (Lee and Jeon, 2008).

1.3.2.1 Array comparative genomic hybridization (aCGH)

aCGH is a method that relays on dual hybridization of test and reference DNA to either immobilized short oligonucleotides or long DNA sequences such as bacterial artificial chromosomes (BACs). The signal ratio between test and reference sample is used for normalization and then to infer copy number. Basically, the reference and test DNA samples are differentially labelled with fluorescent tags (Cy5 and Cy3, respectively),

repetitive-elements are blocked using COT-1 DNA and then hybridised to genomic arrays. After hybridisation, the fluorescence ratio (Cy3:Cy5) reveals copy-number differences between the two DNA samples. If the intensities of the fluorescent dyes are equal on one probe, this is interpreted as having equal quantity of DNA in the test and reference samples; if there is an altered ratio (Cy3:Cy5) this indicates a loss or a gain of that specific genomic region (Feuk et al., 2006). Initial methods using BAC clones (with resolution 100 to 200 kb) provided important understandings into the landscape of structure variation in the human genome (Redon et al., 2006). On the other hand, weak breakpoint resolution can cause overestimation of CNV size. Following studies using long oligonucleotide arrays (with resolution between 0.5 and 2 kb) obtained more reliable picture of the CNV landscape (Conrad et al., 2010; Cantsilieris et al., 2013). This method provides clear indication of the presence of a CNV region on particular location, however, it cannot give accurate information about the precise copy number of a sequence.

1.3.2.2 SNP microarrays

SNP microarrays have the advantage of analysing both single nucleotide differences and in some cases non-polymorphic copy number probes that are not restricted by sequence properties of SNPs. Compared to Array CGH, SNP microarrays analyse a single sample per microarray and compare signal intensities from a sample with clustered intensities from a set of reference samples, or the whole sample population to generate a log ratio (McCarroll et al., 2008). SNP microarrays produce two types of fluorescent information. The first one is the total fluorescence gained from intensity of both alleles. The second is the allelic ratio gained from the relative intensity of each allele. Three different types of clusters should emerge corresponding to genotypes by plotting the normalized intensity of each allele against one another. Results indicative of a deletion can result in three additional clusters corresponding to reduced intensity of homozygous alleles indicative of a heterozygous deletion, or no signal at all indicative of a homozygous event. The intensity of one allele as a proportion of the total allele signal, can also be used as an additional measure to confirm the presence of copy number polymorphisms. This is useful for accurately predicting CNPs of 0–4 diploid copy numbers (Cantsilieris et al., 2013).

1.3.2.3 Application of arrays to CNV genotyping

The deep understanding of complex CNVs is certainly dependent on the genetic properties of the region, the probe density/performance as well as normalization parameters (Conrad et al., 2010). Commercial SNP arrays such as Affymetrix and Illumina SNP arrays do not efficiently cover regions of genome complexity such as SDs. In addition, using customized tiling array CGH could only reliably genotype 61% of CNPs that map to SDs parameters (Conrad et al., 2010). In general, SDs show higher levels of false positive and false negative call rates for both SNP and Array CGH platforms when compared to unique regions of the genome. On the other hand, the Nimblegen and Agilent array CGH platforms have more probes in SD regions than do Affymetrix and Illumina SNP arrays (Pinto et al., 2011). Large studies using customized array CGH platforms showed that duplications and multi-allelic loci are more problematic to detect than deletion variants. Particularity for the multi-allelic CNV regions, the choice of normalization algorithm impacts the resolution of the data and each individual locus must be treated separately to obtain the more reliable results (Cantsilieris et al., 2013).

The microarray platforms have the ability to screen CNVs on a genome-wide level at relatively low cost in large data sets. Generally, the parameters for accurate genotyping are set at five consecutive probes and a minimum size of 1 kb and most single channel array platforms lose sensitivity to identify variants below 10 kb. However, the capability of commercial microarray platforms is generally poor to detect breakpoints and smaller rearrangements. Structural rearrangements that do not affect copy number (such as inversions and translocations) will not be detected. Copy number estimation is relative and it relies on the diploid copy number for a given region is two. That is not always the case especially in regions of segmental duplication. In addition, the characterization of the reference sample is critical for the reliability of Array CGH CNV calls. Because, a loss in the reference sample can be interpreted as a gain in the test sample even though the test sample may have a diploid copy number of two (Cantsilieris et al., 2013).

The compares and contrasts of the above-mentioned methods and techniques for typing copy number variations is summarized in Table 1.2. NGS techniques will be explained in detail in the section 1.3.3.

Table 1. 2 Methods to measure copy number variations.

	qPCR[a]	MLPA[a]	PRT[a]	Array CGH[b]	SNP Array[b]	NGS
Detection	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Absolute copy number
Sample	5-10ng DNA	100-200ng DNA	5-10ng DNA	0.5-1 µg DNA	0.5-1 µg DNA	1-2 µg DNA
Loci	Single	>40	Single	>2 million	>2 million	Genome wide
Minimum theoretical resolution	100bp	100bp	100bp	5-10kb	5-10kb	>1kb
Cost/sample	Low	Low	Low	Moderate	Moderate	High
Time to result	4h	>24h	4h	>24h	>24h	2-3 days
Labour requirement	Low	Low	Low	Moderate	Moderate	High

a Minimum resolution is in general the length of a single prob
b High resolution Array CGH can achieve a minimum resolution of >500 bp
Adapted from Cantsilieris et al., 2013.

1.3.3 Sequence based methods

1.3.3.1 Next generation sequencing

DNA sequencing is identified as the determination of the precise sequence of nucleotides in a particular sample of DNA. HTS or Next-generation sequencing (NGS) is a contemporary sequencing technology which aims to determining the whole genomic sequence and allows to generate fast, cheaper and huge amount of sequence information compared to DNA sequencing by utilizing massive parallel sequencing (Ansorge, 2016).

The general outline for all second-generation platforms is very similar. There are three main steps which are library preparation, amplification, and sequencing. NGS library preparation can be performed in several steps. Fragmenting and/or sizing the target sequences to a desired length, attaching oligonucleotide adapters to the ends of target fragments, and quantitating the final library product for sequencing (Head et al., 2015). Library preparation from DNA samples for sequencing whole genomes, targeted regions within genomes (e.g exome sequencing), ChIP-seq experiments, or PCR amplicons follows the same general workflow. Library preparation from RNA is done by capturing

mRNA, random priming, and complementary DNA (cDNA) synthesis followed by the end polishing and adapter ligation (Ambardar et al., 2016).

The size of the target DNA fragments in the key parameter for NGS library construction. Fragmentation is a crucial step for preparation of template libraries, and this can be done in different ways. Physical/mechanical methods such as nebulization and ultrasonication, enzymatic methods such as non-specific endonuclease cocktail and transposases such as tagmentation reactions. Desired library size is determined by the desired insert size which refers the portion between the adapter sequences because the length of the adaptor sequences is a constant with known sequences. The optimal insert size depends on the limitations of the NGS instrumentation and the specific sequencing application. The known adapter sequences are used to amplify the insert DNA by PCR (Head et al., 2015; Ambardar et al., 2016). Size selection can be commonly done with gel electrophoresis-based method (the adapted fragments are run on a gel to separate the fragments by size and the band corresponding to the size of interest is collected) or bead-based size selection method (magnetic beads are used with varying concentrations of buffers to isolate the DNA fragment sizes of interest). As final step, the library quantification (library concentration and fragment size information) is performed for high quality downstream analyses because accurate library quantification is important for successful template preparation and sequencing.

The amplification process can take place in emulsion (Emulsion PCR) or in solution (Bridge PCR) to create DNA clusters. (Figure 1.7). In emulsion PCR amplification, DNA is fragmented, and adapters are added onto these fragments. The double stranded DNA with adapters is denatured and the reverse strand of single stranded DNA fragments are then attached to the beads which have matched primers. Each bead is emulsified in a water-oil mix with PCR reagents (DNA polymerase, dNTP, buffer, primers) in a micro well. PCR is conducted in these droplets and the process is repeated 30-60 cycles to create the DNA clusters. Following PCR, the emulsion is broken by beads and then products can be enriched and sequenced (Shendure and Ji, 2008). In bridge PCR, DNA is fragmented, and adapters added on both ends. Single stranded DNA molecules are created by denaturing and are then loaded onto flow cells which have corresponding adapter sequences. After adding dNTPs and DNA polymerase for elongation, the single

DNA strands make bonds to the surface of the cell by using complementary primers which constructs bridged structures. The enzyme creates double strands which later stay attached to the surface closely when the denaturation occurs. The step is repeated until it forms dense colonial clusters of identical DNA molecules (Shendure and Ji, 2008). Roche 454 and Ion Torrent rely on emulsion PCR on beads while Illumina relies on bridge PCR on the flow cell surface.

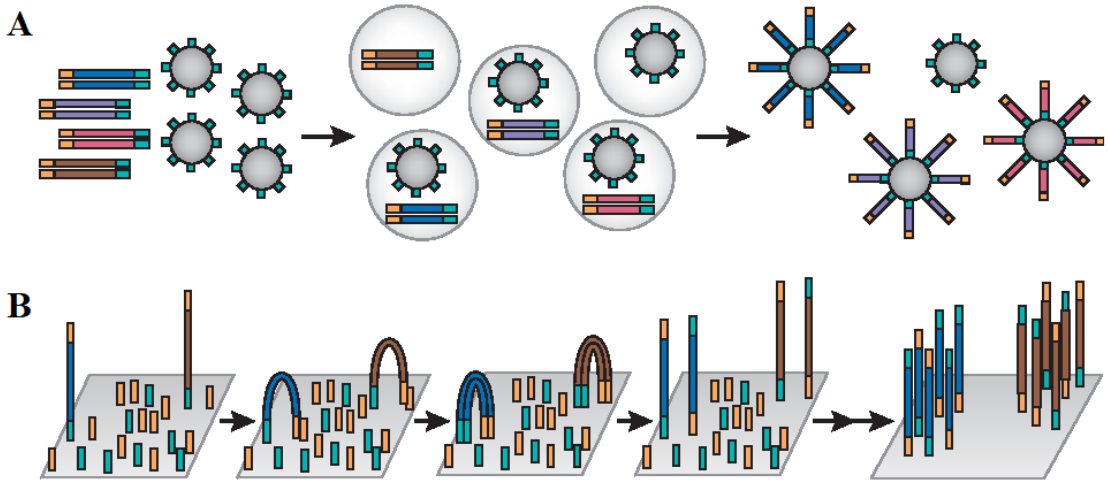


Figure 1. 7 Schematic diagram showing the workflow of NGS. A: An in vitro–constructed adaptor flanked shotgun library (shown as gold and turquoise adaptors flanking unique inserts) is PCR amplified in the context of a water-in-oil emulsion. One of the PCR primers is tethered to the surface (5'-attached) of micron-scale beads that are also included in the reaction. After breaking the emulsion, beads bearing amplification products can be selectively enriched. Each clonally amplified bead will bear on its surface PCR products corresponding to amplification of a single molecule from the template library. B: an in vitro–constructed adaptor-flanked shotgun library is PCR amplified, but both primers densely coat the surface of a solid substrate, attached at their 5' ends by a flexible linker. As a consequence, amplification products originating from any given member of the template library remain locally tethered near the point of origin. At the conclusion of the PCR, each clonal cluster contains ~1,000 copies of a single member of the template library. Obtained from Shendure and Ji, 2008.

The sequencing step performed by the various available methods as per the manufacturer's orders (Figure 1.8). 454 pyrosequencing uses sequencing by synthesis approach during extension. In this pyrosequencing, pyrophosphates are released when nucleotides are enzymatically added to the DNA chain. Each phosphate (per nucleotide) is converted into ATP which later contributes to the chemical activity to produce oxyluciferin from luciferin by catalysing with luciferase so that light can be produced. The signal intensity of light is detected with a camera, and the number of identical

nucleotides used can be measured since light signal is proportional to the incorporated number of nucleotides (Kircher and Kelos, 2010). Ion torrent semiconductor sequencing is also based on the pyrosequencing principle. Instead of phosphates, this system detects H+ ions by using a semiconductor chip. The nucleotides are introduced one at a time. When a nucleotide is incorporated, a hydrogen ion is released. This ion is then detected by a pH sensor. The change in pH/signal intensity is proportional to the amount of nucleotide added (Ambardar et al., 2016). Illumina sequencing relies on sequencing by synthesis approach. Similar to other methods, the genome is fragmented to manageable shorter fragments (>300), and adapters are ligated to these fragments. After denaturation, the single stranded DNA is bound to the surface of flow cell channels. The fluorescently labeled dNTPs (terminators) are then added to the flow cells. After primer binding, DNA polymerase attaches one fluorescently labelled terminator to the strand. The emitted fluorescence on the nucleotide is detected by a camera and to be recorded. Once a base has been added to the new DNA strand, no other bases can be added per cycle. Then, the protection group and the fluorescent molecules are removed from the base. The next cycle repeats the process and it continues until all the clusters have been sequenced. The bridge amplification reactions occur on the surface of the flow cell leading the generation of DNA clusters which later provides strong emitting signal to be detected by a camera (Shendure and Ji, 2008). All Illumina platforms use the same basic idea (sequencing by synthesis), however, they run the output differently.

The system 10X Genomics Linked-Read sequencing, by GemCode Technology, provides sequencing applications with Illumina sequencers. In this platform, GemCode Chips are used for DNA samples to be inserted, including gel beads and a partitioning oil. The long DNA molecules are separated by the oil into different gems inside the instrument's microfluidics. Then, the DNA can be fragmented into short-read libraries with unique molecular barcodes based on the gem origin, leading to Illumina sequencing as usual principal. The 10X software makes use of these molecular tags to generate the original long fragments. The 10X Genome sequencing approach is very promising for understanding of the long-range genetic information (Ambardar et al., 2016; Mostovoy et al., 2016).

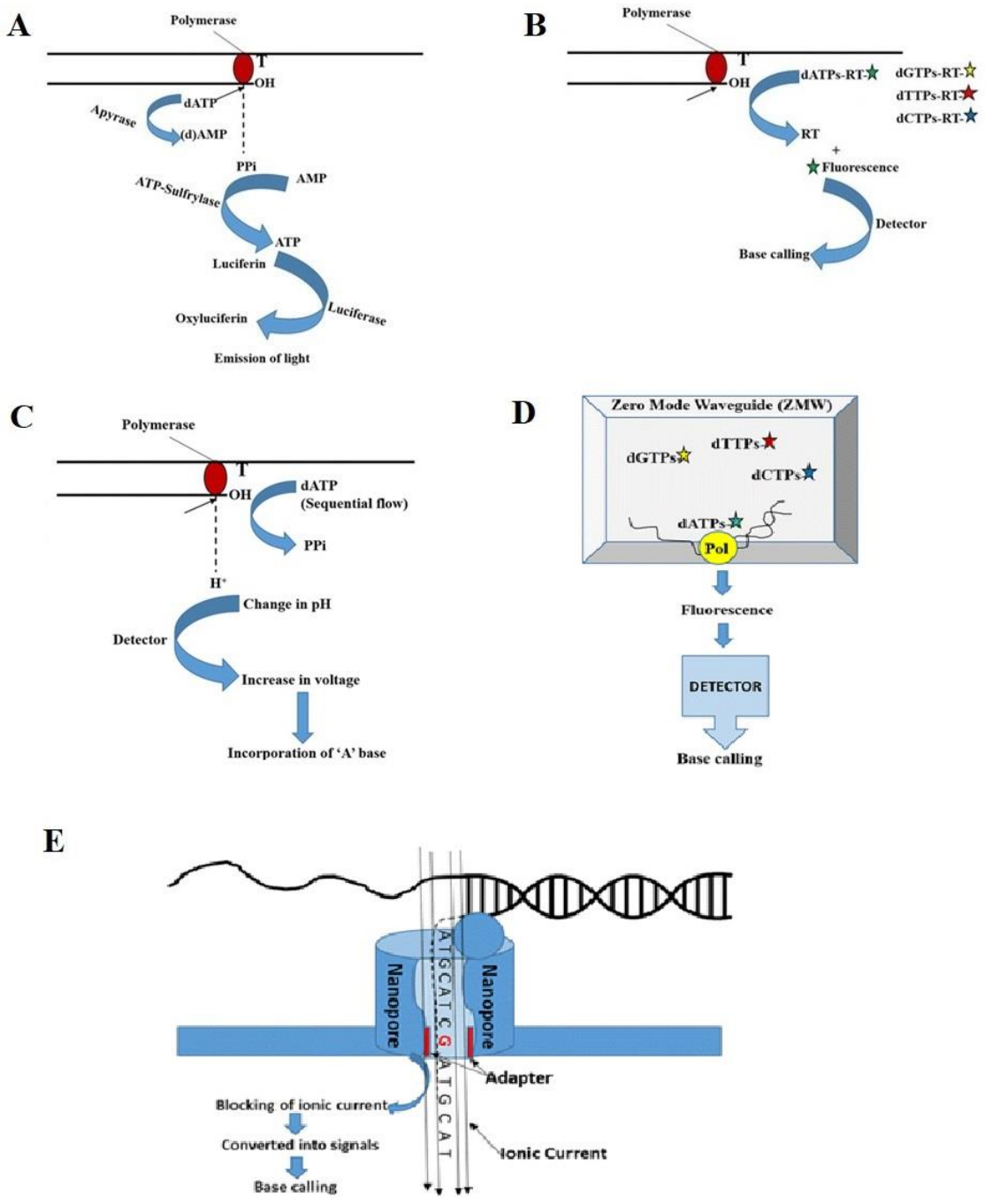


Figure 1. 8 Diagrammatic representation of nucleotide incorporation in different NGS platforms. The representation of (A) enzymatic reactions involved in pyrosequencing, (B) Sequencing by reversible termination (Illumina), (C) pH change involved in Sequencing by detection of hydrogen ions (Ion Torrent), (D) single molecule real time (SMRT) sequencing in Zero-mode waveguide (ZMW), during complementary nucleotide incorporation, and (E) nanopore sequencing during base calling. Adapted from Ambardar et al., 2016.

There are especially two limitations shared by second generation sequencing platforms. First, the short-read length that need to be assembled by using numerous bioinformatic tools/ pipelines into original length template. Second, PCR bias introduced by clonal amplification, for detection of base incorporation signal. The third generation of high throughput NGS technology was developed to handle these limitations (Ambardar et al., 2016). In addition, it promises to increase throughput with decreased costs, run times and error rates with the requirement of minimal input material. These long-read sequencing technologies target to sequence single DNA template instead of sequencing clonally amplified template. This also requires minimal use of biochemicals leading to miniaturization of whole process to nanoscale.

Pacific BioSciences (PacBio) developed an approach which combines nanotechnology with molecular biology to sequence single molecule known as single molecule real time (SMRT) sequencing. The sequence information is captured during the replication process of the target DNA molecule. SMRT cells contain zero-mode waveguides (ZMWs) which provides the smallest obtainable amount for light detection (Figure 1.8). A single polymerase is immobilized at the bottom of each ZMW, which can bind to either hairpin adaptor of the SMRTbell and start the replication. After immobilization of DNA template and polymerase at the bottom of ZMWs, fluorescently labelled dNTPs are added which generate distinct emission spectrums. When nucleotides are incorporated to 3' end of the DNA strand, a light is produced. ZMWs detect and call the fluorescently labelled nucleotides. Then the attached fluorophore is released. The step is repeated for several times The replication processes are recorded by a movie of light pulses, and the pulses corresponding to each ZMW can be inferred to be a sequence of bases (called a continuous long read, CLR) (Rhoads & Au, 2015; Ambardar at el., 2016).

A different long-read technology has been developed by Oxford Nanopore Technologies which incorporate nucleotides using nanopore technology. While the DNA sequence passes through a nanopore that has an internal diameter of 1 nm, the electrical conductance of the pore is changed, and signal is detected (Figure 1.8). This technology involves the use of protein nanopores embedded in the polymer membrane. The sequencing does not need any intervening PCR amplification or a chemical labelling step.

In addition, sample preparation is not needed as cell lysate can be directly sequenced (Ambardar at el., 2016; Lu et al., 2016).

1.3.3.2 Sequencing Technologies

DNA sequencing technologies have existed since the early 1970's. The Sanger method was the primary sequencing technology between 1975 and 2005. High quality sequences (500-1000bp long) can be produced by Sanger sequencing which has long been considered the gold standard for sequencing DNA. The introduction of pyrosequencing technology by 454 Life Sciences in 2005 began the "next generation sequencing" (NGS) revolution. This high throughput technology allowed the generation and detection of thousands to millions of short sequencing reads in a single machine run without the need for cloning. Since then, many other NGS technologies have emerged that generate both short (50 – 400 bp) and long reads (1 – 100 kb) (Besser et al., 2018). Several of NGS platforms, using a variety of different chemistries, are now broadly available for high throughput DNA sequencing. The most commonly used NGS platforms are Illumina (HiSeq and MiSeq), and Ion Torrent (PGM and Proton), which can be classified as second-generation methods and platforms developed by Pacific Biosciences and Oxford Nanopore, which can be classified as third-generation methods. Addition to these current NGS platforms, more sequencing technologies are still being developed (Ansorge, 2016). A brief description of the main platforms and their performance is provided below (Table 1.3). The short-read sequencing platforms can be differentiated based on their engineering, sequencing chemistry, output (length of reads, number of sequences), accuracy and cost (Buermans and den Dunnen, 2014**).** The most widely used family of short-read sequencing technologies is claimed to be manufactured by Illumina. Illumina provide a variety of options which create a range of differing quantities and lengths of DNA sequences depending on the needs of the user (Haynes at al., 2019). For example, Illumina HiSeq X Ten, which is a suite of ten instruments can produce up to 1.8 trillion bases. Different platforms that produce short reads are offered, have their own advantages and disadvantages such as Ion Torrent technology (e.g. cost per base as advantage and higher homopolymer error rate as disadvantage). The primary advantage of long read platforms is that their capability to produce relatively long reads. The Sequel as the most recent PacBio system has the potential to generate a million DNA sequences

per run, however, this is far fewer than that generated by the highest throughput short read platforms. Initial concerns about high error rates have been tackled by circularising the DNA molecule to produce an accurate consensus sequence. However, not all reads produced will be high quality reads (Rhoads & Au, 2015). The device MinION developed by Oxford Nanopore Technologies was introduced has been commercially made available since 2015. This sequencer has 512-2000 nanopores with each nanopore having the sequencing speed of 120-1000 bases per minute. It looks like a USB stick and can be used only one time. This sequencer still has issues surrounding error rate, but as new pore chemistries are being brought online, this is improving (Ambardar at el., 2016; Haynes at al., 2019).

Table 1. 3 Comparison of performances of NGS platforms.

Platform \Instrument	Throughput range (Gb)*	Read Length (bp)	Strength	Weakness
Sanger Sequecing				
ABI 3500/3730	0.0003	Up to 1kb	Read accuracy and length	Cost and throughput
Illumina				
MiniSeq	1.7-7.5	1×75 to 2×150	Low initial investment	Run and read length
MiSeq	0.3-15	1×36 to 2×300	Read length, Scalability	Run length
NextSeq	10-120	1×75 to 2×150	Throughput	Run and read length
HiSeq (2500)	10-1000	1×50 to 2×250	Read accuracy, throughput, low per sample cost	High initial investment, run length
NovaSeq 5000/6000	2000-6000	2×50 to 2×150	Read accuracy, Throughput, Low per sample cost	High initial investment, run and read length
IonTorrent				
PGM	0.08-2	Up to 400	Read length, Speed	Throughput, homopolymers***
S5	0.6-15	Up to 400	Read length, Speed, Scalability	Homopolymers***
Proton	10-15	Up to 200	Speed, Throughput	Homopolymers***
Pacific BioSciences				
PacBio RSII	0.5-1**	Up to 60 kb (Average 10 kb, N50 20 kb)	Read length, Speed	High error rate and initial investment, low throughput
Sequel	5-10**	Up to 60 kb (Average 10 kb, N50 20 k)	Read length, Speed	High error rate
Oxford Nanopore				
MInION	0.1-1	Up to 100 kb	Read length, Portability	High error rate, run length, low throughput

*The throughput ranges are determined by available kits and run modes on a per run basis. As an example of a 15 GB throughput, thirty-five 5 MB genomes can be sequenced to a minimum coverage of 40x on the Illumina MiSeq using the v3 600 cycle chemistry. **Per one SMRTcell ***Results in increased error rate (increased proportion of reads containing errors among all reads) which in turn results in false positive variant calling. Besser et al., 2018.

1.3.3.3 CNV detection methods with NGS

Genome sequencing technologies have become very useful in a wide range of genomic analyses and have helped to detect genomic variations in human genome including CNVs. There are four main methods for detecting CNVs with short read sequencing data: (1) Read-pair (RP), (2) Split- read (SR), (3) Read-depth (RD), and (4) Assembly based (AS) methods (Figure 1.9).

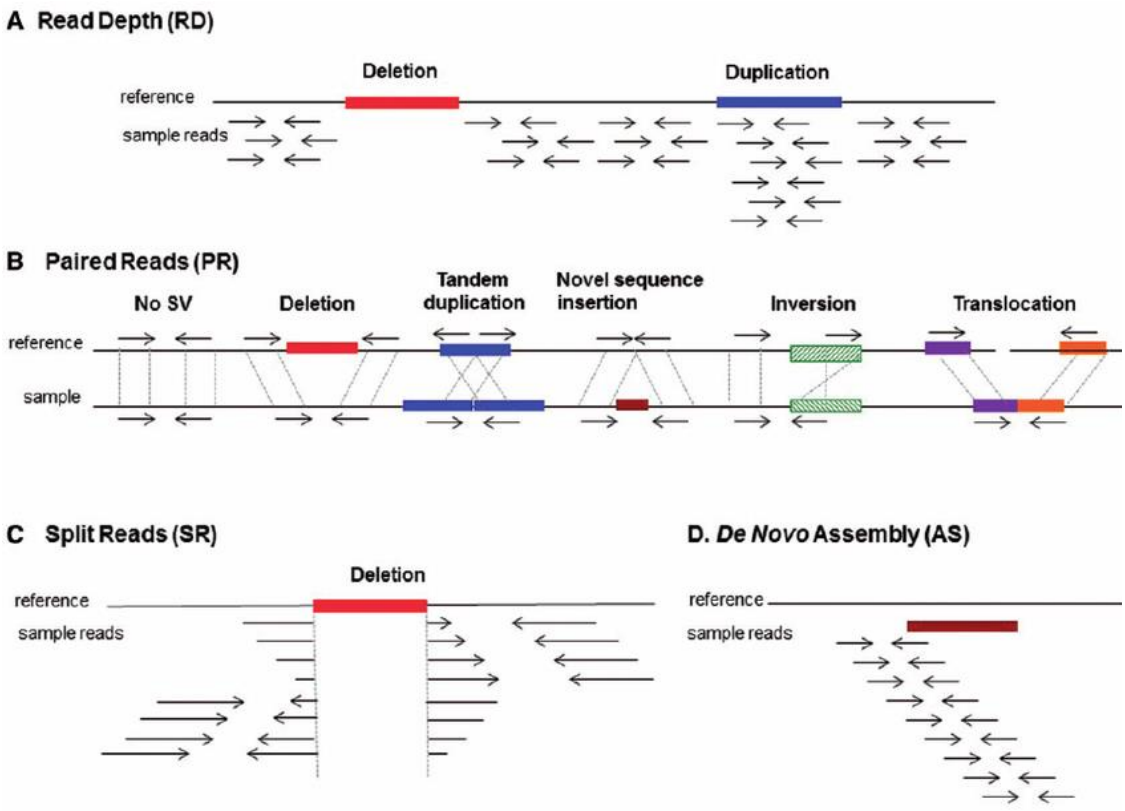


Figure 1. 9 Methods for detecting CNVs with NGS data. (A) Read depth. Reads are aligned into the reference genome and when compared to diploid regions they show a reduced number of reads in a deleted region or higher read depth in a duplicated region. (B) Paired reads. Pairs of sequence reads are mapped into the reference genome (from left to right): (1) no SV, pairs are aligned into correct order, correct orientation and spanned as expected based on the library's insert size; (2) deletion, the aligned pairs span far apart from that expected based on library insert size; (3) tandem duplication, read pairs are aligned in unexpected order, where expected order means that the leftmost read should be aligned in the forward strand and the rightmost read in the reverse strand; (4) novel sequence insertion, the pairs are aligned closer from that expected based on library insert size; (5) inversion, read pairs are aligned in wrong orientation, both reads align either in forward or reverse strand; and (6) read pairs mapped to different chromosomes. (C) Split reads. Sequenced reads pointing to the same breakpoint are split at the nucleotide where the breakpoint occurs. The corresponding paired read is properly aligned to the reference genome. (D) De novo assembly. Sample reads from novel sequence insertions are assembled without a reference sequenced genome. Obtained from Escaramis et al., 2015.

The read pair (RP) approach was the first method used for the detection of CNV. For the paired-end sequencing, the DNA fragments are required to show a specific distribution of the insert size (Korbel et al., 2007). RP method compares the average insert size between the actual sequenced read-pairs with the expected size based on a reference genome. CNVs are identified from discordantly mapped paired-reads whose distances are significantly different from the predetermined average insert size (Zhao et al., 2013; Pirooznia et al., 2015). These types of methods are very suitable to detect structural variations such as deletions, insertions of novel sequences and inversions (Zhao et al., 2013). While they can efficiently identify medium-sized insertions and deletions, they are insensitive to small insertion or deletion from mapped data. Besides, sequence insertions that are larger than the average insert size which cannot be detected (Medvedev et al., 2009; Tattini et al., 2015). Additionally, RP methods often cannot detect CNVs in segmental duplication-rich regions (Pirooznia et al., 2015).

Split read (SR) methods also use pair end sequencing data. While one read of the pair is properly mapped to the reference genome, the other read completely or partially fails to align to the genome (Zhao et al., 2013). The unmapped or partially mapped reads are potential sources for breakpoints at the single base pair level for CNVs. These unmapped and incompletely mapped reads are then divided into multiple fragments. The first and the last fragment of each split read are mapped to the reference genome independently. Therefore, this remapping step shows precise start and end positions of the fragments that are insertion/deletion events (Pirooznia et al., 2015). SR methods are good for detecting deletions and small insertions (Alkan et al., 2011), however, they are limited in their ability to identify large-scale detection of structural variations (Pirooznia et al., 2015). Additionally, these methods rely on the length of reads and are applied to specific regions in the reference genome (Zhao et al., 2013).

Assembly based (AS) methods theoretically can be used for detecting all types of genetic variations including CNVs. AS methods first create contigs from short reads by assembling overlapping reads. Then, these assembled contigs are compared to a known reference genome to identify the genomic regions with discordant copy numbers. AS methods are rarely used to detect CNVs due to extensive computation and perform

weakly on repeat regions (Zhao et al., 2013). Besides, they are powerless to handle different haplotypes, and only homozygous structure variations can be identified.

Read depth (RD) is one of the important advantages of the NGS technologies. Detecting important differences in CNVs among individuals can be achieved by using this approach. RD methods investigate the divergence from a random distribution in mapping depth to determine deletions and duplications (Bailey et al., 2002) and they are based on the assumption that the depth of the coverage of a genomic region is correlated with the copy number of the region (Teo et al., 2012). When sequences are mapped to a reference genome, some reads have higher read mapping (high read depth) which can indicate a duplication region and some reads have lower reads mapping (low read depth) which can indicate a deletion region. Therefore, RD can detect the existence of both deletions and duplications, and high coverage sequence data can be used to detect sequence variations of CNVs (Alkan et al, 2011; Forni et al, 2015). RD methods can be classified into three categories: single sample, paired case/control samples, and a large population of samples. In the single sample category, only the absolute copy number is stated, and the read depth is estimated by using mathematical models since there is no information from other samples or controls. In the presence of case and controls, the relative copies compared to controls are stated. In population-based studies, the discordant copy numbers for every individual are discovered and the overall mean of the read depth from numerous samples is used to detect CNVs (Zhao et al., 2013). The pipeline of the software programs for detecting CNV by using RD are like each other and can be summarized in four steps. Mapping: the raw data (short reads) are taken as input which is later used for mapping. Considering the read length, type of the reads (single or paired), and the platform which reads are generated, the parameters for the mapping are adjusted to minimize the false positives and false negatives. The mapped reference genome is first divided into non-overlapping segments, and then the number of mapped reads is counted. Because the existing methods to find CN in NGS data use depth of coverage, this can be represented as log read counts or read counts in an interval. Normalization: the non-overlapping segments undergoes the sample normalization process and GC-correction based on the assumption of distribution of the reads along the segments. Estimation of copy number: the accurate copy number is

determined by considering the gain or the loss of a region. Segmentation: the genomic regions having the same number of copy number are united to detect discordant copy number regions. While joining consecutive segments with a segmentation algorithm to call CNVs, each RD program uses different or improved statistical approaches (Zhao et al., 2013). Overall, the RP and SR methods only identify the position of the potential CNVs but not the counts of CNVs. RD works more efficiently on large size of CNVs compared to RP and SR methods (Yoon et al., 2009; Pirooznia et al., 2015). Additionally, RD methods are best suited for detecting absolute copy number and detect larger CNVs (Alkan et al., 2009), however, they may not be efficient enough to determine small CNVs and reporting exact breakpoints (Zhao et al., 2013).

A considerable number of tools for the detection of CNV in whole-genome sequencing (WGS) data have been reported in the literature, including CNVnator, CNAnorm, CNAseg, rSWseq, cn.MOPS, JointSLM, ReadDepth, and BIC-seq. Each CNV prediction tool applies a different statistical approach to NGS data, and the tools can be chosen based on study purpose. Even though there are several tools used for CNV detection with different statistical approaches, the combined mapping programs can heal the results better. LUMPY is one of the combined methods which integrate multiple SV detection signals including read-pair, split-read, and prior knowledge if provided to improve sensitivity. Including raw read-depth data to LUMPY is expected to have significant improvements in the detection of CNVs (Layer et al., 2014). The prediction of CNV from exome data is also challenging. In a study, six different CNV tools (ExomeCNV, CONTRA, ExomeCopy, ExomeDepth, CoNIFER, XHMM) were used to predict exonic CNV discovery, short (1- 4 exon) CNVs, on 30 exomes from the 1000 Genomes project and 9 exomes from primary immunodeficiency patients. The validation of the CNV prediction was confirmed by using a custom CGH array. The performance of CNV prediction tools were compared by demonstrating the sensitivity and False Discovery Ratio (FDR). It was concluded that ExomeDepth is more sensitive than CoNIFER and XHMM to find rare exonic CNVs. On the other hand, CoNIFER has the lowest FPR for exonic CNVs detection (Samarakoon et al., 2014).

1.4 Short-read sequencing data processing for variant/SNP calling and genotyping

Meaningful analysis of NGS, produced widely by genetics and genomics studies, significantly depends on the accurate calling of SNPs and genotypes. Evaluating the raw output of NGS technology into a final set of SNP and genotype data involves several of steps. Each step contributes to the accuracy of the final SNP and genotype calls. A whole genome, or targeted regions of the genome, is randomly digested into small fragments (or short reads) that get sequenced and are then either aligned to a reference genome or assembled in short-read sequencing NGS methods (Flicek et al., 2009). Having aligned the fragments of one or more individuals to a reference genome, 'SNP calling' recognizes variable sites, while 'genotype calling' determines the genotype for each individual at each site (Nielsen et al., 2011). The SNP calling pipeline comprises several steps (Figure 1.10). Base calling is the first step of the analysis and is specialized for each NGS platform. As explained briefly above, this step evaluates the images created by an image-capturing device during the sequencing process. It generates the short reads by reading the signals and converting them into nucleotide bases. Additionally, a statistical model is applied to the nucleotide bases providing a measure of certainty for each base with a Phred quality score. The statistical models base their error estimate on information such as signal intensities from the recorded image, the number of the sequencing cycle and distances to other sequence colonies (Altmann et al., 2012). Base calling accuracy, measured by the Phred quality score (Q score), is the most common metric used to assess the accuracy of a sequencing platform and implies the possibility that a given base is called incorrectly by the sequencer. Q scores are defined as a property that is logarithmically related to the base calling error probabilities (P). The formula is:

$$\boldsymbol{Q}\textbf{phred} = \textbf{-10 x } \log_{10}\boldsymbol{P}\textbf{(error)}$$

When Phred gives a Q score of 30 (Q30) to a base, this is equivalent to the probability of an incorrect base call 1 in 1000 times and the probability of a correct base call is 99.9% (Table 1.4). Low Q scores can increase false-positive variant calls leading inaccurate conclusions and higher costs for validation experiments.

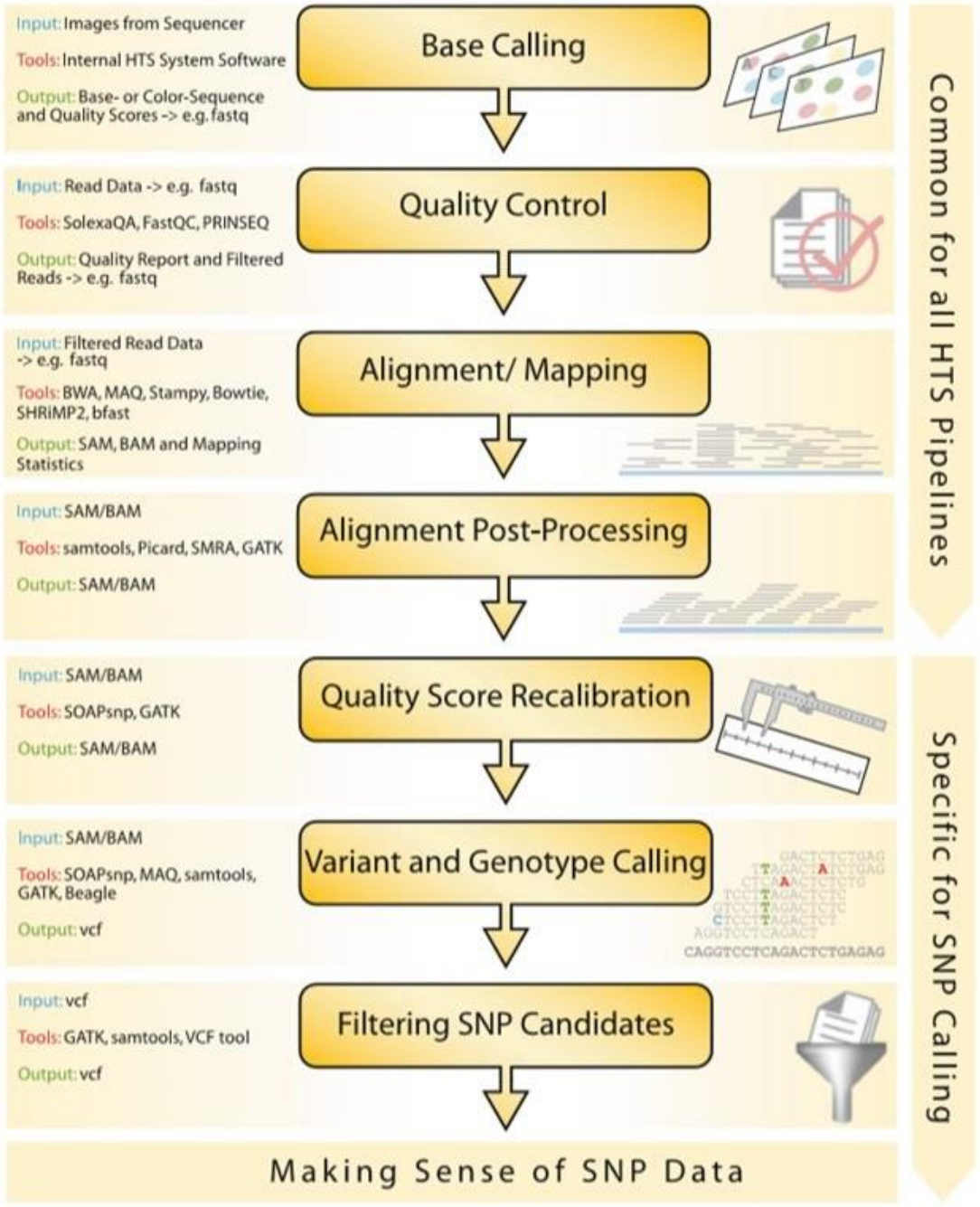


Figure 1. 10 A workflow of the SNP calling pipeline for whole exome sample dataset. The diagram briefly outlines the essential steps in the process of making variant calls. Each step has an input and output format which can be obtained by several recommended tools based on the purpose of the study. Obtained from Altmann et al., 2012.

Table 1. 4 Quality Scores and Base Calling Accuracy.

Phred Quality score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

In order to prevent misguided results and quick identification of the problem, a quality check of the data is a necessary and important step before performing any analyses (Patel and Jain, 2012). The initial base quality check is already done in the base calling step. A general quality check can provide very meaningful and useful information such as the distribution of nucleotides per read position, deviation from the expected GC content, overrepresented sequences, or average base quality per tile. The error probability of a base can increase with read length. This behaviour is typical for short-read sequencing NGS platforms. As a result, the filtering of the reads (trimming) is often performed for removing the bases from the end of the reads that are likely to contain sequencing errors in order to have a higher mappable read after quality check (Altman et al., 2012).

The mapping/alignment of the short reads is the next step after base calling and quality check (probably including trimming). The reliability of the alignment is very important in variant detection because misaligned reads may lead to errors in SNP and genotype calling. Therefore, alignment algorithms should be capable of both detecting sequencing errors and potential real differences between the reference genome and the sequenced genome (both point mutations and indels). Likewise, alignment algorithms need to produce well-calibrated alignment/mapping quality values. Mapping algorithms can help to find corresponding sequences in the reference genome by allowing some mismatches permitting variation detection (Reinert et al., 2015). The optimal choice of the reasonable number of mismatches may differ between different organisms and regions. In general, mapping is more challenging for regions with higher levels of diversity between the reference and the sequenced genome. This difficulty can be improved utilizing longer reads and paired end reads (Nielsen et al., 2011). The alignment algorithms for short-read NGS data are mostly based on either ‘hashing’ or

Burrows–Wheeler transform (BWT) (Burrows and Wheeler 1994) algorithm. BWT-based aligners (e.g Bowtie, SOAP2 and BWA) are specifically designed for efficient data compression. They are also fast, memory-efficient, particularly useful for aligning repetitive reads; however, they tend to be less sensitive than hash-based algorithms. Hashed-based algorithms (e.g Stampy and mrsFast-Ultra) accelerate the alignment step. The use of hashing allows quick access to the information on the location of subsequence in the reference sequence (Altman et al., 2012).

Once the short reads are mapped to the reference genome, the alignment post-processing or BAM file refinement can be performed. All the non-unique alignments (e.g. reads with more than one optimal alignment) needs to be removed. Insertions/deletions (indels) do not present in the reference genome can cause small misalignments. Differences in resolving these indels may cause artificial SNPs in the downstream analysis. Thus, a local realignment can be performed for these regions. PCR duplicates as being introduced during library construction are also removed. Base quality score recalibration will also correct base scores, in turn leading to more accurate SNP calling. Therefore, higher quality and more accurate results can be achieved (Altman et al., 2012).

SNP and genotype calling are the processes of converting base calls and quality scores into a set of genotypes for each individual. Determining in which positions there are polymorphisms, or in which positions at least one of the bases differs from a reference sequence (also called variant calling) is the aim of the SNP calling. The process of determining the genotype for each individual is called genotype calling which is typically only done for positions in which a SNP or a 'variant' has already been called. Thus, this step helps to identify and qualify the differences between the data and the reference genome with an estimate of variant frequency and some measure of confidence (Nielsen et al., 2011).

Filtering is an essential step to reduce the possible number of false-positive SNP calls. This step can lower the error rate in detecting genomic variants by short-read sequencing. Typically applied filters check for deviations from the Hardy–Weinberg equilibrium (HWE), minimum and maximum read depth, adjacency to indels, strand bias, etc. Filtering may also eliminate real SNPs, however, minimizing SNP calling artifacts is also essential. Finally, the meaningful variants are annotated. It is substantial challenge

and effort to perform the post-processing and interpreting the generated SNP data, and this task should not be underestimated (Altman et al., 2012).

As seen in Figure 1.10, the most common file formats used in this type of workflow are the FASTQ, SAM, BAM and VCF files. A simplified workflow and the file formats are seen in image (Figure 1.11). A FASTQ file stores both raw sequences and their quality scores (Phred quality score, Q) with ASCII characters in a single file. FASTQ files have 4 parts, the sequence identifier, the raw sequence letters, a "+" character, optionally followed by the same sequence identifier, and the quality scores. As a typical step, a sequence alignment tool (e.g. BWA) use FASTQ files as input and align a large set of short reads to a reference and create SAM (Sequence Alignment Map) file as output. A SAM file is text-based format which represents biological sequences that are aligned to a reference sequence. SAM files contain two parts, the first part is the header that starts with an "@" character, and the second part is the alignment section that has 11 mandatory fields for each row such as read name, bitwise flag (codes information about the read e.g. mapped/unmapped, paired/not paired, mapped to forward/reverse strand etc.), reference sequence name, starting position of the mapped reads on the reference sequence, mapping quality, CIGAR string (a short description of the alignment), reference name for the mate (for paired data), position of the mate (for paired data), distance between paired reads (for paired data), nucleotide sequence of the read, per base quality of the read, and maybe optional fields. A BAM (Binary Alignment Map) file is the binary file that SAM file is stored into typically by using SAMtools and compressed with the BGZF (Blocked GNU Zip Format) tool (Hosseini et al., 2016). A VCF Variant Call Format (VCF) file is generated by comparing the BAM file to a reference sequence. With the VCF file, variant analysis can be performed because VCF files are used to report sequence variations such as SNPs and indels with rich annotations. VCF files begin with a number of meta-information lines that starts with two hash ('##') characters. Then, a single header line begins with a single hash ('#') character and data lines where each describe with each data line describing a genetic variant at a particular position relative to the reference genome (Ahmed et al., 2017).

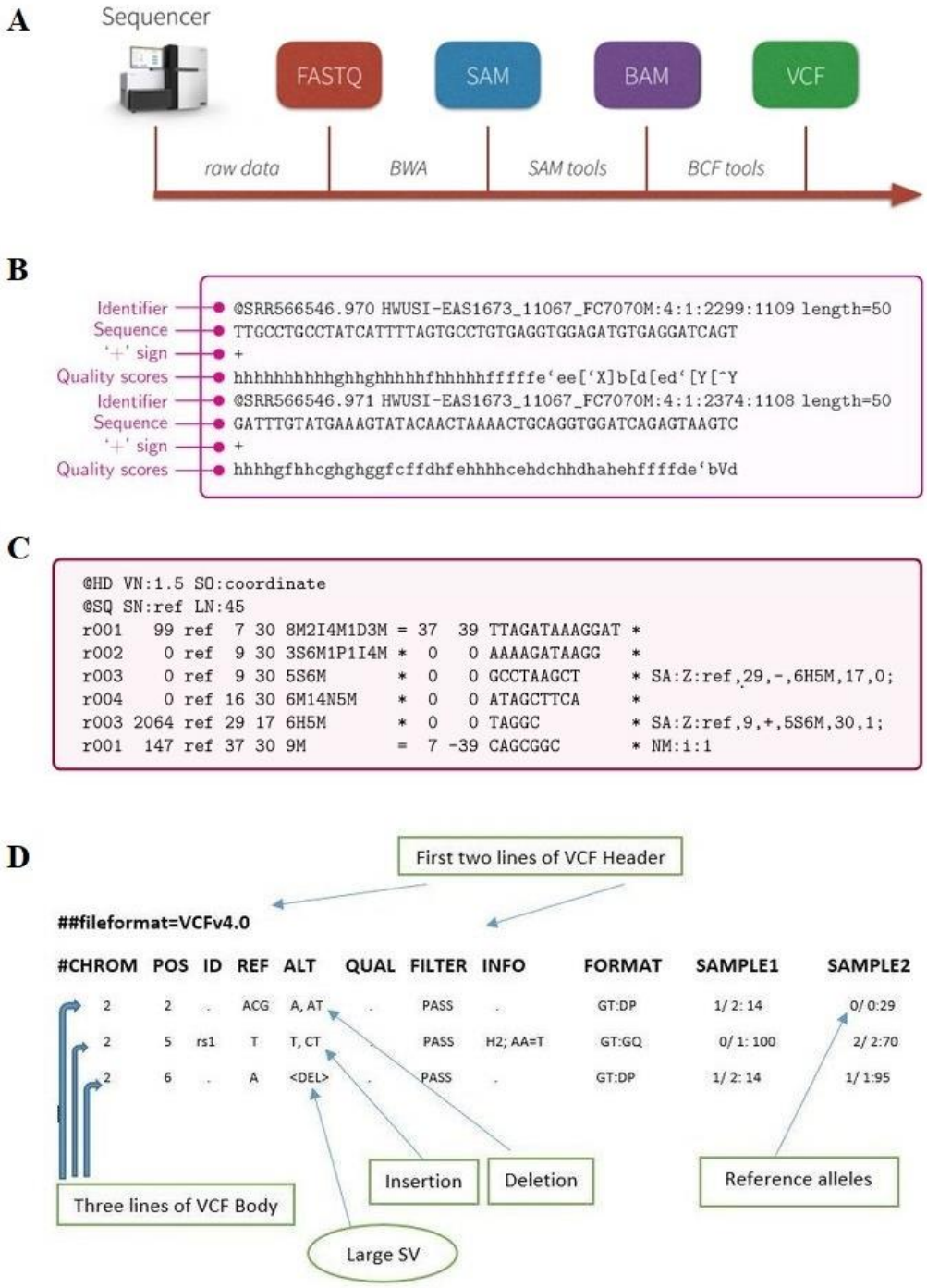


Figure 1. 11 A Simplified workflow to call variants and common file formats. A: A simplified workflow with common file formats and computational tools. B: A sample of the FASTQ file. C: A sample of the SAM file. D: A sample VCF file. Obtained and reproduced from Hosseini et al., 2016; Ahmed et al., 2017.

1.5. Sequence variation of copy number variable regions by short-read sequencing data

The copies of genes or regions can accumulate different variants and these variants can cause different functional consequences such as disease. It is even claimed that recent duplicated sequences are highly similar to each other, they can be complicated to do genome assembly and investigate the sequence variation between the variants as it is explained that SRGAP2 gene clusters are misassembled in the reference human genome in a study (Tyler-Smith and Xue, 2012).

There have been studies about sequence variation in the duplicated DNA regions of the human genome. In a study, short-read paired-end from 1000 Genomes project for 159 human genomes were used by using read-depth approach and demonstrated the estimation of copy number accurately for duplications about 1.9 kb, ranging from 0 to 48 copies. This study also identified 4.1 million “singly unique nucleotide” informative positions to investigate the specific copies as well as genotype the copies for highly duplicated gene families (Sudmant et al., 2010). In another study, the high-coverage phase 3 exome sequences of the 1000 Genomes project were used to study diploid copy number of the beta-defensin genomic region, not only detecting the copy numbers but also calling sequence variants for 1285 samples from 26 populations from all around world by using RD approach (Forni et al., 2015). The sequence variation was investigated by applying haplotype-based variant detector tools which can call sequence variants from samples of different copies. The study concluded 436 sequence variants between the copies as well as the individuals. The 81 % of identified sequence variants are reported as new compared to dbSNP database (Forni et al., 2015).

1.6 Duplicated regions used in this study

Three different regions were choses as initial analyses. The *RHCE*/*RHD* repeat region, the beta-defensin repeat region and the low-affinity FC gamma receptor repeat region. These repeat regions have high sequence similarity (>97%), all containing copy number variable functionally important genes. The breakpoints for the SDs and copy number variability of these regions were previously studied.

1.6.1 Rhesus blood group system genes

The Rhesus factor is clinically the most important protein-based blood group system. It is the largest of all 29 blood group systems with 49 described antigens (Flegel, 2007). There is only one RH gene in most mammals which corresponds to the human RHCE gene. The *RHD* and *RHCE* genes are result of a duplication event about 5 million to 12 million years ago during mammalian evolution. RHD arose by a duplication of *RHCE*. An *RHD* deletion happened during the evolution of hominids, causing complete lack of the RHD gene. Two *RHD* and *RHCE* genes encode the erythrocyte Rh proteins, RhD and RhCE. While RHD carries the D antigen, RhCE carries CE antigens in various combinations (ce, Ce, cE, or CE) (Avent and Reid, 2000). The functional variations of these antigens are caused by Indels, single nucleotide polymorphisms, or gene conversion events (Flegel, 2011; Perry et al, 2012). The deletion of RHD gene on both chromosomes results in D-negative blood group phenotype (Wagner and Flegel, 2000). The presence of either one or two copies of this gene results in D-positive phenotype. Two *RHD* and *RHC* genes are in close proximity on chromosome 1 in opposite orientations. Each gene is ~ 60kb and two genes are separated by ~30kb. While the genes each have ten exons, are 97% identical to each other, RhD and RhCE proteins differ by 32-35 of 416 amino acids.

The location of segmentally duplications in the region of *RHD/RHCE* and NAHR-associated RHD deletion and duplication breakpoints with copy number variations are shown in Figure 1.12. NAHR mechanism frequently results in deletion (and more rarely, duplication) of *RHD*.

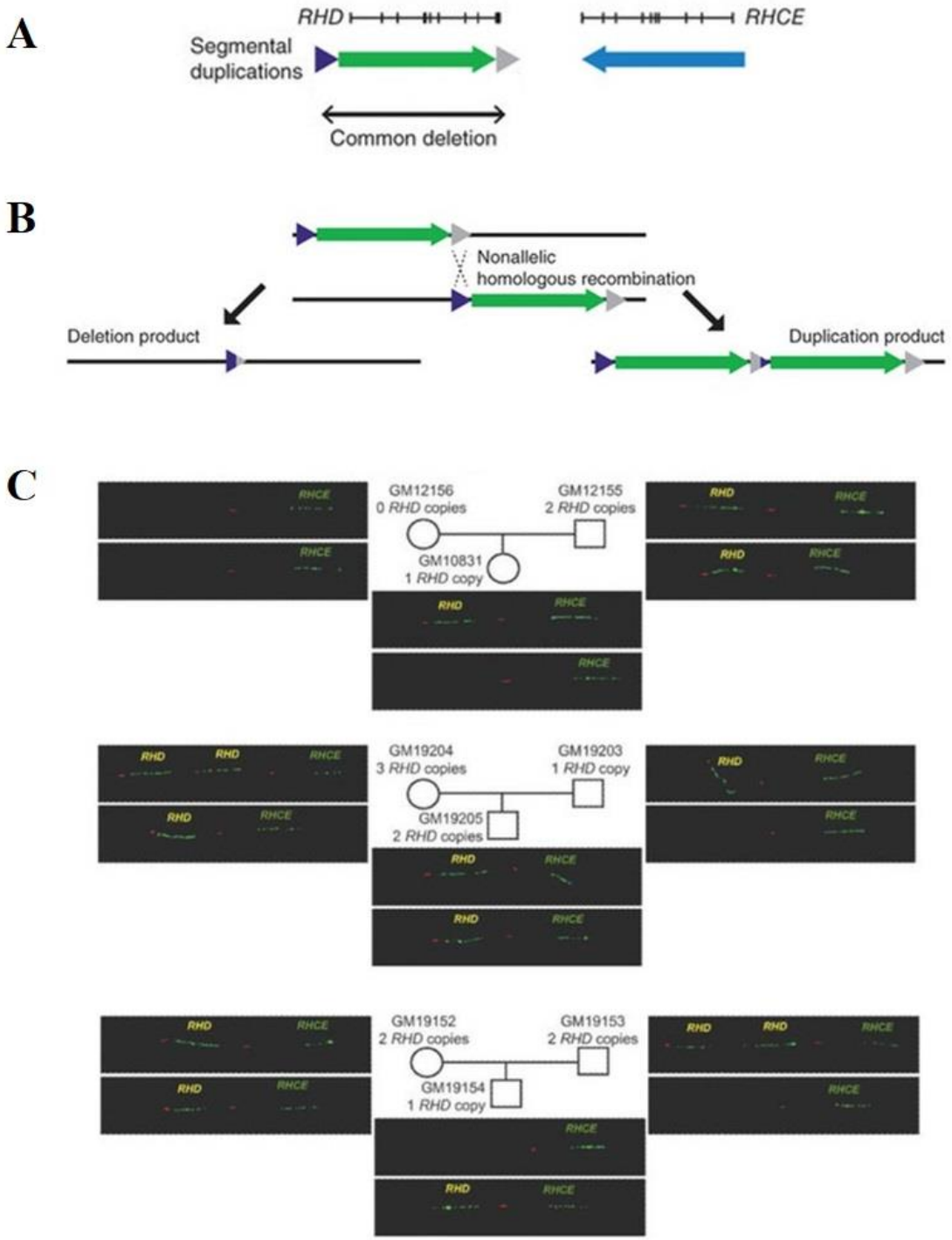


Figure 1. 12 NAHR-associated *RHD* deletion and duplication breakpoints with copy number variations. A: Segmentally duplicated regions of RHD and RHCE are shown as green and blue arrowed lines B: Nonallelic homologous recombination (NAHR)-associated *RHD* deletion and duplication breakpoints. NAHR between the flanking sequences frequently results in deletion (and more rarely, duplication) of *RHD*. Adapted from Nuttle et al., 2013. C: Fiber-FISH validation of copy number of RHD gene for three HapMap trio samples with representative images from the two alleles for each individual. Individuals could have 0 to 3 copies of RHD. For example, in the first scenario the new-born, in which the mother is D-negative, and the father has two copies of RHD, such that all of their offspring would be D-positive. Adapted from Perry et al., 2012.

1.6.2 Beta-Defensins

Defensins are significant components of the vertebrate innate immune system against a range of pathogenic microorganisms as having a role in adaptive immunity. The defensins can be divided into the alpha, beta and theta-defensin subfamilies on the basis of disulphide-bridge formation between the six conserved cysteine residues that define defensins. Beta-defensins are small cationic peptides providing a wide range of antimicrobial activity against bacteria, fungi, mycobacteria, and several enveloped viruses and differ in their expression patterns.

It has been claimed that a primordial beta-defensin is the common ancestor of all vertebrate defensins and this gene family expanded throughout vertebrate evolution even though there is still uncertainty between the evolutionary relationship between vertebrate and non-vertebrate defensins. Alpha-defensin genes and different beta-defensin genes are present on adjacent loci on chromosome 8p22–p23 in humans. The hypothesis of multiple rounds of duplication and divergence under positive selection from a common ancestral gene is consistent with the organization of this gene cluster and diversified paralogous. The current β-defensins might have evolved before mammals diverged from birds generating α-defensins in rodents, lagomorphs, and primates after their divergence from other mammals (Machado and Ottolini, 2015)

The 8p23.1 β-defensin locus has been shown to evolved by duplication and subsequent divergence, to create a diverse cluster of paralogous genes. This cluster is composed of nine genes. The four novel genes (*DEFB105, DEFB106, DEFB107, DEFB108*), a novel pseudogene (*DEFB109p*), and three known genes (*DEFB4, DEFB103 and DEFB104*) (Pazgier et al., 2006). A particular region has a 322kb segmental duplication (Figure 1.13). The copy number of this region changes from 2 to 12 copies in different populations (Bakar et al, 2009; Ottolini et al, 2014).

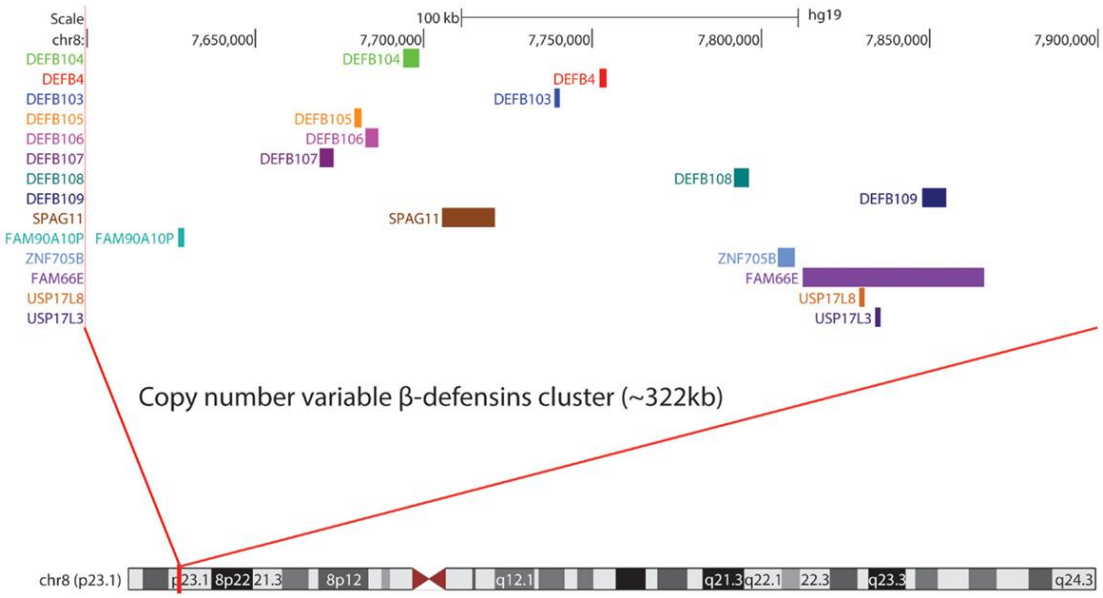


Figure 1. 13 Genome assembly of β-defensin repeat unit at 8p23.1. The human β-defensin CNV region includes several genes. The red arrows indicate the copy number variable repeat (322kb). Obtained from Machado and Ottolini, 2015.

1.6.3 FC Gamma Receptors (FCGRs)

In mammals, four different classes of Fc gamma receptors have been identified: FcRI (high affinity receptors), FcgRII, FcgRIII and FcgRIV (low-affinity receptors) (Nimmerjahn et al., 2015). A common ancestor of rodents and primates has been hypothesized to carried three low-affinity Fc receptor derived from a single ancestral gene (Qiu et al., 1990; Fanciulli et al, 2008). Humans show the most complex system for the low-affinity FCGR genes and the gene duplications were probably accompanied by intergenic recombination. The existence of multiple paralogues for both FCGR3 and FCGR2 in human and chimpanzee but not in macaque suggest that *FCGR3B* and *FCGR2C* arose after the separation from Macaque lineage about 25mya (Fanciulli et al, 2008). However, the high homology between the genes makes the clarifications difficult and cause inaccurate interpretations. The segmental duplication forming the organisation of the FCGR2/3 genes is not present in all primates. It has been suggested that both chimpanzees and gibbons have duplicated FCGR3 (Machado et al., 2012). A recent study searched the genomes of non-human primates for low/medium affinity FCGR orthologs and the *FCGR2C* gene was amplified by using a long PCR approach for human, chimpanzee, gorilla but not for orangutan, or macaque. It was concluded that NAHR

mechanism gave rise to a new block encompassing *FCGR2C* and *FCGR3B* in humans, chimpanzees, and gorillas (Ref).

The genes encoding the low-affinity FcgRs (*FCGR2A, FCGR2B, FCGR2C, FCGR3A, and FCGR3B*) are located in an 82.5-kb segmental tandem duplication on chromosome 1q23.3 and characterized by extensive copy number and sequence variation and high sequence similarity (Nimmerjahn and Ravetch, 2008). The receptors encoded by *FCGR3A* and *FCGR3B* are functionally distinct and shares ~98 % sequence identity. *FCGR2A* and *FCGR2B* are both end of the segmental duplications, with *FCGR2C* which is a fusion gene of partial 5' end of *FCGR2B* and partial the 3' end of *FCGR2A* as a result of unequal crossover, spanning the two segmental duplications (Hollox and Hoh, 2014).

The mechanism generating FCGR deletions and duplications has been shown to be mediated by NAHR (Figure 1.14). CNVs in this region mostly can either result in FCGR3A and *FCGR2C* duplication or deletion together, or either *FCGR3B* or *FCGR2C* together. The *FCGR2A* and *FCGR2B* are not involved in the duplications or deletions as they fall outside the CNV region. The location of the breakpoints determines whether a fusion gene is created and which of the *FCGR3A* or *FCGR3B* gene is deleted (Rahbari et al, 2016). Four combinations of copy number variable regions (CNRs) of FcgR genes have been shown to occur in duplication/deletion.

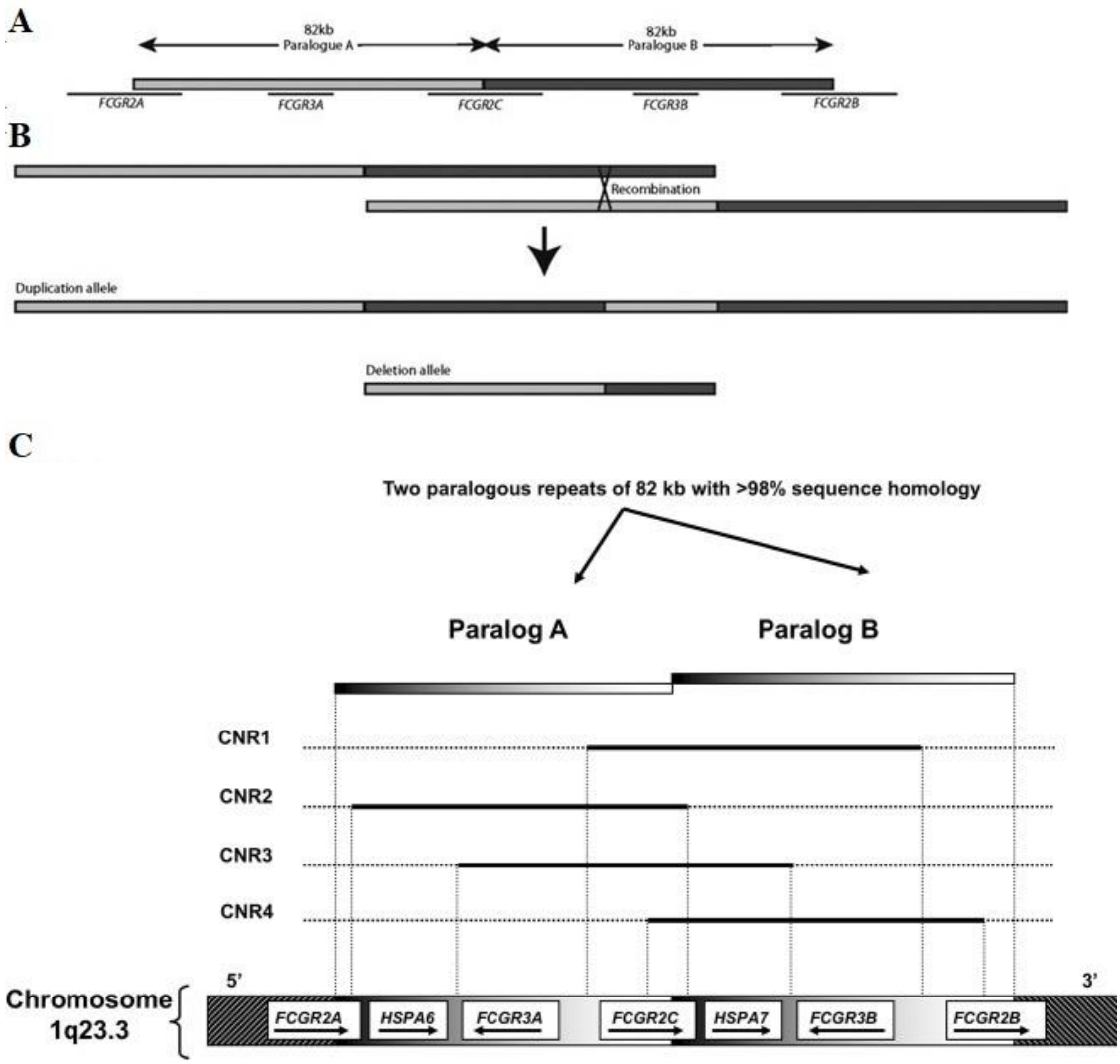


Figure 1. 14 Genetic structure of low-affinity Fc gamma receptor region. A: Two 82 kb paralogues repeats with 98.5% identity at chromosome 1p23.3 carry the low-affinity Fcg receptor genes *FCGR2A* , *FCGR3A* , *FCGR2C* , *FCGR3B*, and *FCGR2B*. B: NAHR between paralogues can generate deletion and duplication alleles where the duplicated allele contains a chimeric copy. Obtained from Rahbari et al., 2016. C: Overview of copy number variation at the locus. The most commonly observed is CNR1 which includes the *FCGR2C* gene and the *FCGR3B* gene. CNR2 involves the 3' untranslated region (UTR) of the *FCGR2A* gene and partially *FCGR2C* gene. CNR3 is involves the *FCGR3A*, *FCGR2C* and *HSAP7*. There is also a rare deletion named CNR4 which involves the *FCGR2C* exon 3 to *FCGR2B* exon 3. Obtained from Nagelkerke et al., 2019.

Aims of this study

- Design a computational pipeline to resolve the ambiguity in the read mapping for the duplicated DNA regions in the human genome by using a reduced reference. Perform mapping analyses of short sequence reads against reduced references by using different computational mapping programs and compare the performance of these programs. (*According to the results of computational pipeline, *RHD/RHCE* and Beta-Defensins repeat regions will not be investigated further).
- Confirm the copy number of previously studied samples and call the copy numbers of the samples used in this study for the low-affinity FCRs locus by using robust PRT assay.
- Design a gold standard long range PCR of *FCGR2B* and use Ion Torrent semiconductor sequencing platform for sequencing of the samples as a wet lab validation.
- Investigate the sequence variation of the PCR samples and construct haplotypes. Investigate sequence variation between the copies of the duplicated regions based on predicted copy number for the low-affinity FcgR locus for the mapped samples and construct haplotypes. Compare the results of computational analyses and wet lab validation results for the sequence variation and haplotype construction.
- Create a good indication of variant list for *FCGR2B* gene and inspect the consequences of these variants and investigate any variants which variants might explain any GWAS hits.
- Develop a PCR assay for genotyping the gene conversion of *FCGR2B* and to Investigate the gene expression profile in *FCGR2B*
- Inspect for any signature of selection on *FCGR2B* concerning the importance of rs1050501.

CHAPTER 2 Materials and methods

2.1 Resources

2.1.1 Sequence data used: Platinum pedigree and high coverage data

The 1000 Genomes project aimed to create a public catalogue of human variation and genotype data. The final data set contains data for 2,504 individuals from 26 populations. Low coverage and exome sequence data are present for all the individuals. 24 unrelated individuals were sequenced to high coverage for variant validation purposes. In this study, the publicly available 24 samples (sequenced 20-60x coverage, 2x250bp paired-end reads sequenced by Illumina technology in fastq format) of 1000 Genomes project were downloaded from the European Bioinformatics Institute (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/) Additionally, the CEPH (Centre d'Étude du Polymorphisme Humain)/UTAH pedigree 1463 samples (sequenced 50x coverage) were used to generate comprehensive pedigree-validated variants. The dataset comprises 11 children from two parents who are the children of the four founders of the pedigree (Figure 2.1). The DNA samples of the platinum data and the 24 unrelated individuals were also used in this study as indicated in Table 2.1.

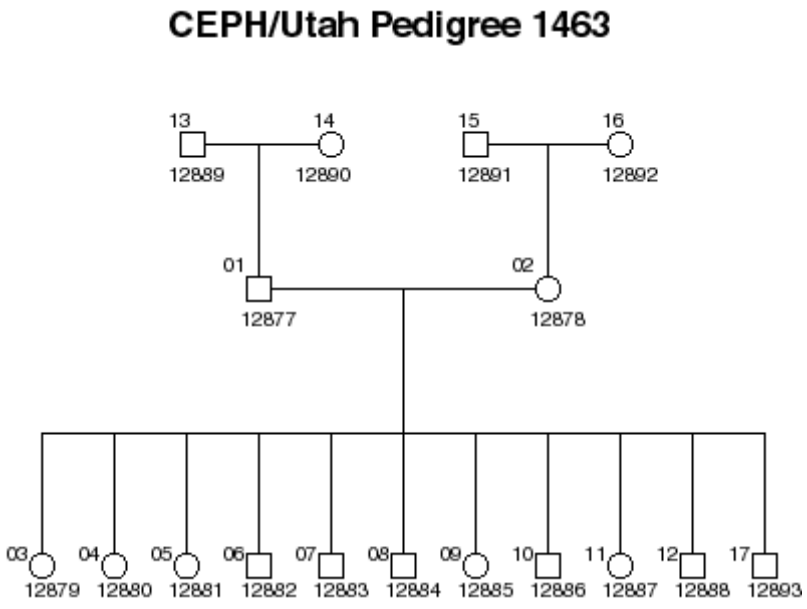


Figure 2. 1 The pedigree of the family sequenced in the Illumina Platinum Genomes Project.

Table 2. 1 The list of the high coverage data samples used in this study.

Sample ID	Sub Population Code	Population Code	Population Description	DNA sample
HG00419	EAS	CHS	Southern Han Chinese	+
HG00759	EAS	CDX	Chinese Dai in Xishuananna, China	+
HG01595	EAS	KHV	Kinh in Ho Chi Minh City, Vietnam	+
NA18525	EAS	CHB	Han Chinese in Bejing, China	+
NA18939	EAS	JPT	Japanese in Tokyo, Japan	+
HG00096	EUR	GBR	British in England and Scotland	+
HG00268	EUR	FIN	Finnish in Finland	+
HG01500	EUR	IBS	Iberian Population in Spain	+
NA20502	EUR	TSI	Toscani in Italia	+
HG02922	AFR	ESN	Esan in Nigeria	+
HG01879	AFR	ACB	African Caribbeans in Barbados	+
HG02568	AFR	GWD	Gambian in Western Divisions in the Gambia	+
HG03052	AFR	MSL	Mende in Sierra Leone	+
NA19017	AFR	LWK	Luhya in Webuye, Kenya	+
NA19625	AFR	ASW	Americans of African Ancestry in SW USA	+
HG01051	AMR	PUR	Puerto Ricans from Puerto Rico	+
HG01112	AMR	CLM	Colombians from Medellin, Colombia	+
HG01565	AMR	PEL	Peruvians from Lima, Peru	+
NA19648	AMR	MXL	Mexican Ancestry from Los Angeles USA	+
HG01583	SAS	PJL	Punjabi from Lahore, Pakistan	+
HG03006	SAS	BEB	Bengali from Bangladesh	+
HG03642	SAS	STU	Sri Lankan Tamil from the UK	+
HG03742	SAS	ITU	Indian Telugu from the UK	+
NA20845	SAS	GIH	Gujarati Indian from Houston, Texas	+

2.1.2 DNA samples used

2.1.2.1 HapMap samples

The International HapMap project was a worldwide project to study the genetic diversity of four different populations with a total of 270 DNA samples from the African Yoruba from Nigeria (YRI), the Japanese from Tokyo (JPT), the Han Chinese from Beijing (CHB) and the European-descent collection from Utah (CEU) (The International HapMap Consortium, 2003). All HapMap DNA samples were incorporated into the 1000 Genomes Project and sequenced at low coverage (http://www.internationalgenome.org/). The blood samples were converted into cell lines by Epstein Barr Virus transformation of peripheral blood lymphocytes and the appropriate ethics committees were provided by the Coriell Institute (http://ccr.coriell.org). The seven lymphoblastoid cell lines NA18517, NA18507, NA18956, NA19240, NA18555, NA12156 and NA12878 (which was sequenced as a part of the platinum pedigree) from Coriell Cell Repositories were used to grow and to extract DNA/RNA by MSc student Poonam Thakkar as a part of her Masters' thesis. The DNA samples were used for the gene conversion assay. The RNA samples were used for the expression profile of two SNPs in *FCGR2B* locus. These cell lines correspond to the samples where fosmid libraries were created and sequenced (Kidd et al., 2008; Mueller et al, 2012). The fosmid sequences were combined with the other sequence data samples for the population genetics analyses of the *FCGR2B* locus.

2.1.2.2 ECACC Human Random Control (HRC) samples

The Human Random Control (HRC) DNA samples represents a control population of 480 European origin blood donors characterised by gender and age at venesection. (https://www.phe-culturecollections.org.uk/products/dna/hrcdna/hrcdna.jsp). The donors have given written, informed consent for their blood to be used for research purposes. The DNA samples were extracted from lymphoblastoid cell lines derived by Epstein Barr Virus (EBV) transformation of peripheral blood lymphocytes from single donor blood samples. The genomic DNA was provided as a solution at a standard concentration of 100ng/µl in 10mM Tris-HCl buffer (pH 8.0) with 1mM EDTA. The six control samples CO121, CO081, CO909, CO210, CO744 were used for the paralogue ratio test (PRT) assay.

2.1.3 Computational resources

All the computer analyses were performed on the ALICE 2 which is an HPC cluster. It is a research facility available to all staff and research students who have registered as members of an HPC project at the University of Leicester. The HPC system runs CentOS Linux that offers both a command line and graphical environment. Many applications were installed on the HPC system and commonly there is a module available for each application. The command "module" starts the application. All the applications used in this study are listed in Table 2.2.

Table 2. 2 Applications locally installed and installed on ALICE HPC cluster.

Application	Version	Module	Description
FastQC	0.11.5	fastqc/0.11.5	Quality control tool for high-throughput sequence data
Trimmomatic	0.35	trimmomatic/0.35	Illumina NGS read trimmer
BWA	0.7.12	bwa/0.7.12	Burrows-Wheeler Aligner Mapping tool
mrsFast-Ultra	3.4.0	mrsfast/3.4.0	Micro read substitution-only Fast Alignment Search Tool
GATK	3.6	gatk/3.6	Genome Analysis Toolkit
Picard	2.6.0	picard/2.6.0	Java tools for next generation sequencing data
SAMtools	1.8	samtools/1.8	SAM-format sequence alignment tools
Perl	5.24.0	perl/5.24.0	Perl programming language interpreter
FreeBayes	1.1.0	FreeBayes/1.1.0	Bayesian haplotype-based polymorphism discovery and genotyping
HapCompass	0.8.2	hapcompass/0.8.2	Creation of haplotype assemblies
BCFtools		bcftools/1.9	Reading/writing BCF2/VCF/gVCF files
VCFtools	0.1.14	vcftools/0.1.14	Tools for using Variant Call Format files
Tabix/bgzip	0.2.6	tabix/0.2.6	Generic indexer for TAB-delimited genome position files

2.2 Quality checking and filtering of the sequencing data

2.2.1 Quality check of sequencing reads

FastQC is a tool which provides several types of report such as basic statistics, per base quality check, sequence length distribution, or adapter content. It is mostly used to identify any problems that many have originated from the sequencer or library preparation process. The program can also take different formats of sequencing data as input (Andrews, 2010). The data from 24 unrelated individuals of 1000 Genomes project and 17 of the CEPH/UTAH 1463 pedigree samples were used to run on FastQC (v0.11.5) for the quality check.

2.2.2 Trimming of sequencing reads

Trimmomatic is quality filtering tool used for both paired-end and single-ended data (Bolger et al., 2014). It is mainly used to cut adaptors from the reads. It cuts low quality bases off from the start and end of the reads and performs a sliding window trimming. Finally, it drops all the sequences below a defined threshold or removes bases with a defined low-quality score (Figure 2.2). As a result, it produces trimmed/filtered paired forwards and reverses (both reads survived), and unpaired forwards and reverses (one of the pairs is survived but the other pair did not due to the quality and minimum read length).

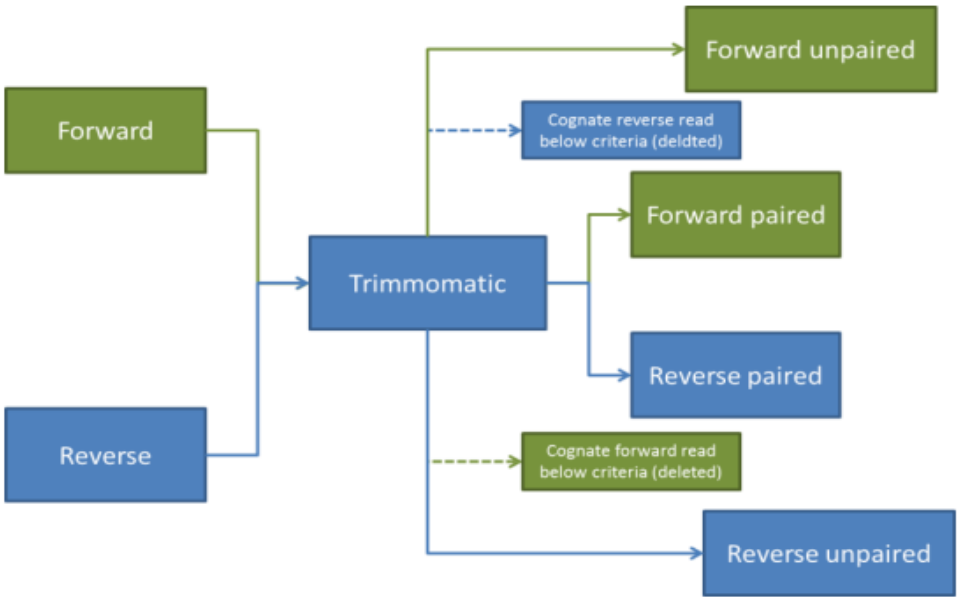


Figure 2. 2 "Flow of reads in Trimmomatic Paired End mode". Trimmomatic User Manual v0.32

The unrelated samples are 250 base-long short-read sequences from different population all around the world and were filtered by quality with trimmomatic (v0.35). The following parameters were used for filtering :LEADING: 3, reads have low quality below 3 from the start were removed; TRAILING: 3, reads have low quality below 3 from the end were removed; MINLEN: 250, reads below 250 bases long were dropped; SLIDINGWINDOW: 4:10, reads were scanned with a 4-base wide sliding window and cut if the average quality is below 10 (Phred Score 10). ILLUMINACLIP option is used to remove adapter. Adapters were already removed for the unrelated and pedigree samples so that this option was not used for quality filter as a parameter. The pedigree samples were 100 base-long short-read sequences in paired-end format and that had been already filtered by quality so that trimming was not applied.

2.3 Mapping of the sequencing reads against reference

Burrows Wheeler Aligner (BWA) (Li and Durbin, 2010) and mrsFAST-Ultra (Micro-read substitution-only Fast Alignment Search Tool), as a hash-based mapper (Hach et al, 2014) are the two mapping tools used in this study. MrsFAST-Ultra (v3.4.0) and BWA-MEM (v0.7.12) mappings were performed using a paired end approach on both trimmomatic quality filtered and unfiltered version of 24 unrelated samples and platinum data. A general workflow of BWA and mrsFAST-Ultra are shown in Figure 2.3.

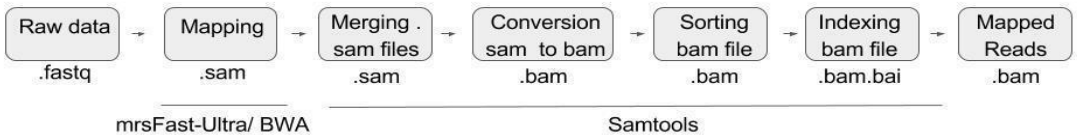


Figure 2. 3 A workflow of mapping process.

2.4 BAM Refinement of mapped reads

BAM refinement includes local realignment and PCR duplicate removal. Insertions/deletions that are not present in the reference genome can cause small misalignments. By doing local realignment for these regions, the number of mismatched bases can be reduced. PCR duplicates as introduced during library construction are also removed. Local realignment was done by using Genome Analysis toolkit (GATK) (v3.4-0) and PCR duplicates were removed by using SAMtools (v1.8) Picard (v2.1.0) was used to

mark duplicates and for the flag statistics. A general workflow of BAM Refinement is shown in Figure 2.4.

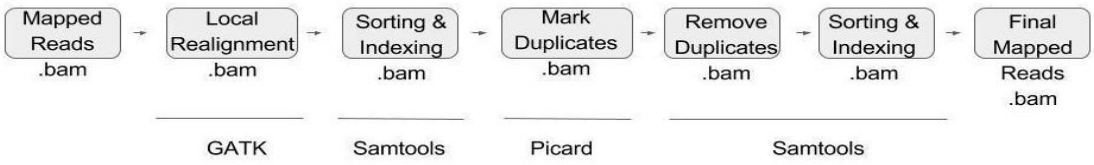


Figure 2. 4 A workflow of BAM refinement.

2.5 Analysis of mapping results

Upon the mapping results, a perl script (Appendix 9) scripted by Diego Forni (former postdoctoral researcher) is used to count how many sequences are mapped to a region from the BAM files for the mapping results of all samples. The script takes the final mapped reads (the final BAM file for a sample) with a BED file (showing the borders of the CNV or Non-CNV regions for each gene cluster) by calling the counting option SAMtools to find how many short sequences are mapped to a region. Then a ratio is generated by dividing the number of mapped reads of CNV region by the number of mapped reads of non-CNV regions for each gene cluster separately (will be named RHD, FCGRs and BDEF for the rest of the analyses). The ratios are used to create strip plots by using the samples for which we know the copy numbers known from the previous studies for *RHD*, FCGRs and Beta-defensins regions (Handsaker al., 2015; Forni et al., 2016).

2.6 Paralogue Ratio Test (PRT) for the low-affinity FCGR locus

2.6.1 Primer design and PCR

The primers listed in Table 2.3 were previously designed and used in another study to determine the copy umber of the low-affinity FcgR region (Niederer et al., 2010). The PCR components and PCR conditions are listed in Table 2.4 and Table 2.5 below, respectively. The forward PRT primers are labelled on their 5' end by fluorescent dyes; HEX (hexachloro-fluorescein) and FAM (Fluorescein Amidite). **T**he PCR for the desired regions was optimised in the Veriti® Thermal Cycler PCR machine.

Table 2. 3 Primer used for PRT Assay.

Primer Name	Primer sequence (5' to 3')	Product Size	5' Modification
PRT-2C3-Forward-H	CTT CAT GAA TTG CGC CTC AG	274*,279**	HEX
PRT-2C3-Forward-F	CTT CAT GAA TTG CGC CTC AG	274*, 279**	FAM
PRT-2A3-Reverse	GCT AGA GGC CAG AAG TTC GAG		
*FCGR2C, **FCGR2A			

Table 2. 4 PCR Components for PRT Assay.

PCR Components	Final Concentrations
10x low dNTP Mix*	1X
Forward-FAM (10µM) primer	0.05µM
Forward-HEX (10µM) primer	0.05µM
Reverse primer	0.1µM
Taq DNA Polymerase (5U/µl)	0.025U/l
DNAase free H2O	As needed
gDNA(~10ng/ul)	~1ng/ul
*Appendix 1	

Table 2. 5 PCR Conditions for PRT Assay.

Step	Process	Temperature °C	Time
1	Initial denaturation	96	2 minutes
2	Denaturation	96	30 seconds
3	Annealing	56	30 seconds
4	Extension	70	30 seconds
5	Repeat step 2 to 4 (for 22 cycles)		
6	Final Extension	70	5 minutes
7	Hold	4	For ever

2.6.2 Capillary electrophoresis

In order to distinguish size or nucleotide differences between the test and the reference to infer the copy number calculation, the PCR amplicons have resolved using capillary electrophoresis. For each sample, 1µLPCR product was added to the mixture of 10µl HiDi Formamide containing 0.01% MapMarker X-Rhodamine (Rox1000XL) (Bioventure, Murfreesboro, TN). The mix was denatured for 3 minutes at 96°C, followed by ice shock. Then, the PCR products were run on ABI 3130xl Genetic Analyser to be distinguished based on their size differences. Fragment analysis was carried out by electrophoresis on an ABI3100 36 cm capillary using POP-4 polymer with an injection time of 30 s. All the peak sizes are obtained and analysed on Gene Mapper® (Applied Biosystems).

2.6.3 Data normalization

The raw paralogue ratios (test/reference) were calculated from each dye by using peak heights. The data were then used to create a calibration curve by performing a linear regression of the control sample ratios against expected integer copy number, shown in Figure 2.5. The DNA samples of five unrelated individuals from the HRC Panel were used as control samples and optimize the PRT assay of the low-affinity FCGR locus. Controls are CO121 (3 copy), CO081 (4 copy), CO909 (4 copy), CO210 (5 copy), CO744 (6 copy). they have been used throughout as PRT controls and initial copy number was inferred from aCGH data results (Redon et al 2006). The PRT assays were applied to the 24 of 1000 Genomes project and 17 Platinum data (CEPH/UTAH 1463) DNA samples. The calibration curve was created using the known copy number positive controls samples. Data normalisation was performed in Excel by linear regression of the control sample ratios against their previously estimated integer copy numbers as shown in Figure 2.5 below.

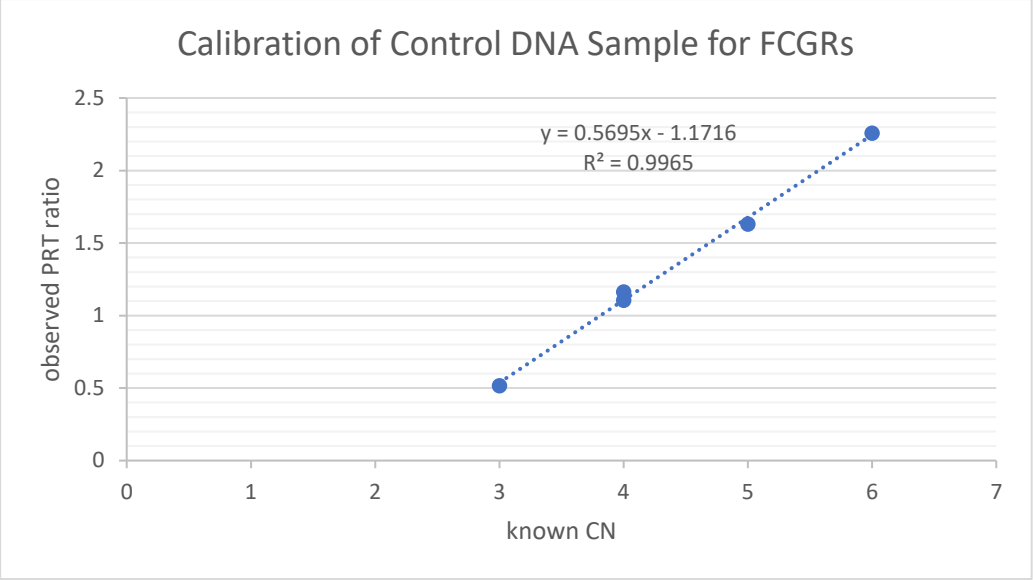


Figure 2. 5 The calibration curve of controls for low-affinity FCGR locus. X-axis represents the predicted relative CN ratio for this PRT assay and y-axis represent the known CN for the same samples from the previous studies.

2.6.4 Copy number estimation

CNVtools is a software programme that can be used to perform robust case-control and quantitative trait association testing of CNVs (Barnes et al., 2008). A histogram and a fitted text file were created for each PRT assay using CNVtools in the R program (Appendix 10).

2.7 Amplicon Sequencing with the Ion Torrent (PGM) platform.

2.7.1 Primer design and PCR

The desired sequence of *FCGR2B* region was retrieved from the GRChr37/hg19 human reference genome assembly on UCSC Genome Browser Home (http://genome.ucsc.edu/). The primers were designed by using the web-based tool Primer3 (http://primer3.ut.ee/) and then checked using the *In-silico* PCR tool of the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgPcr) to ensure that primer pairs were unique to the sequence to be amplified and no common sequence variants were observed. These primers were used in PCR (Table 2.6) and then the PCR for the desired region was optimised in the Veriti® Thermal Cycler PCR machine. The PCR components and PCR conditions are listed in Table 2.7 and Table 2.8 below, respectively.

PCR products were resolved on 0.8 %(w/v) agarose gel (0.5X TBE) with an ethidium bromide concentration of 0.5µg/ml. Theof DNA was mixed with 1µLof 6X loading dye and ran alongside 5µLof DNA ladder (HyperLadder™ I, Bioline) at 100V for several hours. The gel was then observed under the UV light transilluminator.

Table 2. 6 Primers used for the long-range PCR of *FCGR2B.*

Primer sequence (5' to 3')		Product Size
Forward	TGG TAA TGA GGA TGA TGA CAG A	16834bp
Reverse	TGT TGC ACG ATG TTA CTG CG	

Table 2. 7 PCR Components for the long-range PCR of *FCGR2B.*

PCR Components	Final Concentrations
11.1x PCR Buffer Mix*	1X
Forward primer (10µM)	0.2µM
Reverse primer (10µM)	0.2µM
Taq DNA polymerase (5U/µl)	0.025U/l
Pfu DNA polymerase (2.5U/ul)	0.0033U/l
DNAase free H2O	As needed
gDNA	~10ng/ul
Appendix 2 (Kauppi et al. 2009)	

Table 2. 8 PCR Conditions for the long-range PCR of *FCGR2B.*

Step	Process	Temperature °C	Time
1	Initial denaturation	94	1 minute
2	Denaturation	94	15 seconds
3	Annealing	68	10 minutes
4	Repeat step 2 to 3 (for 22 cycles)		
5	Denaturation	94	15 second
6	Annealing	68	10 minutes *A
7	Repeat step 5 to 6 (for 12 cycles)		
8	Extension	72	16 minutes
9	Hold	4	For ever
***A: 15 second elongation is added for each successive cycle**			

2.7.2 Ion Torrent Sequencing of *FCGR2B* region

All amplicons were sequenced using Ion Torrent Sequencing technology. This was performed with the Ion Xpress plus gDNA Fragment Library Kit according to User Bulletin. The amplicons were cleaned with Agencourt® AMPure XP beads and used to make individual-specific libraries using the Ion Xpress™ Library kit and barcodes according to the manufacturer's instructions. Libraries were size selected on 1.8% (w/v) 0.5X TBE agarose gel and purified by using a Zymoclean™ Gel DNA Recovery Kit. The purified samples were quantified using an Agilent 2100 Bioanalyzer with Agilent DNA 1000 chip and all the samples were pooled equimolar to 100pmol/ul. Sequencing templates were prepared using the Ion PGM™ Hi-Q View OT2 Kit and sequencing was performed according to manufacturer's instructions in two runs on an Ion Torrent™ PGM™ 316 Sequencer, using the Ion PGM™ HI-Q Sequencing Kit and Ion 316Chip Kit v2 BC (Thermo Fisher Scientific). Reads were mapped to the human reference sequence (GRChr37/hg19) of *FCGR2B* using the Torrent Suite™ Software 5.0.2.

2.8 Haplotype estimation

2.8.1 BEAGLE for diploid data

Haplotype estimation (phasing) was inferred using BEAGLE (v4.1) (Browning and Browning, 2007) for the PCR samples of the 24 unrelated samples and the platinum data samples. All the unphased samples were combined in a multi VCF file. A BED file was created showing the ancestors of each individual for this data set. Then the multi VCF file was phased together by BEAGLE.

2.8.2 HapCompass for polyploid data

Haplotypes can be assigned to separate the different copies of the segmentally duplicated regions. In this study, haplotype estimation was inferred using HapCompass (Aguiar and Istrail, 2013) for the 24 unrelated samples and the 17 pedigree samples mapped against reduced references. The unphased VCF files were obtained from FreeBayes calls on the BAM files.

2.9 Prediction of consequences of the variants in *FCGR2B*

A single VCF file was created with all the variants found for *FCGR2B* region from the FreeBayes call on diploid PCR samples. Then it was run on the Variant effect predictor (http://grch37.ensembl.org/Homo_sapiens/Tools/VEP) (VEP) to obtain the location of the novel/existing variants, identifiers for the variants, the types of the consequence of the variants, the allele frequency (AF) data for existing variants from several major genotyping projects, the 1000 Genomes Project, the NHLBI-ESP (NHLBI GO Exome Sequencing Project) and gnomAD (Genome Aggregation Database), and to search for any clinical significance found previously.

2.10 Description of the variants in *FCGR2B*

The variant descriptions were obtained using the VariantValidator (https://variantvalidator.org/) which is a web-based tool for the validation of human gene sequence variant descriptions to confirm that they comply with the HGVS variant description nomenclature (Freeman et al., 2018). On the website, the VCF to HGVS (powered by VariantValidator_vv0.2.2) option was selected. A single VCF file with all the novel and existing variants found for *FCGR2B* region was uploaded. The genome GRCh37 was chosen. After submission, results were returned by email once processing was complete.

2.11 Sequence alignment of SNP haplotypes

After performing haplotype estimation by BEAGLE, total of 56 chromosome sequences were identified and combined with another 11 chromosome sequences from the fosmid sequences of six samples (Rahbari et al., 2016). For the fosmid sequences, there was only one chromosome sequence available for some samples and the other chromosome

sequence was not fully spanning the *FCG2B* region so that only one chromosome sequence was used (Appendix 17). Total number of 67 samples including *Pan troglodytes* (chimpanzee) (Assembly access code: GCF_002880755.1) and *Gorilla gorilla gorilla* (western lowland gorilla) (Assembly access code: GCA_900006654.3) were aligned in ClustalW, multiple alignment program, that is implemented in MEGA (v7) software. A final multiple sequence alignment fasta file was created.

2.12 Population genetics analysis

Sequence polymorphism analysis and neutrality tests were performed by using the fasta file of the estimated haplotypes in the software DnaSP (v6.2) (Rozas et al., 2017). The nucleotide diversity, haplotype diversity, the average number of nucleotide differences were used as estimates of the genetic diversity of the total dataset and within the sub populations.

2.13 Network analysis

NETWORK is a software program used to reconstruct phylogenetic networks and trees, infer ancestral types and potential types, evolutionary branchings and variants, and to estimate datings (NETWORK User Guide, 2012). Median joining (MJ) is one of most commonly used methods in this software package. It finds a minimum spanning tree first. Then, it adds new median vectors (i.e., hypothetical, or missing ancestors) to a single network. This method handles large datasets and works fast (Bandelt and Forster, 1997). The unrooted network was computed for the generated data set by using the parameter epsilon=0 (default) and epsilon=10. The epsilon states a weighted genetic distance to the known sequences in the data set, within which potential median vectors may be constructed. If epsilon is less than the greatest weighted genetic distance within the data set, then theoretically the MJ network will not contain all possible shortest trees. If epsilon is set equal to (or greater than) the greatest weighted genetic distance, the MJ algorithm is guaranteed to produce a full median network.

2.14 Gene conversion assay

2.14.1 Growing and storage of the lymphoblastoid cell lines

The seven lymphoblastoid cell lines were grown up in RPMI media with the addition of Fetal Bovine Serum (FBS), Penicillin and Streptomycin. After three to four days the cell cultures were subcultured with fresh medium depending on cell growth. The number of counted cells in one 4x4 squared well was multiplied by 10,000. Approximately $5x10^6$ cells were transferred to a conical tube and centrifuged at 200g for 5 minutes at 4°C. The supernatant was discarded, and the cells were suspended in 5ml of ice-cold phosphate buffered saline (PBS). The cells were then centrifuged at 200g for 5 minutes at 4°C. The supernatant was discarded, and the cells were immediately frozen on dry ice. The pelleted cells were stored at -80°C. Cells were also frozen down in liquid nitrogen for any future work. 10% DMSO was added to the normal growth media (RPMI with FBS, Penicillin and Streptomycin) and stored on ice. The cells were then transferred to a 15ml tube and pelleted at 700g for 5 minutes. The supernatant was discarded, and the pellet was suspended in the cold media containing the DMSO. The cells were the stored in a freezing tub at -80˚C for at least 24 hours and then transferred to liquid nitrogen after the 24 hours.

2.14.2 Total DNA and RNA extraction

The DNA and RNA were extracted using the Qiagen AllPrep DNA/RNA Mini Kit according to the User Bulletin. The AllPrep DNA/RNA Mini Kit purifies genomic DNA and total RNA simultaneously from a single sample. Lysate from homogenized cells or tissue is first passed through an AllPrep DNA spin column to isolate DNA, then through a RNeasy® spin column to isolate RNA.

2.14.3 Primer design and PCR for detecting the gene conversion

The primers were designed using the web-based tool Primer3 (http://primer3.ut.ee/) and then checked using the *In-silico* PCR tool of the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgPcr) to ensure that primer pairs were unique to the sequence to be amplified and no common sequence variants were observed. These primers were used in PCR (Table 2.9) and then the PCR for the desired region was optimised in the Veriti® Thermal Cycler PCR machine. The PCR components and PCR conditions are listed in Table 2.10 and Table 2.11 below, respectively.

Table 2. 9 Primer used for the gene conversion assay.

Primer Name	Primer sequence (5' to 3')	Type	Purpose
ForwardA	TATGAAAGATGAATTGGAACGG	LNA*	Detects the presence of the gene conversion on the left side
ReverseB-P	CAAACCATTAATTACAACCATAC	LNA*	Detects the presence of the gene conversion on the left side on one allele.
ReverseB-A	CAAACCATTAATTACAACCATAT	LNA*	Detects the absence of the gene conversion on the left side on one allele.
***LNA (locked nucleic acid)**			

Table 2. 10 PCR Components for the gene conversion assay.

PCR Components	Final Concentrations
11.1x PCR Buffer Mix	1X
Forward primer (10µM)	0.2µM
Reverse primer (10µM)	0.2µM
Taq DNA polymerase (5U/µl)	0.025U/l
Pfu DNA polymerase (2.5U/ul)	0.0033U/l
DNAase free H2O	As needed
gDNA	~10ng/ul

Table 2. 11 PCR Conditions for the gene conversion assay.

Step	Process	Temperature °C	Time
1	Initial denaturation	94	1 minute
2	Denaturation	94	15 seconds
3	Annealing	59	10 minutes
4	Repeat step 2 to 3 (for 22 cycles)		
5	Denaturation	94	15 second
6	Annealing	59	10 minutes *A
7	Repeat step 5 to 6 (for 12 cycles)		
8	Extension	72	10 minutes
9	Hold	4	For ever
***A: 15 second elongation is added for each successive cycle**			

2.14.4 cDNA synthesis by reverse transcription–PCR (RT-PCR) for RNA samples

Before the cDNA synthesis, the RNA samples were treated DNase I. 2µL of 10x DNase I buffer was added to 1g of the RNA samples followed by 2µL 1U/µL DNase I. The samples were then made up to 20µLusing DEPC water and incubated at 37°C for 30 minutes. 2µLof EDTA was added and then the sample was incubated at 70°C for 10 minutes. The RNA samples were then stored at -80°C.

30ng of RNA was transcribed into cDNA using Invitrogen's SuperScript III First – Strand Synthesis Super Mix kit as per the kit's instructions. Total RNA isolated from the cell lines was primed for first-strand synthesis by using the Oligo(dT)20 primers provided with the kit in a total volume of 20ul. The following components were mixed thoroughly on ice for per RNA samples and the protocol was followed as indicated in Table 2.12.

Table 2. 12 The protocol for RNA to cDNA reaction by RT-PCR.

Components	Amount
Up to 5 µg total RNA	n µL
Primer (50 µM oligo(dT)20)	1 µL
Annealing Buffer	1 µL
DEPC water	Up to 8 µL
The mixture was incubated at 65°C for 5 minutes, then immediately placed on ice for at least a minute. Briefly centrifuged and content was collected. Following components were added to the tube on ice.	
Components	**Amount**
2X First-Strand Reaction Mix	10 µL
SuperScript™ III/RNaseOUT™ Enzyme Mix	2 µL
The samples were then incubated at 25°C for 10 minutes followed by 50 minutes at 50°C. The reaction was terminated at 85°C for 5 minutes and cDNA was stored at -20°C for later use.	

2.15 SNP targeted PCR from cDNA samples

The samples of NA18517, NA19240, NA18555 and NA19129 were used to amplify the desired regions where the two SNPs (rs60519172 and rs844) are located. The desired sequence of *FCGR2B* region was retrieved from the hg38 human reference genome assembly on UCSC Genome Browser Home (http://genome.ucsc.edu/). The primers were designed by using the web-based tool Primer3 (http://primer3.ut.ee/) and then checked using the In-silico PCR tool of the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgPcr) to ensure that primer pairs were unique to the sequence to be amplified and no common sequence variants were observed. These primers were used in PCR (Table 2.13) and then the PCR for the desired region was optimised in the Veriti® Thermal Cycler PCR machine. The PCR components and PCR conditions are listed in Table 2.14 and Table 2.15 below, respectively.

Table 2. 13 Primer used for SNP targeted PCR assay.

Primer sequence (5' to 3')		Product Size
Forward	CTG ATG AGG CTG ACA AAG TTG G	529bp
Reverse	CAT AAG CAT TTC CCA AGT TGC	

Table 2. 14 PCR Components for SNP targeted PCR assay.

PCR Components	Final Concentrations
10X Buffer+MgCl2 (Bioline)	1X
Forward-(10µM) primer	0.4µM
Reverse(10µM) primer	0.4µM
Taq DNA Polymerase (5U/µl)	0.025U/l
DNAase free H2O	As needed
gDNA(~10ng/ul)	~1ng/ul

Table 2. 15 PCR Conditions for SNP targeted PCR assay.

Step	Process	Temperature °C	Time
1	Initial denaturation	96	1 minutes
2	Denaturation	96	30 seconds
3	Annealing	62	45 seconds
4	Extension	72	45 seconds
5	Repeat step 2 to 4 (for 35 cycles)		
6	Final Extension	72	5 minutes
7	Hold	4	For ever

PCR products run on the gel with HyperLadder™ II, Bioline and the gel was then observed under the UV light transilluminator.

2.15.1 Window Gel PCR product extraction

The PCR products were cleaned up using the window gel extraction method (Ma and DiFazio, 2008). This method consists of running the PCR products on a higher concentrated gel with a lower concentrated gel (0.1-0.2% (w/v)). A diagram is shown in Figure 2.6. The PCR products were tun on 0.8 %(w/v) agarose gel (0.5X TAE buffer, ethidium bromide concentration of 0.5µg/ml) with a 0.2% agarose gel. About 20µL of PCR products were mixed with 3.5µLof 6X loading dye and ran alongside 5µLof DNA ladders (HyperLadder™ I, Bioline). When the products were in the middle gel frame, the bands were cut out under the blue light and frozen overnight.

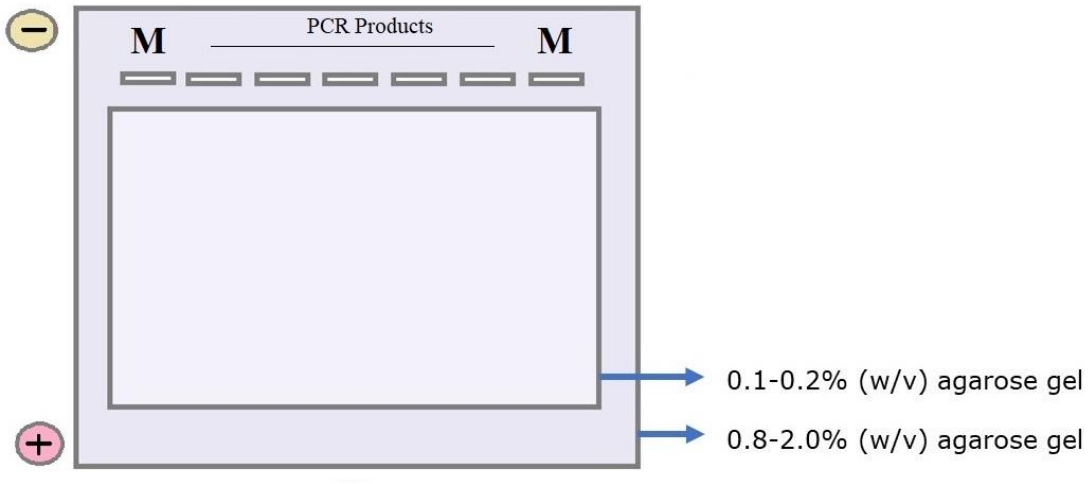


Figure 2. 6 The diagram of the agarose gel for the window gel extraction method. M: DNA ladder

The removed gel (frozen) was defrosted at room temperature for 10 minutes. It was then centrifuged for 30 minutes at 13200 rpm. The supernatants were extracted in order to separate the PCR product from the agarose gel. The PCR products were then quantified on agarose gel, and the sequencing protocol was carried out.

2.15.2 Sanger sequencing of PCR products

The automated sequencing reaction with BigDye ready protocol was used for the sequencing. Purified PCR product was prepared for a sequencing reaction (total volume 20 µl/reaction) using approximately 20-30 ng/kb of purified PCR products. The reaction components are listed in Table 2.16 The samples were run on the Veriti® Thermal Cycler PCR machine at 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.

Table 2. 16 Automated sequencing with BigDye ready reaction components.

Sequencing Components	Volume
BigDye Terminator ready reaction mix	1ul
5X BigDye Terminator buffer	2 ul
Forward primer (3.2µM)	1ul
Reverse primer (3.2µM)	1ul
PCR Product	~20-30ng/ul
DNAase free H2O	As needed

After the sequencing reaction, excess dye removal was performed by following protocol. The samples were then cleaned up by adding 10µl of H2O and 2µl of 2.2% SDS. The samples were heated to 98°C for 5 minutes and then cooled to 25°C for 10 minutes. 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. Finally, the ready reaction then was run on an Applied Biosystems 3700 Genetic Analyzer by the Protein and Nucleic Acid Laboratory of the University of Leicester (PNACL), and data returned via email. The sequencing data were then analysed using the programme MEGAv7.0 (Molecular Evolutionary Genetics Analysis) (Tamura et al., 2013) (http://www.megasoftware.net/)

2.15.3 Analysis for expression profile

QSVanalyser is a program that calculates the relative levels of two sequence variants from sequence trace files. The analysis is performed in batches which can contain several trace files over each other (Carr et al., 2009). Upon amplification and sequencing of the desired region on cDNA samples of NA18517, NA19240, NA18555 and NA19129, the trace files were loaded to the QSVanalyser to get the raw data, variant data, intensity and averages of the variants for each sample. Figure 2.7 below shows examples of the sequencing traces of the two SNPs. As each sample was sequenced multiple times, averages and standard deviation were calculated. The averages were normalised to one to allow better visualisation and comparison of the averages of the variants across the samples.

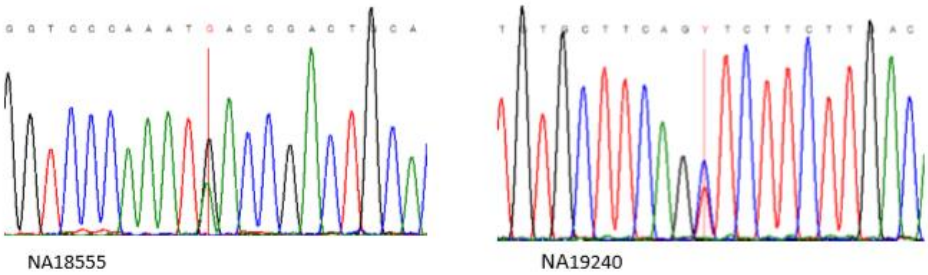


Figure 2. 7 A screenshot of QSV analyser showing heterozygosity.

The sequencing trace of NA18555 shows that the cell line is heterozygous for rs844. The sequencing trace of NA19240 shows that the cell line is heterozygous for rs60519172.

CHAPTER 3 Sequence read depth analysis of CNV using a reduced reference

3.1 Introduction and the study rationale

Segmentally duplicated regions have various sizes, orientation, content, and most of them show extreme organizational complexity. Characterisation of sequence variation within these regions using NGS can be challenging due to extensive copy number variation, gene conversion, long identical repetitive elements, chimeric repeats, mismapping of sequence reads to alternative paralogues and the diploid nature of the human DNA source (Zhang et al., 2011b). In the process of mapping short reads to a reference sequence, some reads will possibly match with several loci in the reference sequence and not uniquely to any specific region using the current mapping programs (Tattini et al., 2015, Mostovoy et al., 2016). This may cause problems with the correct identification of sequence variation within the segmentally duplicated regions. This issue can in theory be resolved by mapping reads to a reduced reference sequence that has only one variant of the duplicated regions in the human genome (Figure.3.1). Thus, the short-read sequences from high coverage data can be mapped by using different mapping tools under a range of different mismatch parameters to investigate the sequence variation between the copy variants of the duplicated regions. High-coverage data is desirable because the short-read coverage must be sufficient for a complete and accurate assembly of the genomic sequence since it increases the possibility of identifying the possible variant at a genome region (Sudmant et al., 2010; Zhang et al., 2011b). While mapping the short reads of high coverage data against reduced reference, the read depth (RD) can be used because RD approaches assume a random distribution (Poisson or modified Poisson) in mapping depth and investigate the divergence from this distribution to explore deletions and duplications. They are based on the assumption that the depth of the coverage of a genomic region is correlated with the copy number of the region, namely, higher copy number of a particular region is expected to have a higher read depth when compared to normal (e.g., diploid) regions. (Magi et al., 2012; Tattini et al., 2015).

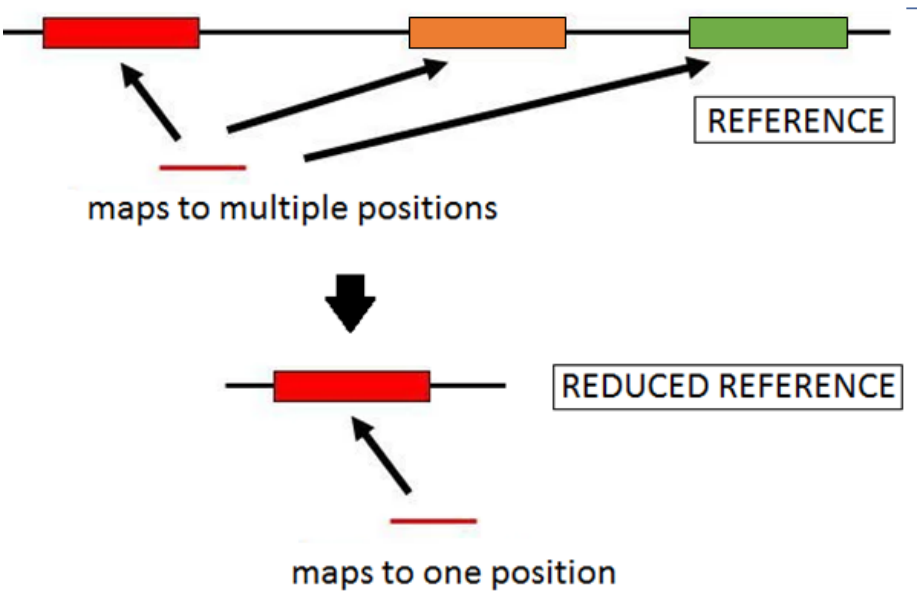


Figure 3. 1 Problem of the mapping approach with a reference sequence. When the short reads are mapped to an original reference sequence which itself is segmentally duplicated or copy number variable(coloured by horizontal rectangles), it may not be possible to identify if these reads are mapped to the correct copy of sequence because the short reads can be mapped to multiple positions. If a reduced reference sequence is used for mapping with different mismatch parameter, all short reads can be forced to be mapped to one copy. This can lead us to identify possible sequence variation between the copies through the first copy of the variants.

This chapter is the first part of a pipeline to resolve the ambiguity in read mapping to the duplicated DNA regions in the human genome by using a reduced reference so that the sequence variation between the copies can be investigated. In this this chapter, I constructed a reduced reference as a test case and obtained the publicly available high coverage data as short paired-end reads. Subsequently, the short reads were mapped against the reduced reference using tools based on different mapping algorithms. Also, I aimed to confirm the copy numbers of the genome regions used in the reduced references, and to compare the different mapping tools on both quality filtered and unfiltered datasets for the high coverage samples for which copy numbers are known from previous studies. Sequence variation calls from the reads mapped to reduced references will be discussed in the next chapter (Chapter 4).

3.2 Mapping against reduced references

3.2.1 Constructing the reduced reference

In order to test the accuracy of the approach, a reduced reference (RR) was constructed using three previously defined segmentally duplicated regions in the human genome; Rh Blood group genes (*RHD* and *RHCE*), low-affinity FCGR locus, Beta-defensin gene cluster. This was done by including only one copy of segmentally duplicated regions with 10kb flanking sequences from both ends and excluding the other copy based on the coordinates of the GRCh37/hg19 human reference genome. According to coordinates given in Table 3.1, the DNA sequences were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu). For each region, the coordinates were selected on the track page. From the 'View' option, the 'DNA' was selected. The 'Mask repeats' option was set to 'N' as sequencing formatting option, then the sequences were downloaded. UCSC uses the latest versions of RepeatMasker (http://www.repeatmasker.org/), a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences, and repeat libraries available on the date. Repeats are common sequence motifs duplicated hundreds of times in the genome. Repeat sequences lead to false overlaps while mapping reads to reference sequence. To avoid false matches while read mapping, repeats therefore needed to be masked. Non-copy number variant regions in the human genome were also included to normalise for copy number analysis. These non-copy variable regions correspond to the 10kb flanking sequences on the upstream/downstream or between of the segmental duplicons. Once RR is constructed, another reduced reference, named alternative reduced reference (ARR) was also constructed. Because both copies of these segmentally duplicated regions are assembled in GRCh37/hg19, the ARR is the same as the RR but with the alternative assembled duplication region. While RR has first assembled copy of the segmental duplication (Figure 3.2), the ARR has the second assembled copy of the segmental duplication (Figure 3.3).

Table 3. 1 The coordinates of the GRCh37/hg19 human reference genome for constructing the RR and ARR.

Regions for RR	Reference coordinates	Regions for ARR	Reference coordinates	Description of the region
R1	chr1:25,584,516-25,594,515	R1	chr1:25,584,516-25,594,515	10kb flanking sequence
R_C1	chr1:25,594,516-25,655,519	R_C2	chr1:25,655,521-25,688,913	Copy number variable region
R2	chr1:25,655,521-25,688,914	R2	chr1:25,688,914-25,751,819	Non copy number variable region
R3	chr1:25,751,819-25,761,819	R3	chr1:25,751,820-25,761,819	10kb flanking sequence
F1	chr1:161,050,791-161,237,982	F1	chr1:161,050,791-161,237,982	Non copy number variable region
F2	chr1:161,469,969-161,479,969	F2	chr1:161,469,969-161,479,969	10kb flanking sequence
F_C1	chr1:161,479,970-161,565,133	F_C2	chr1:161,565,134-161,647,427	Copy number variable region
F3	chr1:161,647,428-161,657,427	F3	chr1:161,647,428-161,657,427	10kb flanking sequence
B1	chr8:7,102,590-7,112,589	B4	chr8:7,559,589-7,569,589	10kb flanking sequence
B_C1	chr8:7,112,590-7,442,590	B_C2	chr8:7,569,590-7,891,090	Copy number variable region
B2	chr8:7,442,591-7,452,590	B5	chr8:7,891,091-7,901,090	10kb flanking sequence
B3	chr8:21,173,211-21,893,410	B3	chr8:21,173,211-21,893,410	Non copy number variable region

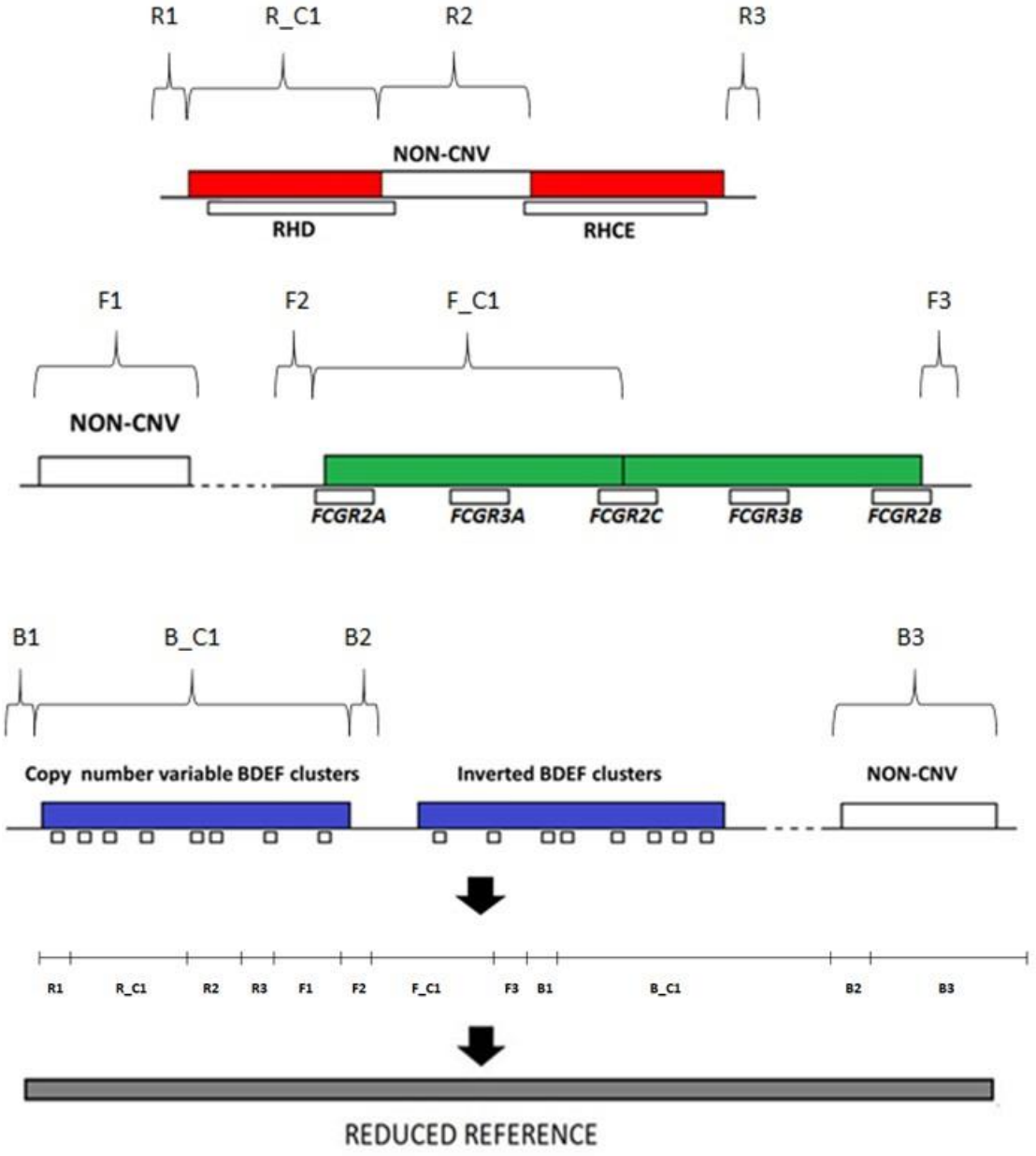


Figure 3. 2 Construction of the reduced reference sequence (RR). The segmentally duplicated regions for each gene cluster are coloured. Red horizontal bars: The *RHD/RHCE* repeat regions. Green horizontal bars: The low-affinity FC gamma receptor repeat regions. Blue horizontal bars: The beta-defensin repeat regions. The genes are represented as colourless boxes underneath the duplicated regions. A final reduced reference (RR) was produced by adding all the selected regions as indicated with a grey horizontal bar at the bottom.

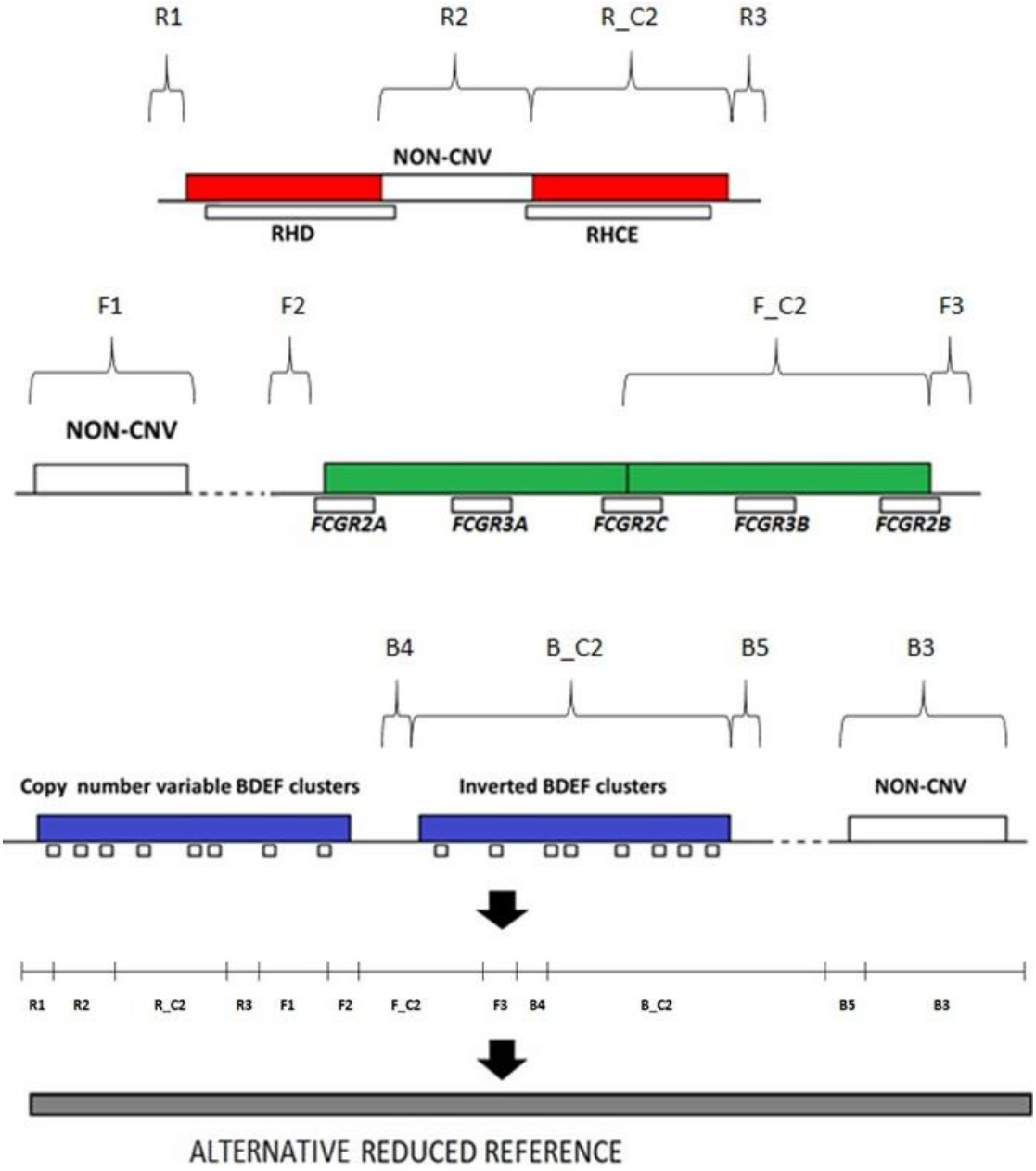
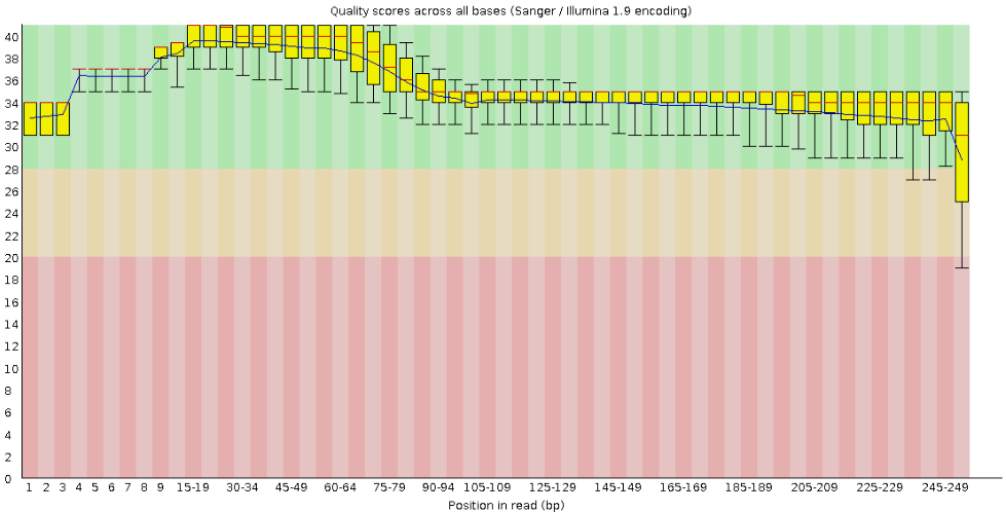


Figure 3. 3 Construction of the and alternative reduces reference (ARR). The segmentally duplicated regions for each gene cluster are coloured. Red horizontal bars: The RHD/RHE repeat regions. Green horizontal bars: The low-affinity FC gamma receptor repeat regions. Blue horizontal bars: The beta-defensin repeat regions. The genes are represented as colourless boxes underneath the duplicated regions. A final reduced reference (RR) was produced by adding all the selected regions as indicated with a grey horizontal bar at the bottom.

3.2.2 Obtaining high coverage data and quality check

The publicly available high coverage 2x250bp paired-end reads sequenced in fastq format for 24 individuals and the CEPH/UTAH 1463 pedigree samples (17 individuals) were downloaded to the HPC of the University of Leicester storage from the FTP site of the 1000 Genomes Project as fastq format (Chapter 2.1.1). Paired-end sequencing sequences both ends of a fragment and produces twice the number of reads compared to single-read data. Sequences aligned as paired allow generation of high-quality data, more accurate read alignments and have more capability to detect insertion-deletion (indel) variants. In order to avoid misleading results and quick identification of any problems, a quality check of the data is a necessary and important step before performing any analyses (Patel and Jain, 2012). An example of sample after trimming can be seen in Figure 3.4 where both pairs were filtered by quality. The quality of all the samples were checked and were filtered (Chapter 2.2.2). Bases in short reads have low quality below 3 (Phred quality score) from the start and the end were trimmed. The minimum length was kept as 250 and 150 bases long for mrsFAST-Ultra and BWA-MEM, respectively. Reads were also scanned with a 4-base wide sliding window and cut if the average base quality is below 10. 'N' bases were also removed. However, trimming these sequences caused a high amount of data loss shown in appendix 3. Even more than 50% of the total reads had to be trimmed because of low quality. Eventually, two sets of quality filtered data generated (first 250bases long paired end data for mrsFAST-Ultra mapping, second is >150 base long paired end data for BWA-MEM mapping). Both filtered and non-filtered data were used for the mapping to compare the results to each other.

A



B

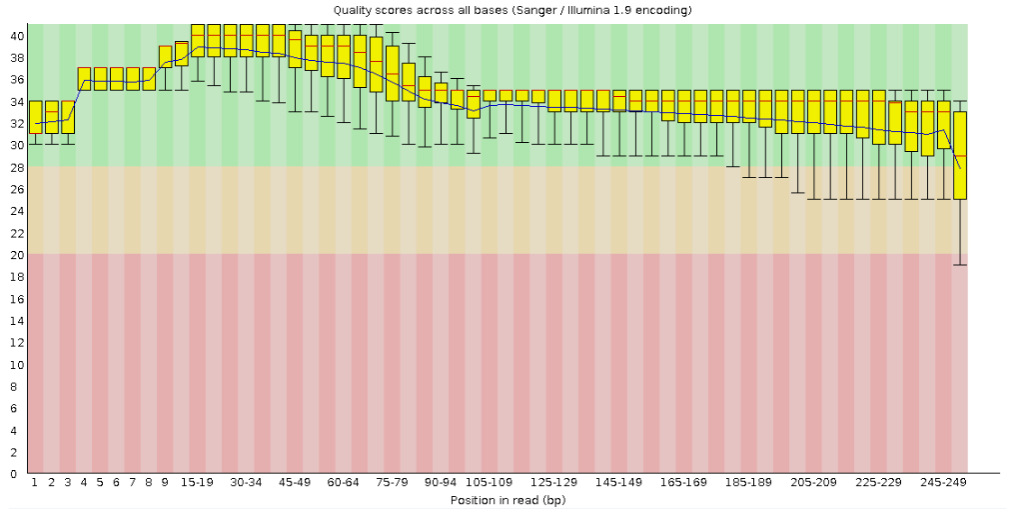


Figure 3. 4 An example file of 'per base sequence quality' after filtering a sample. For both the figures, the x axis shows the positions of the sequences and the y axis shows the quality of the bases. Each vertical yellow bar represents the quality of corresponding bases. If the yellow bars especially in the pink area (indicates quality lower than Q20, namely, the accuracy of the base is less than 99%), this means the sequence quality is bad. Once the reads were filtered as seen in the A and B (pair 1 and pair 2, respectively, sample is HG00096), the bars are over the Q20 line (light pink and green area), meaning the accuracy of the bases is over 99%. If the yellow bars mostly in the green area, this means accuracy of the bases is over 99% (Q30).

3.2.3 Mapping of short reads to reduced reference

Mapping algorithms help to find corresponding sequences in the reference sequence (Reinert et al., 2015). Over the years, many alignment programs have been developed based on different algorithms. The objective of alignment is to correctly determine each read's corresponding location in the reference genome (Ruffalo et al., 2011). In this study, two different of mapping tools were used; BWA and mrsFAST-Ultra.

BWA is a software package that maps highly similar sequences to a reference genome. BWA uses an index built with the Burrows Wheeler transformation (Ruffalo et al., 2011). The data structure, named FM-index, supports exact matching (Ferragina and Manzini 2000). By transforming the genome into an FM-index, the search performance of the algorithm improves where a single read matches multiple location in the genome. However, this significantly creates a large index build up time compared to hash tables (Hatem et al., 2013). BWA reports a meaningful quality score for the mapping that can be used to remove mappings that are not well supported due to e.g. a high number of mismatches (Ruffalo et al., 2011).

mrsFAST-Ultra reports all mappings of a read to a genome rather than a single 'best' mapping. This ability is thought to be useful in the detection of copy number variants (Bailey et al., 2002). mrsFAST-Ultra use a seed-and-extend method for alignment. First, it builds an index from the reference genome for exact 'anchor' matching. Second, it computes all anchor matchings for each of the reads in the reference genome through the index and extends each match to both left and right. At the end, it reports the total alignment within the user defined error threshold (Hach et al., 2014).

Overall, both mapping programs start by indexing the reference genome, however, they use different algorithm approaches for the mapping as mentioned above. Both programs support multithreading so the speed can be accelerated. BWA-MEM was used instead of BWA-ALN because BWA-MEM shows a better performance for longer reads (more than 100bp) since it has a higher accuracy at different mismatch levels (Li, 2013) while BWA-ALN is more suitable for shorter sequences (less than 100bp). Therefore, BWA-MEM was used for mapping paired-end reads (>150 bases long) for the dataset used in this study.

Picking the right tool for short read mapping is essential. Selecting the right parameter is also as important as picking the right tool. Error threshold is a parameter while looking for a match between a reference and short reads. If the error threshold is set at 0, the downstream analysis will not find any differences between the reference and the sequenced genome. Thus, the variants which exist will be missed. If the error threshold is too high, allowing many mismatches between reference and read, many incorrect will result with a high number of false-positive SNPs in downstream analyses Mismapping may lead to both false-positive and false-negative calls such as mutations being missed and mutations being erroneously called (Altman et al., 2012; Yohe and Thyagarajan, 2017). In this study different mismatch parameters were used. The short reads in fastq format were mapped to the RR with 6, 8 and 10 error thresholds, meaning that reads were mapped to reference with a minimum of 94%, 92% and 90% similarity, respectively. Once the reads were mapped to reference genome, SAM/BAM files were created, and BAM refinement was performed. Insertions/deletions which are not present in the reference genome can cause small misalignments. By doing local realignment for these regions, the number of mismatching bases is reduced. PCR duplicates as being introduced during library construction are also removed. These are often removed because there is concern that they can lead to false positive variant calls (Ebbert et al., 2016). Therefore, higher quality and more accurate results can be achieved (O'Rawe et al, 2013). Only the results with 92 % similarity of short reads to reference in sequence on quality filtered and unfiltered version of the data were used because there were no substantial differences between the calculated ratios among the different mismatch thresholds of the mapping for all the samples and the gene clusters (Appendix 7).

A FLAGSTAT file provides read counts for different categories such as the number of quality passed/failed reads, the number of singletons, the number of paired reads in sequencing. The FLAGSTAT statistics was performed to do a full pass through BAM files to calculate and print summary statistics for each sample. The number of duplicates (if any) is also shown so that it can be confirmed the PCR duplicates were removed. The total number of mapped reads can be obtained. Thus, the total number of mapped sequences was used to calculate the average coverage for each sample by using the

formula; (total number of mapped short reads X length of the short reads) / the length of the reference sequence for the RHD, FCGRs, and BDEF regions. The final BAM files are used to count how many reads are mapped according to the RR and ARR defined regions for quality filtered and unfiltered versions of the samples for both mapping tools. By using a BED file (defining the borders of 10kb flanking sequences, the CNV or Non-CNV and regions for each gene cluster) and by calling the counting option of SAMtools, reads mapped to different defined regions were counted. Then a ratio is generated by dividing the number of mapped reads of the CNV region by the number of mapped reads of non-CNV regions for RHD, FCGRs and BDEF separately. An example of the calculation is shown in appendix 11.

3.2.4 Comparison of mapping tools

To investigate and compare the performance of both mapping programs, the generated ratios were compared by using samples for which we know the copy numbers from the previous studies for RHD, FCGRs and Beta-defensins regions (Handsaker al., 2015; Forni et al., 2016). We looked for if there is a correlation between the generated ratios and the known copy number. These ratios are the result of mapping with 92% mismatch of the short reads and the RR. According to Figure 3.5, 3.6 and 3.7, there are four groups of comparison for each plot. These are quality filtered and unfiltered data results with BWA-MEM mapping, quality filtered and unfiltered data results with mrsFAST-Ultra mapping. The number of samples used for each gene cluster is 7, 5 and 13 for RHD, Fc gamma receptors repeat, and Beta-Defensins repeat regions, respectively.

The ratios generated from BWA-MEM mapping are clustered precisely better than the ratios from mrsFAST-Ultra mapping for RHD where the samples only have copy number 1 and 2. There is a clear separation and clustering among the corresponding the copy number 4 and 5 and the ratios of BWA-MEM mapping for the filtered data of FCGRs. For Beta-Defensins, all the ratios are clustered less precisely based on the corresponding copy numbers except the ones have copy number 5 for BWA-MEM mapping which shows better separation and clustering compared to mrsFAST-Ultra mapping. Overall, the plots show that there is a consistency between the known copy numbers and the generated ratios from the BWA-MEM mapping compared to mrsFAST-Ultra mapping for both filtered and unfiltered version of data. Also, the unfiltered version of the samples

relatively shows more robust correlation compared to the filtered version. The filtering of the data was performed by using minimum length parameter 250 for mrsFAST-Ultra since it requires uniform length for the pair-end reads. On the other hand, BWA-MEM can perform without uniformity of paired reads. As a result, trimming short reads and keeping short reads as 250 base long (for mrsFAST-Ultra mapping) resulted that many reads were filtered because of quality issues even though they would have been mapped to desired region. Thus, this data loss may affect the mapping and, therefore, the generated ratios from the mapping results are not clustered and not separated precisely for mrsFAST-Ultra mapping for filtered version of the data compared to the quality unfiltered version. As a result of these comparison on mapping tools, BWA-MEM with 92% mismatch ratio mapping on filtered data results were decided to be used for further analysis.

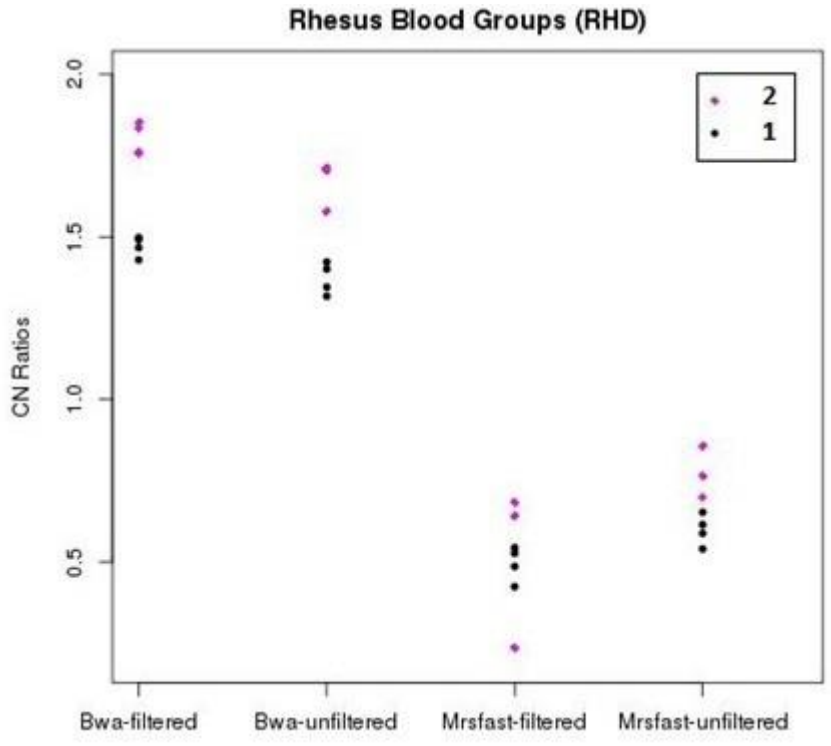


Figure 3. 5 Comparison of mapping tools for *RHD/RHCE* repeat region. Each coloured symbol in the plots corresponds to an individual with known copy number 4 and 5. A total number of 7 individual with known copy were used for each group.

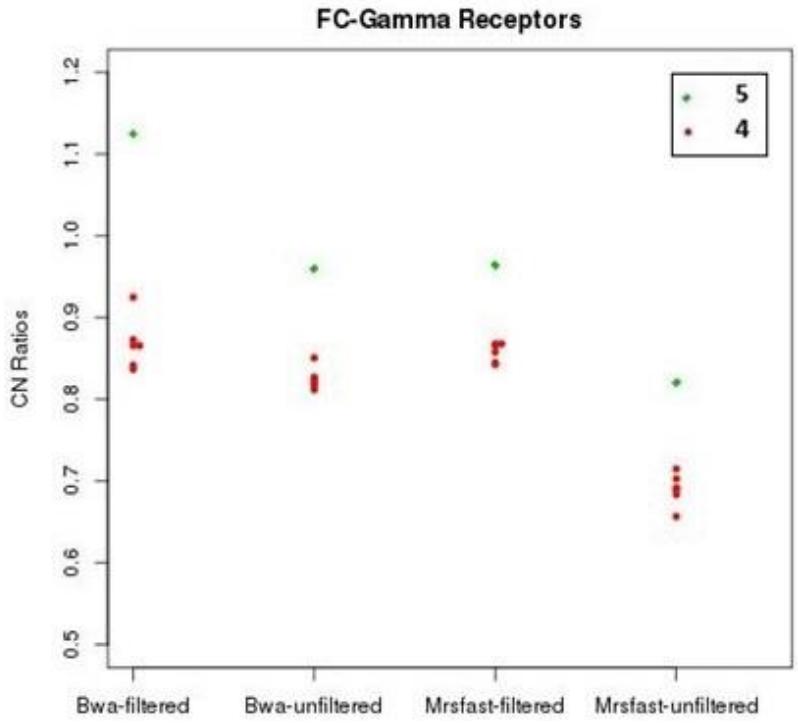


Figure 3. 6 Comparison of mapping tools for the low-affinity FCGRs repeat region. Each coloured symbol in the plots corresponds to an individual with known copy number 4 and 5. A total number of 5 individual with known copy were used for each group.

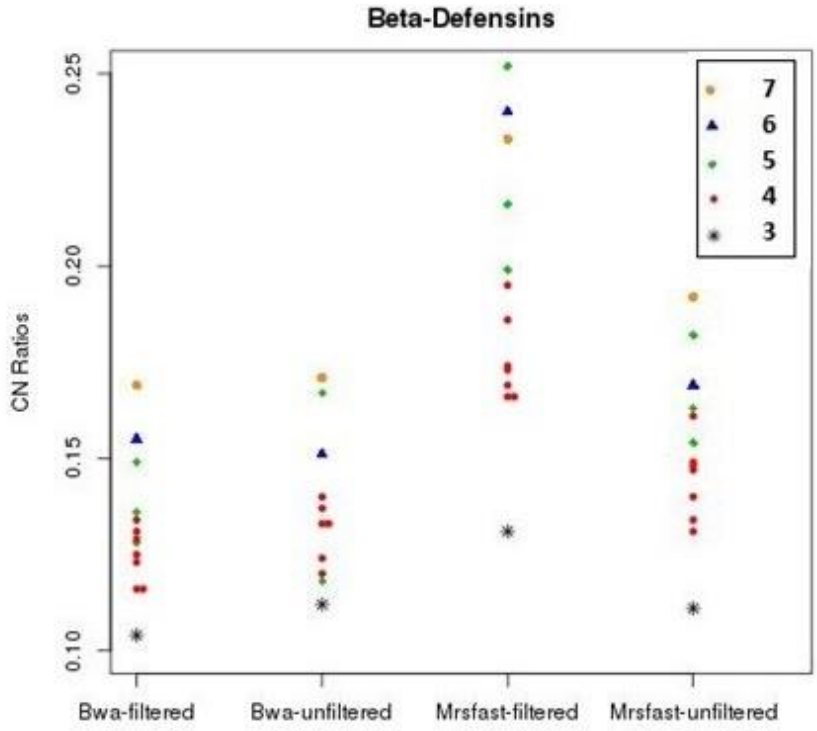


Figure 3. 7 Comparison of mapping tools for the Beta-Defensins repeat region. Each coloured symbol in the plots corresponds to an individual with known copy number 3, 4, 5, 6, and 7. A total number of 13 individual with known copy were used for each group.

3.3 Copy number calling

3.3.1 Copy number estimating by PRT for the low-affinity FCGR locus

To find the copy number of the low affinity FCGRs region for the 1000 genomes project and the platinum pedigree samples, a previously developed PRT assay was used (Niederer et al., 2010). If the copy numbers of these samples are detected, then these copy numbers can be used to confirm previously studied samples and can be used to compare the copy numbers found from the bioinformatics pipeline used here.

The PRT was used to determine the copy number of the low-copy FCGRs locus with the aim of confirming the samples that copy numbers were already known and estimating the copy numbers for the remaining samples of the data set used in this study. The copy number variable (test) region was chosen in the 3'-untranslated region (3'UTR) of *FCGR2C* as 274bp and non-variable (reference) was chosen in the 3'UTR of *FCGR2A*.as 279bp (Niederer et al., 2010). Both loci were amplified by using one pair of primers to estimate total the copy number of the samples for FCGRs region (Figure 3.8).

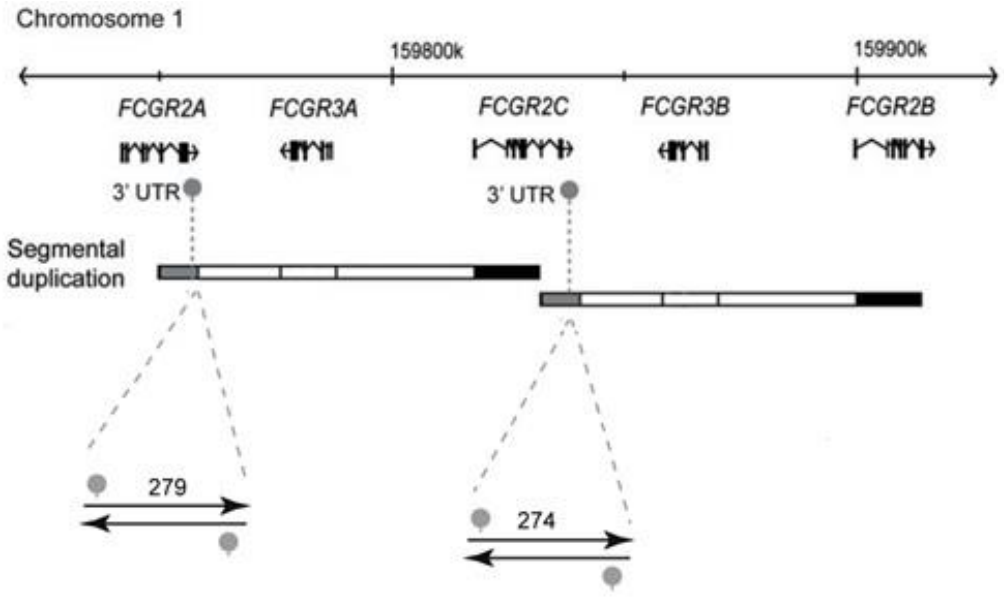


Figure 3. 8 The amplified loci for the PRT assay of FCGRs region. The copy number variable (test) region was chosen in the 3'-untranslated regions (3'UTR) of *FCGR2C* as 274bp and non-variable (reference) was chosen in the 3'UTR of *FCGR2A*. Modified from Niederer et al., 2010.

By using the calibration curve (Figure 2.5 in Chapter 2.6.3) generated from samples of known copy number, the raw copy numbers of the 24 of 1000 Genomes project and 17 Platinum data (CEPH/UTAH 1463) DNA samples were estimated. The raw copy numbers generated for both assays were compared using scatter plots to observe distinct clusters (Figure 3.9). This comparison provides a clear understanding of the data quality and shows the assays are good to use for copy number calling. The scatter plot shows the comparison between raw data normalized to known copy number of FAM assay and raw data normalized to known copy number of HEX assay in this dataset. According to Figure 3.10, both FAM and HEX assays give similar raw copy numbers and shows clear separation in clusters.

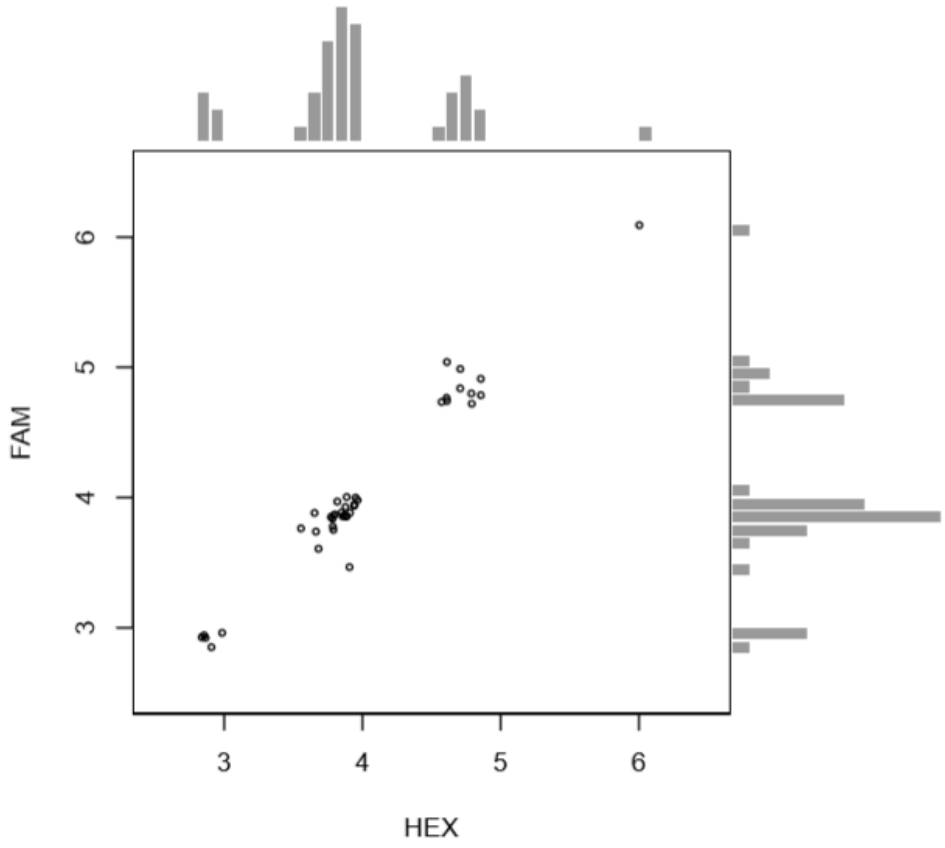
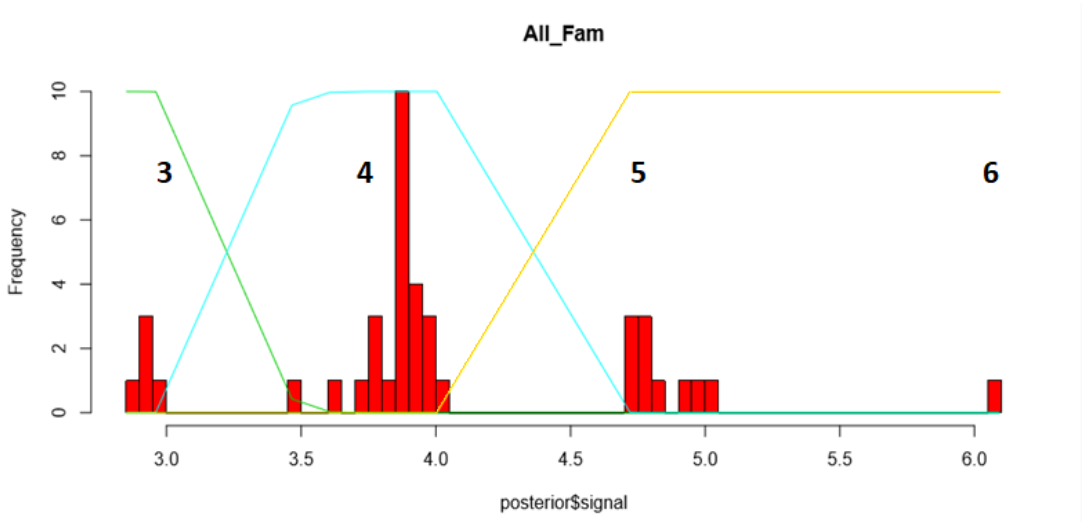


Figure 3. 9 The comparison of raw PRT data for the whole dataset. The scatter plot shows the comparison between raw data obtained from HEX and FAM labelled primers. In total 41 samples used.

The software CNVtools (within R) was (see Chapter 2.6.4) used to create a histogram of PRT ratio data for both HEX and FAM assays (Figure 3.10). The histograms of both assays indicate four clusters. The number of clusters was used to classify integer copy number for the samples. The clusters were counted as 3, 4, 5 and 6 based on control samples.

A



B

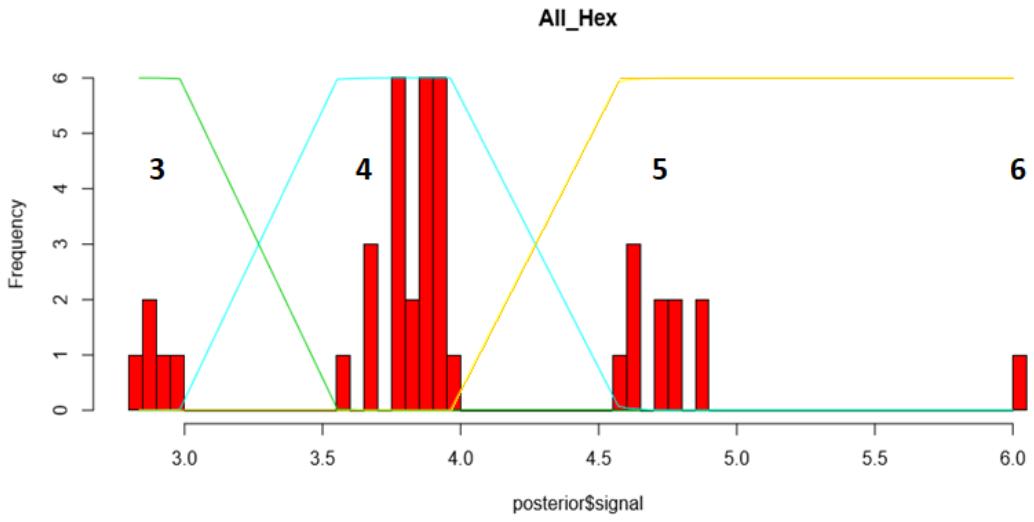
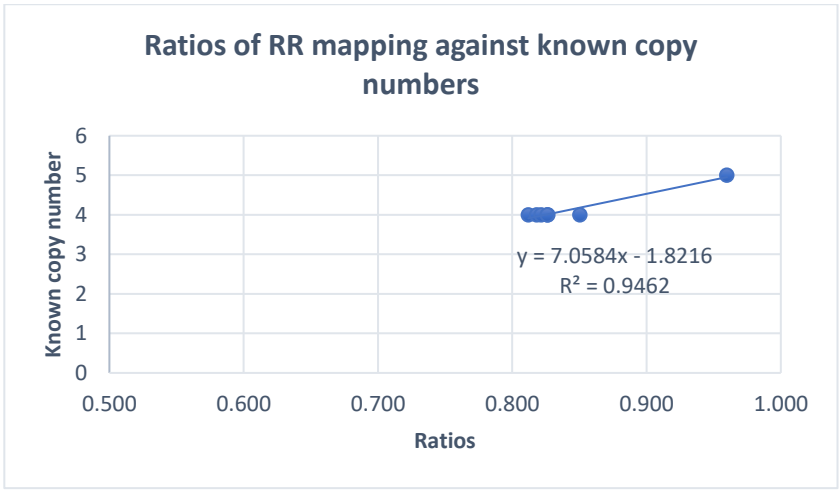


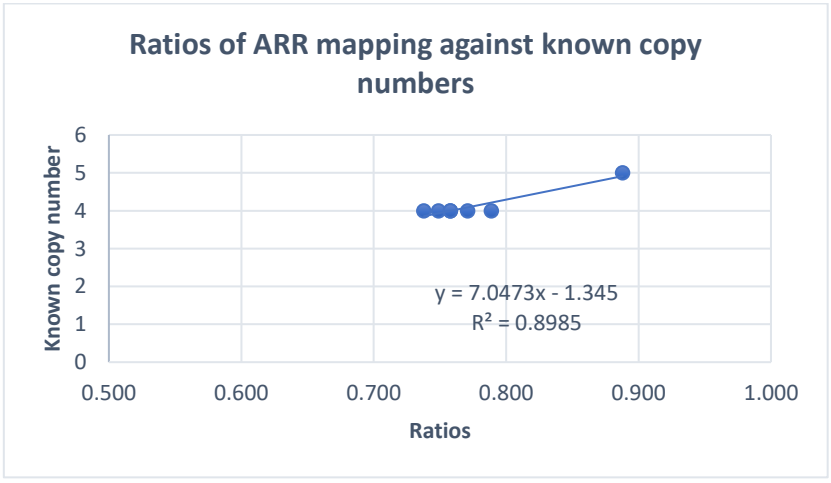
Figure 3. 10 Distribution of low affinity FCGRs copy number across 1000 Genomes samples. A: The distribution of the estimated copy numbers from the FAM-labelled PRT Assay B: The distribution of the estimated copy numbers from the HEX-labelled PRT Assay .The coloured lines for each figure show the Gaussian distributions for each of the copy number classes (copy number =3, 4, 5 and 6). The x-axis indicates diploid copy number data. The y-axis represents the frequency scale. An estimation of the integer copy number is achieved by combining all raw values for both assays.

3.3.2 Copy number comparison for the FCGRs

The low affinity FCGRs repeat regions were used to estimate the copy number of the individuals from the BWA-MEM mapping results. The ratios generated from the BAM files of the BWA-MEM mapping result for all the samples were used for the copy number prediction and this was performed by regression against a standard curve of regions of known copy (Figure 3.11). By using the regression plot the raw copy numbers were estimated and the integer copy numbers were detected by using the CNVtools.



A



B

Figure 3. 11 Standard curves of repeat regions of known copy against generated ratios for the FcgRs. While A shows the linear regression plot of the ratios from RR mapping against known copy numbers, B shows the linear regression plot of the ratios from ARR mapping against known copy numbers.

Once the integer copy numbers are obtained, these copy numbers were compared with the previously known copy numbers and the copy numbers obtained from PRT assay

used in this study. According to Table 3.2, the estimated copy numbers of all the 24 of 1000 Genomes project samples are the same from different sources except two samples which are HG01583 and HG01879. These samples have the same copy number from both mapping approaches; however, the copy numbers did not match with the PRT assay. For the samples HG01583 had very low mapping coverage compared to other samples in the whole data set. It was 7.9X with RR mapping and 7.8X with ARR mapping. This low coverage could be because the percentage of short reads left after trimming this sample by quality (Appendix 3). More than 80% of the total reads were filtered. Both low coverage and having most of the reads filtered for this ample can cause having misleading mapping results. On the other hand, HG01873 (29.1X with RR mapping and 28.5X with ARR mapping) has a higher coverage and the percentage of short reads left after trimming for this sample by quality for BWA is about 50% which is higher compared to the other samples. However, this sample was repeated for the PCR and ABI running for several times and failed for many times which can explain why the copy numbers from the PRT assay and mapping results do not match each other and this could be a DNA sample problem.

Table 3. 2 The comparison of the copy number of the samples from different approaches.

Sample	Handsakar et al 2015	RR Mapping	ARR Mapping	PRT
HG00096	4	4	4	4
HG00268	4	4	4	4
HG00419	5	5	5	5
HG00759		4	4	4
HG01051	4	4	4	4
HG01112	4	4	4	4
HG01500		4	4	4
HG01565		5	5	5
HG01583**		**6**	**6**	**4**
HG01595		3	3	3
HG01879**		**4**	**4**	**3**
HG02568		3	3	3
HG02922		3	3	3
HG03006		4	4	4
HG03052		4	4	4
HG03642		4	4	4
HG03742		4	4	4
NA19625		4	4	4
NA19648	4	4	4	4
NA18525		4	4	4
NA18939		4	4	4
NA19017		4	4	4
NA20502	4	4	4	4
NA20845		4	4	4
**Samples not consistent with mapping and PRT results.				

For the pedigree samples, the same process was performed, and the integer copy numbers were obtained from both mapping approaches. Then, the bam files from two parents (NA12877, NA12878) who are the children of the four founders of the pedigree were downloaded from the 1000 Genomes project website. The bam files of the 11 children were not available; therefore, these samples were not processed. The same regions which were defined by constructing the RR and ARR were used to count the total number of mapped reads to generate ratios for the six samples, and the copy number were estimated. All the copy number predictions from the different approaches are compared in Figure 3.12 for the pedigree samples. As seen, the estimated copy numbers from RR and ARR mapping, PRT results and the 1000 genomes project bam files are the same and do support each other for the six samples. For most of the children the copy numbers also support each other except for 3 samples. Although one of the mapping results is the same copy number as PRT assay, the other mapping result is not giving the same copy number. For two samples (NA12880 and NA12882), while PRT and ARR mapping results support each other, RR mapping results show different copy number. For one sample (NA12893), while RR mapping and PRT results support each other, a different copy number obtained from ARR mapping results. As previously mentioned in the introduction, the low affinity FCGRs region has a deletion/duplication of FCGR3A/ FCGR2C or FCGR2C/FCGR3B. For these samples, if there is a duplication, it is not known which paralogue has the duplication. If the duplication is a FCGR3A/ FCGR2C on the paralogue A, then almost two times higher reads will be mapped to the RR. If the duplication is on the FCGR2C/FCGR3B on the paralogue B, then almost two times higher the reads will be mapped to the ARR. This might be the reason why the copy number of one of the mapping results is higher than the other. Therefore, it can be concluded that the duplication of the FCGR3A/ FCGR2C is on the paralogue A for NA12880 and NA12882, and the duplication of the FCGR2C/FCGR3B is on the paralogue B for NA12893.

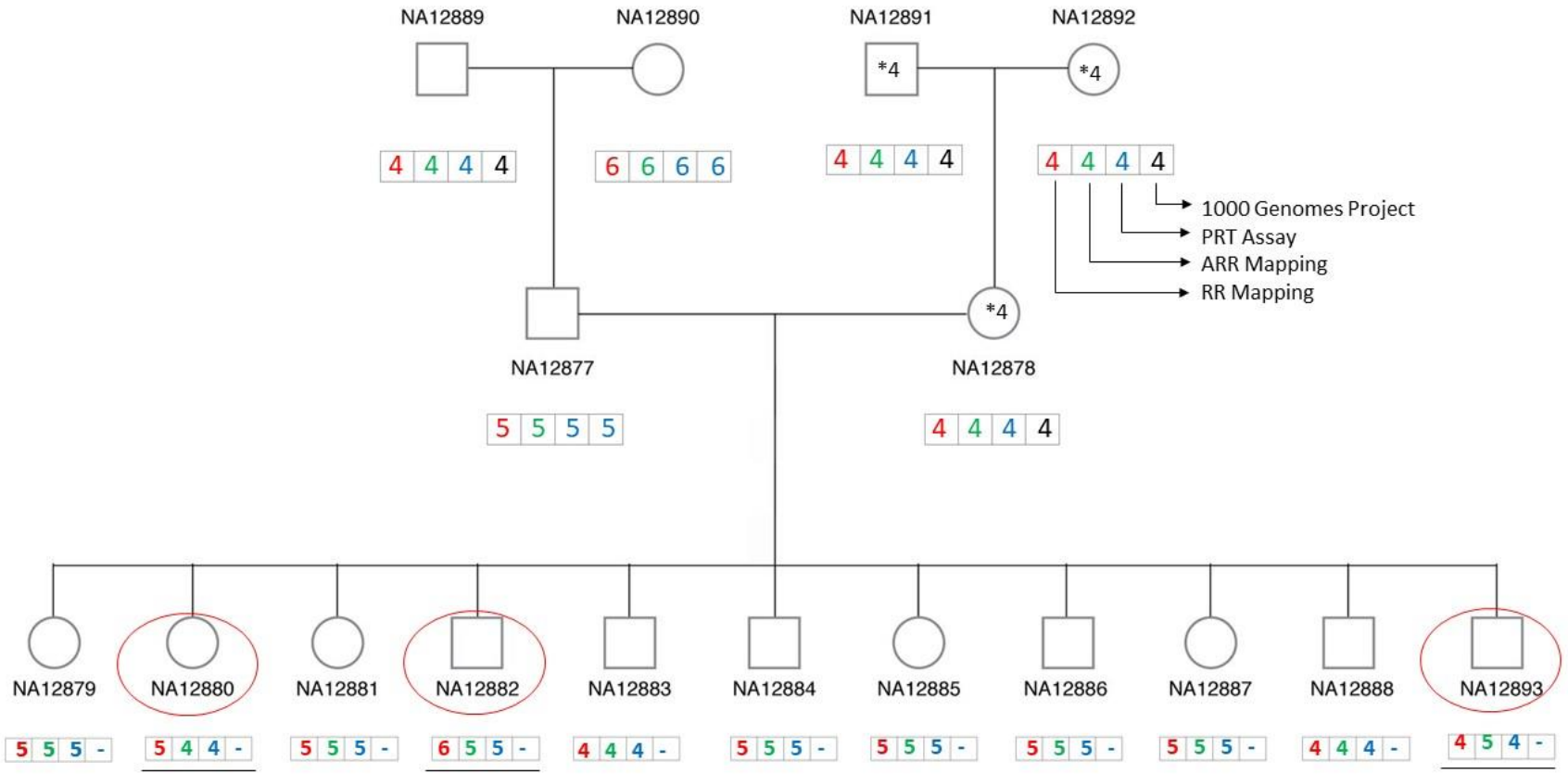


Figure 3. 12 The FCGR copy numbers found by RR mapping and ARR mapping on the platinum data samples. *The copy number found by previous study (Hollox et al., 2008a). The estimated copy numbers were represented in boxes below the individuals as the RR mapping, ARR mapping, PRT assay, and the copy number calculated from 1000 genomes project bam files, respectively. The individuals with different copy number estimation are indicated with red circles.

3.4 Discussion

This chapter is the first part of the pipeline to investigate the sequence variation of the duplicated DNA regions in the human genome by using a reduced reference. The publicly available 41 high coverage paired–end Illumina samples were mapped to reduced reference sequences using two different mapping tools with different mismatch ratios. The mapping results were used to compare the mapping tools and strip plots were generated for each DNA repeat region to see whether a positive relationship/correlation between the generated ratios from the mapping analyses and already known copy numbers. According to the figure 3.5, 3.6 and 3.7, approach cannot make a clear separation and clustering for the Beta-defensins repeat region. For other gene regions (RHD and FCGRs), BWA-MEM mapping results show a better performance compared to mrsFAST-Ultra even though this tool is designed to map short reads generated with the Illumina platform to reference genome assemblies.

mrsFASTaligner (a previous version of mrsFast-Ultra) was used for mapping short-reads from 1000 Genomes project for 159 human genomes with using read-depth approach and demonstrated the estimation of copy number accurately for duplications (Sudmant et al., 2011). MrsFAST-Ultra was also used to study diploid copy number of the beta-defensin genomic region not only detecting the copy numbers but also calling sequence variants for 1285 samples from 26 populations by using read-depth approach with the high-coverage phase 3 exome sequences of the 1000 Genomes (Forni et al., 2015). In these studies, mrsFAST versions show significant performance for read mapping, however, BWA-MEM mapping for the filtered version of the data shows better performance for whole genome studies compared to BWA-MEM unfiltered version of the data and mrsFAST-Ultra as a result of the comparison in this study.

The robust PRT assay was applied to measure copy number variations of the low-affinity FcgR region for all the samples. The copy numbers of previously known samples were estimated the same and the copy numbers of the remaining samples were also detected. The copy numbers were also determined based on the RR and ARR mapping results. The copy number estimations from different sources support each other for most of the samples. 36 samples (out of 41) show the same copy number estimation from the PRT

assay, RR and ARR mapping. As explained previously, two samples have different copy umber estimation from the PRT assay although they have the same copy number from the mapping results. Three samples from the pedigree have different copy number estimation from different mapping results (summarized in Table 3.3). For NA12882 copy number 6 is not possible because the parents of this samples have 5 and 4 copy number respectively (NA12877(2,3) and NA12878(2,2)). NA12882 cannot be 3,3 unless there is a *de novo* mutation which is unlikely. Therefore, using a pedigree is a powerful way of testing the approach used in this study because the segregation of the allele provides consistent information.

Table 3. 3 The samples have diverse copy number estimation from different assays.

Sample	PRT	RR	ARR
HG01583	4	6	6
HG01879	3	6	6
NA12880	4	5	4
NA12882	5	6	5
NA12893	4	4	5

There is not enough evidence and insufficient samples to decide which mapping approach is better than other. There is also some disagreement between the mapping approaches and PRT assay. It can be concluded that trying to force all reads of a sample to map to a reduced reference does not work too well, at least for this region with a similarity of 97%, for some samples. It can be suggested that both approaches should be applied to segmentally duplicated regions and a wet lab validation method should be applied to estimate the copy number of a sample as final decision.

For the further analysis and the second part of this pipeline, the estimated copy numbers of the PRT assay were used to investigate sequence variation because the RR and ARR mapping results gave different copy number assessment for the listed samples above.

CHAPTER 4 Single nucleotide variant calling of Fc receptor region

4.1 Introduction and study rationale

Variant calling is the identification of positions where the sequenced sample is different from the reference sequence and it is performed to obtain information on the positions of genetic variants and their corresponding genotypes. The genomes of many individuals have been sequenced during the past decade. The sequence information generated by the projects such as the 1000 Genomes, the Cancer Genome Atlas and whole exome sequencing projects has clarified many genetic causes of many human diseases and has increased our understanding of genetic diversity (Cho et al., 2018).

A haplotype is defined as a sequence of alleles from the same chromosome. For the complete description and interpretation of the human genome, it is essential not only to discover the variation but also to arrange it onto haplotypes. Construction of the haplotypes from genotypes at multiple loci can provide a natural data reduction. This provides a more discriminative state of a chromosomal region. Haplotype data can increase the statistical power of assigning individuals to populations compared to individual SNP data. The importance of haplotyping information is increasing as we move into the era of large-scale sequencing. Applications of haplotype phase allows us to understand the interplay of genetic variation and diseases (Tewhey et al., 2011), impute untyped genetic variation (Marchini et al., 2007; Browning and Browning, 2009), genotypes in microarray and sequence data (Yu et al., 2009), detecting genotype error (Scheet and Stephens, 2008), inferring human demographic history (Tishkoff et al., 1997), inferring points of recombination and recurrent mutation(Kong et al. 2008), identifying signatures of selection (Sabeti et al., 2002) and modelling cis-regulation of gene expression (Tao et al., 2006; Browning and Browning, 2011). Whole-genome sequencing is becoming increasingly routine, however, individual genomes are mostly unresolved with respect to haplotype, particularly for rare alleles, which remain poorly resolved by inferential methods. Many bioinformatic pipelines have been developed to call variants from NGS data and construct haplotypes. However, in duplicated regions of the genome, the copies of genes or regions can accumulate

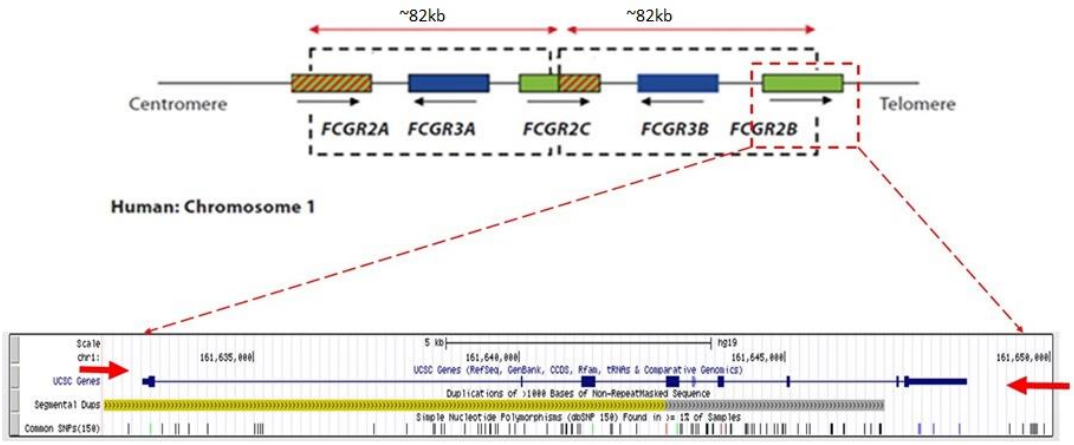
different variants and these variants can cause different functional consequences. In addition, recent duplicated sequences are highly similar to each other, they can be complicated to do genome assembly and investigate the sequence variation between the variants. Therefore, differentiating the variants in duplicated regions of the human genome and constructing haplotypes is a challenging process.

This chapter is the second part of a pipeline to resolve the ambiguity in the read mapping for the duplicated DNA regions in the human genome by using a reduced reference so that the sequence variation between the copies can be investigated and the haplotypes can be constructed across different copies. In this chapter, I aimed to call the variants from the samples mapped against the reduced reference based on their copy number found by PRT assay. I aimed to validate these variants by performing Ion semiconductor sequencing of the same samples for the *FCGR2B* locus. *FCGR2B* encodes the Fc gamma receptor IIb (FcgRIIb) (CD32B), the only inhibitory Fcg receptor which is located on the paralogue B of the segmentally duplicated region of the low-affinity FcgRs. Thus, the amplification of this locus is possible by long-range PCR. Furthermore, I aimed to construct haplotypes from both mapped samples and the Ion semiconductor sequenced samples so that both haplotypes from different source can be compared to each other to test the functionality of the pipeline.

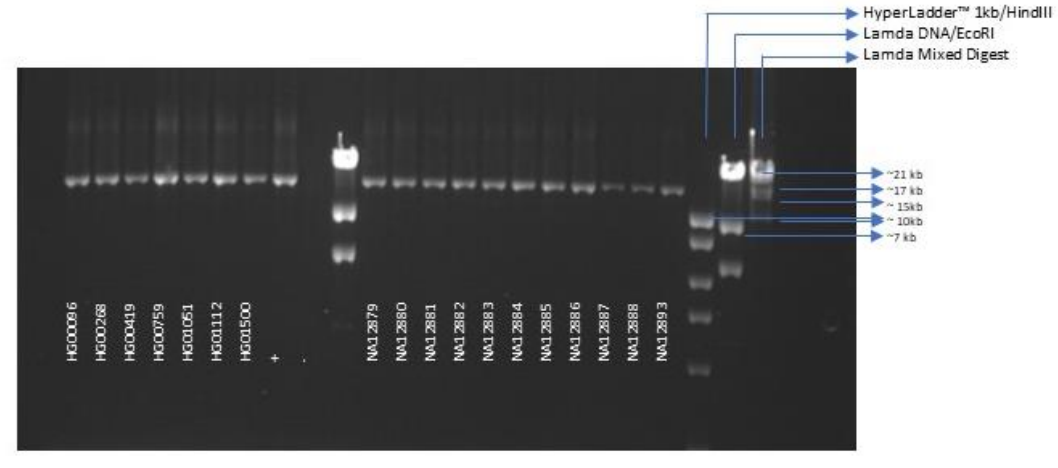
4.2 PCR of *FCGR2B* and Ion Torrent sequencing.

4.2.1 Amplification of *FCGR2B* locus by long-range PCR

In order to validate the variants from the mapping analysis, the complete *FCGR2B* gene (15540bp) was amplified and sequenced as a gold standard method. Because this gene is partially positioned outside of the segmental duplicated regions, one of the primers can be designed in the single copy region. Thus, the primers can only amplify the *FCGR2B* region and it becomes possible accurately genotype and discriminate the DNA sequence polymorphism from other FCGR genes. The complete *FCGR2B* gene with flanking sequences from both ends (16834 bp) was amplified for 24 of 1000 Genomes Project and 17 Platinum data samples. The reverse primer was specifically designed in the single copy region (Figure 4.1A). All the samples were amplified for the desired region and some are shown in the figure 4.1B.



A



B

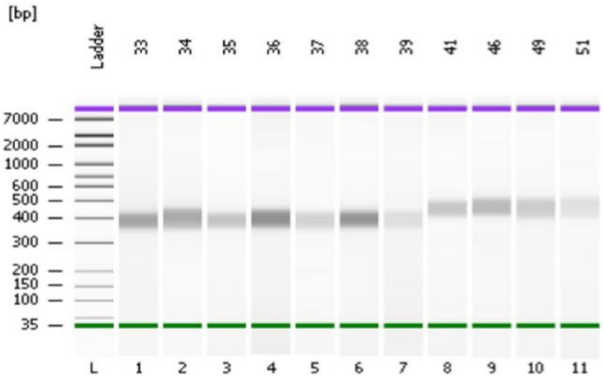
Figure 4. 1 Designing and the amplification of the long-range PCR assay of *FCGR2B*. A Primer design for the desired region. The coordinates are based on GRCh37/hg19 genome assembly. The red arrows indicate where the primers are designed. B Amplification of the desired region. The PCR product of *FCGR2B* region (16834bp) for some of the samples used in this study.

4.2.2 *FCGR2B* sequencing on the Ion Torrent (PGM) platform

Amplicon sequencing was performed with the Ion Xpress plus gDNA Fragment Library Kit according to User Bulletin. The Agilent Bioanalyzer™, a microfluidics-based platform for sizing, quantification and quality control of DNA and RNA, was used to analyse the size distribution of the sequencing libraries. The electropherogram (Figure 4.2) shows a range of 350-550 bp insert size for all the samples. Size selection may have a considerable impact on quality of results. If the size selection of the library has a certain range, it can be ensured that your clusters will almost all have the same insert size. High-quality size selection can boost sequencing efficiency, improve assemblies, and even allow sequencing of low-input samples. Sequencing was performed on an Ion Torrent

PGM with an Ion 316v2 chip using the Ion PGM Hi-Q Sequencing Kit. Based on the Ion torrent semiconductor sequencing run report, the loading percentage reached up to 83% of chip capacity getting a final throughput of 1.42 gigabases with an average coverage depth of more than 78071X. More than 4,368,813 filtered usable reads were obtained. A read length histogram was produced with a mode at 308 bases and a raw read accuracy of 99.4% (Figure 4.3).

A



B

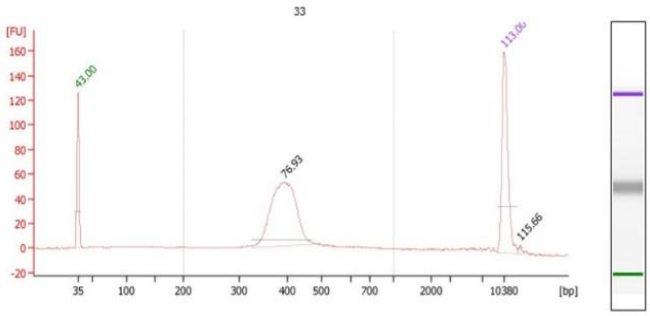
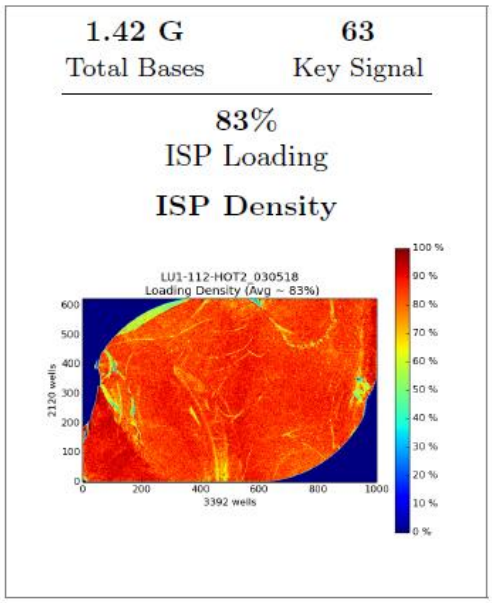
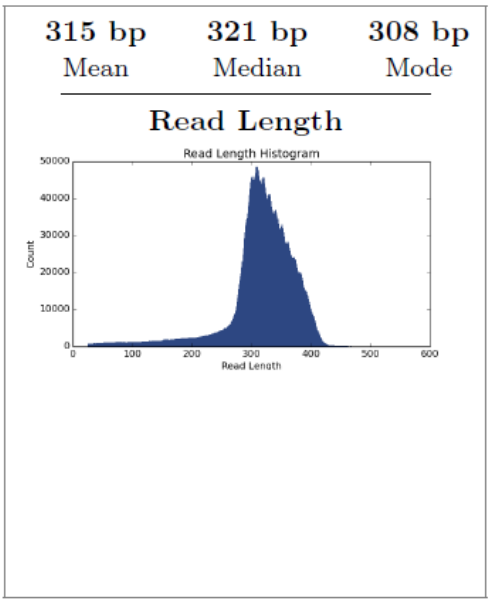


Figure 4. 2 The 2100 Agilent Bioanalyzer Electrophoresis file run summary. A Representation of the size and concentration of 11 samples from a single High Sensitivity DNA chip. B The size (bp) and the area/concentration (ng/uL) of the sample HG00096 (barcode 33). The area under the upper marker peak is compared with the sample peak areas. The concentration for each sample can be calculated because the concentration of the upper marker is known. To calculate the concentration of the individual DNA fragments in all samples, the upper marker, in conjunction with a method/assay-specific concentration against base-pair size calibration curve, is applied to the individual sample peaks in all sample wells.

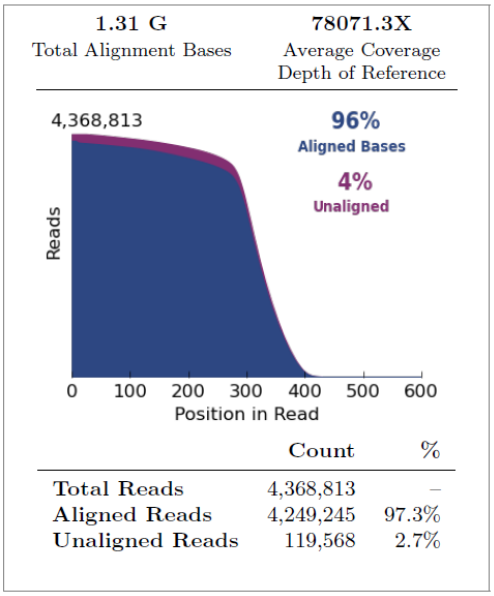
A



B



C



D

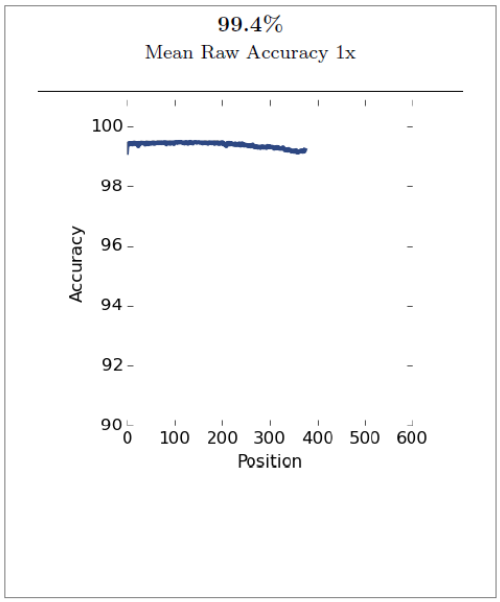


Figure 4. 3 *FCGR2B* sequencing on the Ion Torrent Platform run results. Summary section of the analysis report gives summary statistics of Ion Sphere Particle performance which includes a chip loading image. A: Ion Sphere Particle (ISP) Identification: total bases (1.42G) is shown as the total number of filtered and trimmed base pairs reported in the output BAM file. Key signal is the percentage of Live ISPs with a key signal that is identical to the library key signal. The ISP Density image is the image of the Ion Chip Plate showing percent loading across the physical surface (the percentage of chip wells that contain a live ISP). B: Read length Histogram which is a histogram of the trimmed lengths of all reads present in the output BAM file. C and D: Alignment summary that displays the number of millions of base pairs that have been aligned/unaligned/total to the genome at the specified quality level and mean raw accuracy of the reads.

Upon Ion Torrent sequencing, the mapped BAM and VCF output files were obtained for each sample from the Ion Reporter Software. The BAM files were visualized on the IGV following the manufacturer's manual. All the defined SNPs were confirmed, and the ambiguous heterozygous variants were checked if exited.

Ion Torrent Variant Caller (TVC) is a genetic variant caller for Ion Torrent sequencing platforms. It is designed to call single-nucleotide polymorphisms (SNPs), multi-nucleotide polymorphisms (MNPs), insertions, deletions, and block substitutions. The VCF files obtained from TVC. The detected variants were confirmed with the corresponding BAM file of each sample by visual inspection on IGV. These files were used in the software program BEAGLE (v4.1) for haplotype estimation.

4.3 Prediction of haplotypes of *FCGR2B* locus

In order to estimate the haplotypes from genotype data, computational methods pool information across individuals. Unrelated individuals can be phased by considering sets of common haplotypes that can explain the observed genotype data. The number of unrelated individuals in a dataset is an important factor to determine how well phase can be estimated. Thus, more individuals help a better estimation of haplotype phasing. Related individuals can be phased by considering haplotypes that are shared identically by descent between individuals within families. This within-family information is much more enlightening for phase estimation than the haplotype frequency information. However, haplotype frequency information across families or from the population can also be used to fill in the gaps in haplotype phase that are not determined by within-family information.

BEAGLE is a haplotype-inference method that performs a high computational speed and measures of accuracy for large whole-genome data sets. The software includes applications for haplotype and missing-data inference, single-marker and multilocus association analysis, and permutation testing using the early HMM-based phasing algorithm approach (Browning and Browning, 2007). BEAGLE was one of the tools used to estimate haplotypes for the 1000GP phase 1 array-based genotype data (1000 Genomes Project Consortium, 2010). In BEAGLE phasing, pedigree data can also be used to significantly increase the accuracy of the imputation.

In this study, the accuracy of the haplotypes through BEAGLE, the pedigree samples were analysed, and each haplotype was checked through founders of the pedigree to children. In Figure 4.4, only six of the SNPs were listed for the representation of the haplotypes of the pedigree, which are rs780467580, rs182968886, rs1050501, rs6427615, rs72704062, rs844. rs182968886 and rs6427615 are GWAS SNPs (Chapter 5.2.2) used in this study (Appendix 16 for the detailed version of the haplotypes). rs1050501 is a clinically important variant that has been studied in the literature (Smith and Clatworthy, 2010, Chapter 5.2.1 for details of rs1050501). rs844 has also been used in this study to investigate allelic imbalance in gene expression of *FCGR2B* (Chapter 5.4.3). rs780467580 and rs72704062 were chosen to show how the haplotypes differ from each other.

Each sample has two haplotypes because generally everyone has two copies of each chromosome inherited from his/her parents. NA12892 only has one haplotype instead of two even though this sample was repeated twice by long-range PCR and Semi-conductor sequencing. It is considerable this sample is homozygous for this region. NA12883 seems to have the haplotypes same as his father NA12877 which is not possible. There are several SNPs that distinguish the haplotypes from each other. However, there is only one SNP difference between two haplotypes, rs72704062. This could be because of an individual mutation in NA12893. Instead of C at that position, there is G which make the NA12883 seem to have the NA12877 haplotypes (C<->G Transversion mutation). Even though transversion mutations are less likely to occur compared to transition mutations, this can also be explained by spontaneous mutations during the cell culture growing. If there was a C at that position, the haplotypes would be the same and the genotype would be homozygous for this region such happens in NA12892.

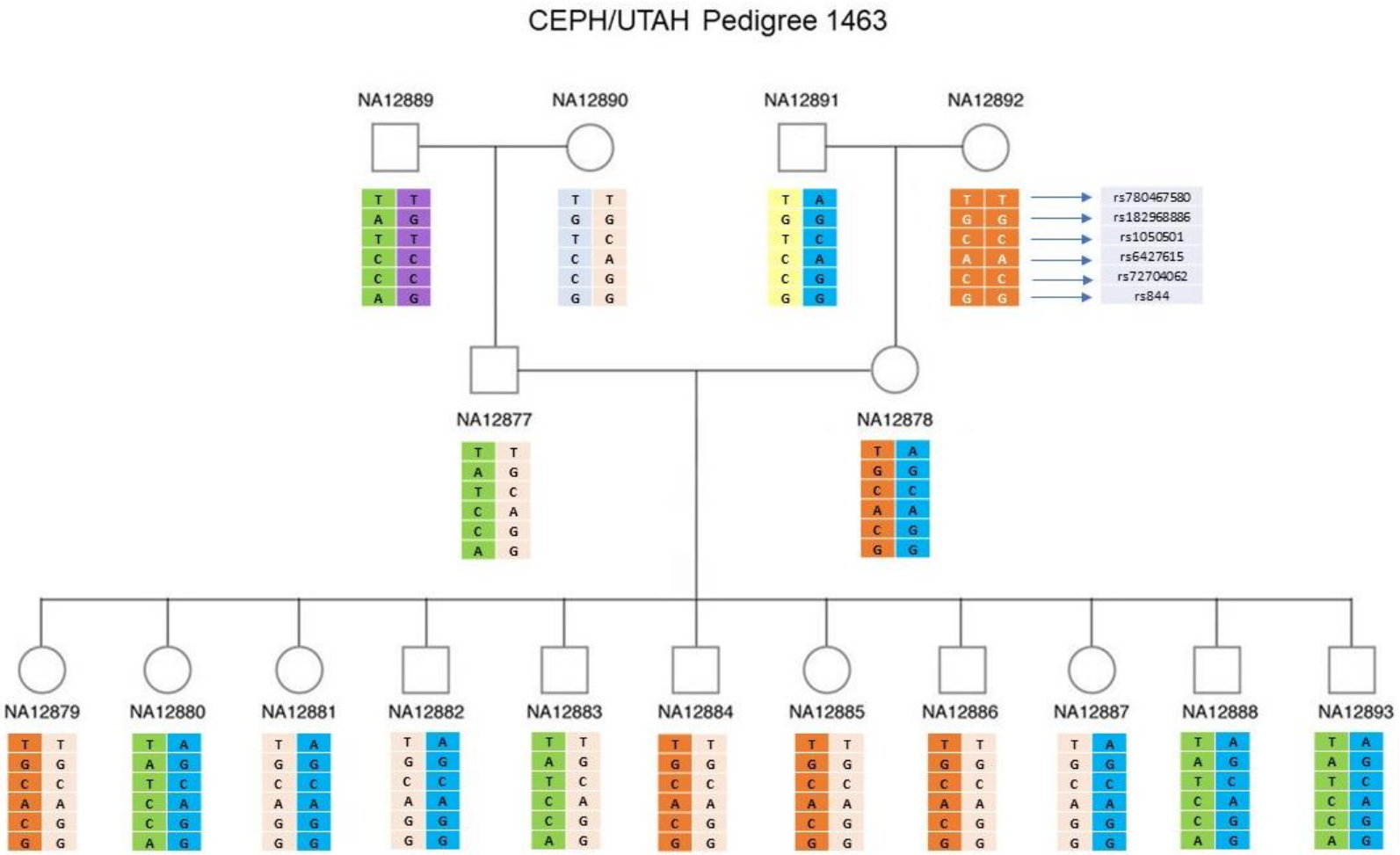


Figure 4. 4 The estimated haplotypes for the samples of the CEPH/UTAH pedigree 1463. Only six SNPs were used to represent the haplotypes. The colour of each column represents a haplotype.

4.4 Comparison of variants from different variant callers

4.4.1 Comparison of variants between Ion Torrent Platform and FreeBayes on diploid data

FreeBayes, a genetic variant detector, can discover SNPs, indels, Multiple nucleotide polymorphisms (MNPs), and complex substitutions and insertions shorter than a short-read sequencing alignment. A Bayesian statistical approach is generalized to allow multiallelic loci and non-uniform copy number across the samples (Garrison & Marth, 2012). The variant calling identifies and qualifies the differences between a dataset and the reference genome and provides an estimate of variant frequency and some measure of confidence.

FreeBayes was run on the bam files that were produced by Ion Torrent Platform to call the variants. These variants were compared to the variants obtained from Torrent Variant Caller (TVC) for the *FCGR2B* to test whether FreeBayes can identify the variants on diploid data. As seen in Figure 4.5, there are some differences between two variant callers regarding to the total number of variants. In most cases the number of variations is higher in Figure 4.5A. The variants are matching to TVC variant calls mostly single base calls or single base indels except some multiple nucleotides calls which were detected by FreeBayes. FreeBayes can successfully call the variants but it also detects the homopolymer calls as variants even though quality of these variant calling is very low. There are some reads showing there is a base insertion at that specific position on IGV, however, the count of the insertion is not enough to call it as a variant regarding the coverage. Therefore, the number of variant differences in Figure 4.5A is because variants from the FreeBayes includes the predicted homopolymer variants with low quality. Homopolymer errors are difficult to handle and it may cause false positive variant calls. In the process of the Ion semiconductor sequencing, multi nucleotides are incorporated and more hydrogen ions are released in a single cycle at the genomic loci with homopolymer repeats of the same nucleotide. TVC plays an important role in this repeat region due to its homopolymer error catcher and handles the low quality homopolymer calls. As seen in the 4.5B, the homopolymer variants were excluded. The number of the variants for FreeBayes also indicate the total number of variants which are the same for both variant callers.

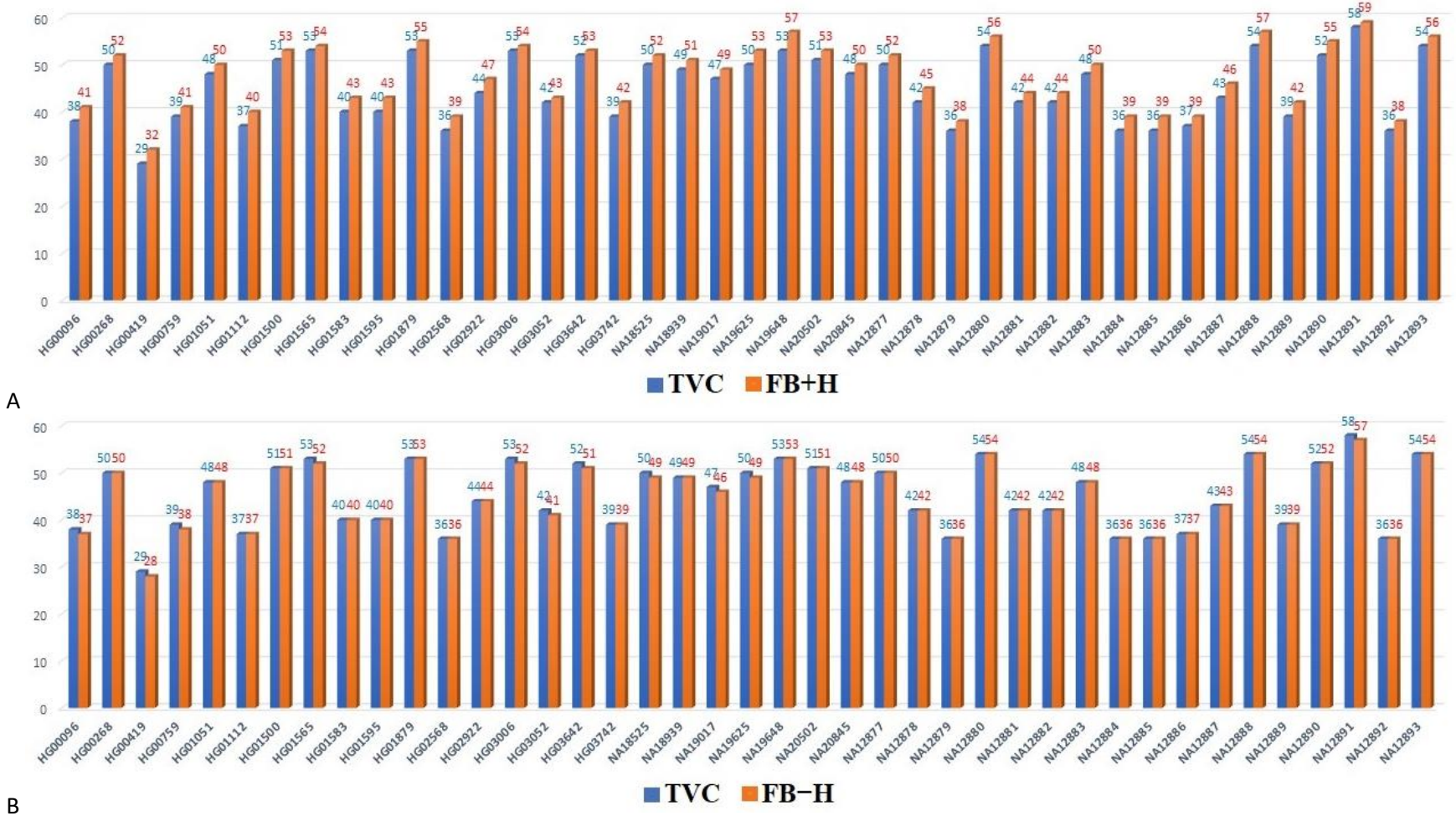


Figure 4. 5 Comparison of the total number of variants identified by TVC and FreeBayes. Total number of variants predicted from each variant caller is represented as vertical bars in the histogram. TVC: The total number of variant calls from Torrent Variant Caller. FB+H and FB-H: The total number of variant calls with homopolymer variant calls, without homopolymer variant calls, respectively.

In some samples, only 1 variant differs for the TVC calls which is also shown in the Figure 4.6. FreeBayes could not detect this variant in any sample which is an AAAAT deletion rs200504085 on dbSNP151 (rs200504085 was merged into rs3039548 on October 11, 2018 (Build 152)).

Figure 4. 6 The deletion AAAAT rs200504085 on dbSNP151. A: The heterozygous deletion for this variant on IGV for two samples; HG01565 and NA12891. B: The location of the variant on GRCh37/hg19 (chr1:161639365 -161639369). Other samples are HG00096, HG00419, HG00759, HG01565, HG03006, HG03052, HG03642, NA18525, NA19017 NA19625, NA12891.

Even though TVC and FreeBayes predict the same variants without homopolymer calls, there are some differences regarding how a variant is called (Table 4.1). While some variants were called as single base by TVC, FreeBayes makes a variant call including three variants together as a single call (Case 1). When there are two alternative variants exist, FreeBayes evaluates these two together at the same position. However, TVC makes two different calls where there is more than one alternative for the same position (Case 2). When there is a single base deletion/ insertion, FreeBayes shows this with a small group of bases together (Case 3). In conclusion, homopolymer variant calls or the quality of the calling variants needs to be visually inspected to avoid any false positive variant calling. These variants should be carefully filtered before further analysis for FreeBayes and the indel-realigner of the FreeBayes can be improved.

Table 4. 1 The cases where TVC and FreeBayes call variants differently.

Position		TVC				FreeBayes				
GRCh37/hg19	**PCR product**	**Ref**	**Variant**	**Quality**	**Allele Call**	**REF**	**ALT**	**Quality**	**Genotype (Unphased)**	**Example Sample**
Case 1:										
161643527	11698	A	C	38317.3	Homozygous	AAATGTA	CAGTGTG	53547.7	1/1	HG01583
161643529	11700	A	G	38469	Homozygous					
161643533	11704	A	G	38287.6	Homozygous					
161646112	14283	C	T	6193.03	Heterozygous	CGCATGTCTG	TGCACGTCCA	19914.2	0/1	HG01500
161646116	14287	T	C	5899.42	Heterozygous					
161646120	14291	TG	CA	5944.8	Heterozygous					
Case 2:										
161643025	11196	C	A	30874.6	Heterozygous	C	A, G	67323.4	1/2	HG02568
	11196	C	G	30874.6	Heterozygous					
161643546	11717	TG	CA	19191.4	Heterozygous	TG	CA, CG	41652	1/2	HG03742 HG00268 HG20502
	11717	T	C	19191.4	Heterozygous					
161644307	12478	GA	TG	17777.8	Heterozygous	GA	TG, GG	31187.9	1/2	NA19625
	12479	A	G	17777.8	Heterozygous					
Case 3:										
161646787	14959	C	-	4055.38	Heterozygous	ACCCCCCT	ACCCCCT	1514.63	0/1	NA20502
161646500	14672	-	C	5388.77	Heterozygous	AT	ACT	17268.8	0/1	HG02568

4.4.2 Comparison of variants between Ion Torrent Platform and FreeBayes on polyploid data

After testing Freebayes on the Ion Torrent sequenced data of *FCGR2B*, FreeBayes were used to call variants for the final mapping results of the BWA-MEM on ARR. This was done for the 24 of 1000 Genomes Project and 17 Platinum data samples for the region of *FCGR2B*. For the PCR products of *FCGR2B*, the estimated variants were identified based on a diploid sequence because PCR is diploid source. On the other hand, the variants from the mapping analyses were called based on copy numbers found from the PRT assay considering the copy number of the whole low affinity *FCGR2B* region which can be 3, 4, 5 and 6. FreeBayes is a useful tool, which provides easy set up and run on haploid and polyploid genome so that it can be run with copy number higher than 2. Even though the human genome is not polyploid, our strategy is to apply the polyploid variant calling approach on duplicated regions. The reads that are mapped to ARR/RR come from multiple copies of the region and therefore there will be more variants, as fixed variants between duplicated sequences will also be called. Here, the variant calling is made on ARR because the *FCGR2B* is located on the paralogue B of the segmentally duplicated region of the low-affinity FcgRs. The expectation is the variant calls predicted from PCR source should also be predicted by Freebayes variant call of the BWA-MEM mapping on ARR. Thus, the number of variants found from the mapping analyses with ARR should be much higher than the number of variants of Ion semiconductor sequencing on the same sample (Figure 4.7) because variant predictions was made based on the copy number variability of the region. However, not all the variants from the PCR source were predicted with the variants calls from the mapping analyses. Two examples of the samples from each data group is shown in Figure 4.8. In addition, there are some samples where even the number of variants found for the mapped samples is lower, for example HG02568 (Figure 4.8).

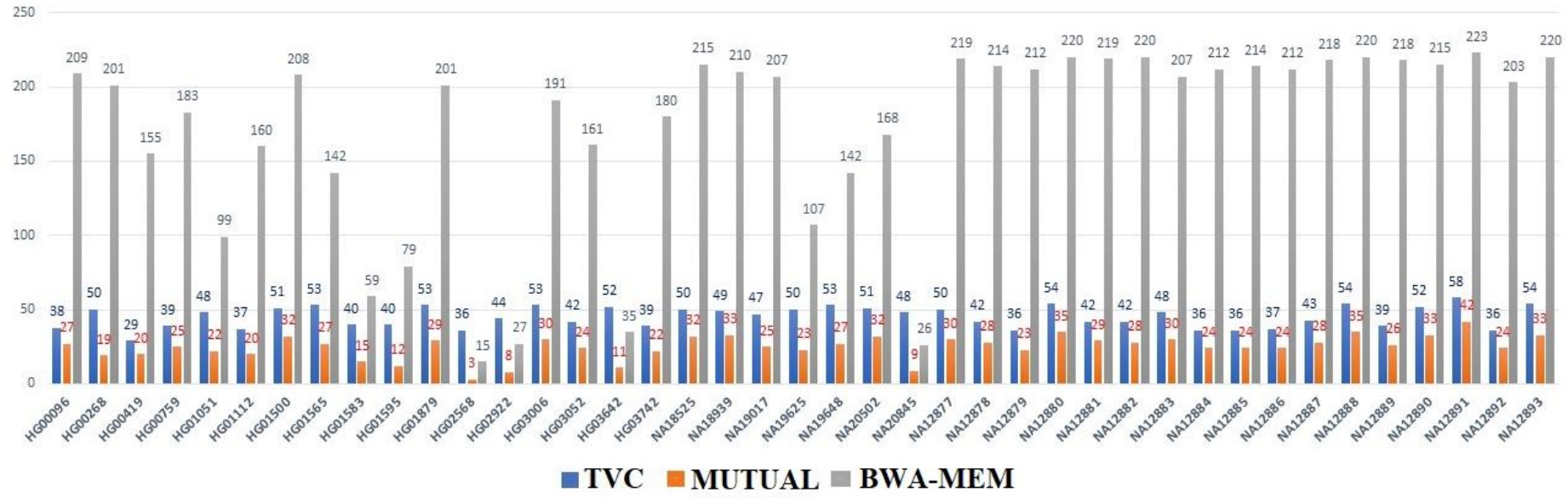


Figure 4. 7 The comparison of the total number of the variants between TVC on PCR product and FreeBayes on mapped samples. TVC: Ion torrent variant caller, Mutual: the variants are the same for both variant callers, BWA-MEM: FreeBayes variant call on mapped samples.

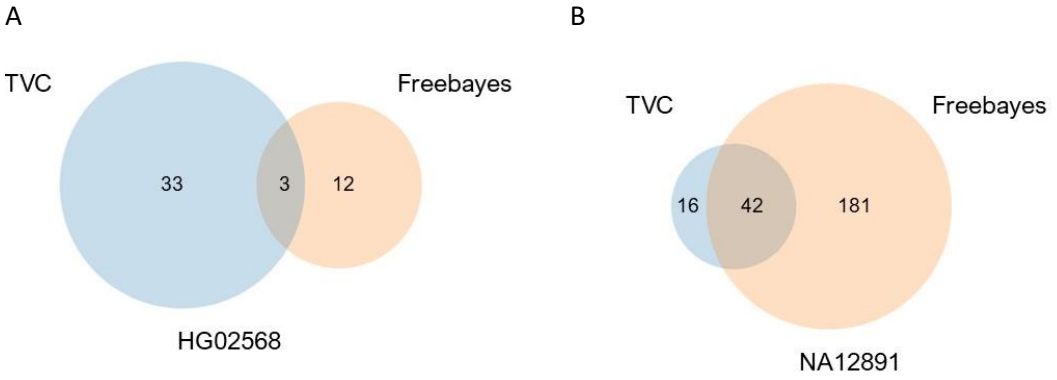


Figure 4. 8 The comparison of variants from different sources on the same two sample. The lowest(A) and the highest (B) variant match between two sources is shown. TVC: Torrent Variant Caller on the PCR product of *FCGR2B* locus. FreeBayes: FreeBayes variant call on the mapped sample with BWA-MEM of ARR reference.

In this case, first, the quality of the reads can be still low even they were filtered as FASTQ files before mapped to the references. Second, the low quality may result in low coverage of the mapped data which decreases the confidence of the variant calling. Third, because the reduced references used in this study are masked due to the repeats before the mapping analyses, neither mapping nor variant calling can be performed for these masked regions in the reduced references (Figure 4.9). Therefore, FreeBayes cannot identify all the variants.

Figure 4. 9 A screenshot of two samples mapped against ARR on IGV with gaps. The samples HG00096 (top) and HG00419 (bottom) are shown on IGV. In both samples, the gaps between mapped regions corresponds to the repeat masked region of the alternative reduce refence.

4.4.3 Comparison of variants among Ion Torrent Platform, FreeBayes and the phase 3 of the 1000 Genomes project on polyploid data

The 1000 Genomes project performed variant calls so that genotypes for a specific individual or population from VCF files can be obtained. These variants are listed in the Essembl Data Slicer which provides an interface to get subsections of either VCF or BAM files based on genomic coordinates. The 24 of 1000 Genomes Project samples were searched on the Data Slicer for the region chr1:161632905-161648444 where *FCGR2B* (15540 bp) is located. Only five SNPs were found for this region in the phase 3 of 1000 Genomes project as listed in Table 4.2.

Table 4. 2 The list of SNPs found by phase 3 of the 1000 Genomes project.

Chromosome	Position **(GRCh37/hg19)**	ID	REF	ALT
1	161633506	rs556421297	G	A
1	161633627	rs560795289	A-	AT
1	161633638	rs576163081	G	A
1	161633774	rs1832738	G	T
1	161634890	rs562055782	G	A

Only the rs1832738 was found to be heterozygous or homozygous alternative for some of the samples. The rest of the SNPs were found to be homozygous reference. The variants found by Ion semiconductor sequencing, variants found by FreeBayes from the mapped samples and the variants found by 1000 Genomes project for the *FCGR2B* region were compared to each other. The comparison of the rs1832738 genotype calls is listed in Table 4.3. There are 10 samples found to be either heterozygous or homozygous for the alternative variant of rs1832738 (161633774G>T) for the 1000 Genomes project samples. Only six of the samples from the Ion semiconductor sequencing and only 3 samples from the BWA-MEM mapping have this variant. From these 6 samples, two of them (HG01879 and HG03052) are shown on IGV in Figure 4.10. While variants of rs1832738 was found on HG01879, it is not found on HG03052 as it happens for the other four samples. Sample HG01879 were found to have one of the variations of this SNP from the Ion semiconductor sequencing and BWA-MEM mapping results but not 1000 Genomes project.

Table 4. 3 The detection of rs1832738 variation from different approaches.

Sample	1000 Genomes Project (Phase 3)	Ion Torrent	BWA-MEM ARR
HG00096	+	+	+
HG00419	+	-	-
HG00759	+	-	-
HG01595	+	+	-
HG01879	-	+	+
HG02568	+	+	-
HG02922	+	+	-
HG03052	+	-	-
HG03742	+	-	-
NA18525	+	-	-
NA20845	+	+	+

The whole genome sequencing of the 1000 Genomes project samples was performed with 7-8X coverage in the phase 3. However, this study performed the mapping against reduced reference with higher coverage for more of the samples (Appendix 5). While some samples also have 7-8X coverages, some samples have 30-32X coverage which increases the possibility of predicting the variants more accurately. The 1000 genome project found only five SNPs for the region where *FCGR2B* is located. However, this study found more SNPs/ variants (minimum 29 (HG00419), maximum 53 (NA19640)) for the 1000 genomes samples. This could be because the sequence reads may map to several places when there is more than one copy of a particular region so that this may prevent the identification of the variants with high confidence.

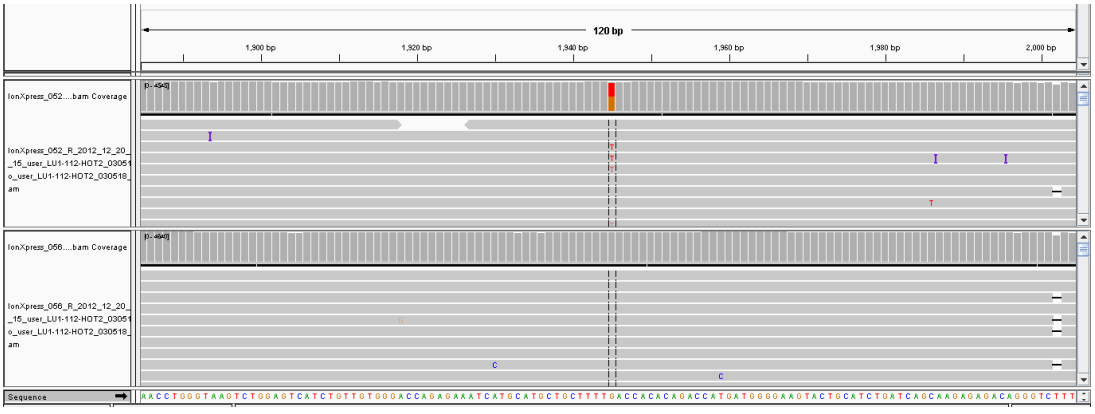


Figure 4. 10 A screenshot of rs1832738 SNP visualization on IGV.

The samples HG01879(top) and HG03052(bottom) are shown on IGV. While HG01879 is heterozygous for the rs1832738, HG03052 does not show any variation at this SNP.

4.4 Haplotype estimation of *FCGR2B* locus polyploid data

HapCompass is for haplotype assembly of densely sequenced human genome data. Its algorithm operates on a graph where SNPs are nodes and edges are defined by sequence reads and viewed as supporting evidence of co-occurring SNP alleles in a haplotype. This program does not make any prior assumptions and the advantage of this program is that it is applicable to polypoid genomes (Aguiar and Istrail, 2012). As a haplotype assembly tool, HapCompass was used in this study to assign haplotypes to different copies of the gene clusters. However, the haplotypes constructed from the mapped sample for FCGRs region did not match the haplotypes constructed from the Ion semiconductor sequenced sampled of *FCGR2B* locus as indicated for sample HG00096 in Table 4.4. This figure only represents some of the data for HG00096. The rest of the data similar to these results as well as other samples used in this study. The table shows the variants on diploid PCR products (called by FreeBayes) with BEAGLE phasing results and shows the variants on the mapped data with HapCompass phasing. Some of the variants are missing for the mapped data and consequently phasing is not possible for the missing variants. All these missing variant calls are in the regions where gaps exit because of masked repeats. This may cause the missing information of these region so that the haplotype assembly program cannot have all the variant information to construct the haplotypes accurately.

Table 4. 4 Comparison of variant calls and phasing approach by BEAGLE and HapCompass (HG00096).

Position		PCR Product						Mapped sample							
GRCh37/hg19	PCR product	Ref	Variant	Allele Call	BEAGLE phasing	B1	B2	Position on ARR	Ref	Variant	HapCompass phasing	H1	H2	H3	H4
161633277	1448	C	A	Heterozygous	0|1	C	A	381638	C	A	1|1|0|0	A	A	C	C
161633527	1698	T	C	Heterozygous	0|1	T	C	381888	T	C	1|1|0|0	C	C	T	T
161633774	1945	G	T	Heterozygous	0|1	G	T	382135	G	T	0|0|1|1	G	G	T	T
161634189	2360	A	G	Heterozygous	0|1	A	G	382550	A	G	0|0|1|0	A	A	G	A
161634945	3116	T	C	Homozygous	1|1	C	C	383306	-	-	-	-	-	-	-
161635125	3296	T	C	Heterozygous	0|1	T	C	383486	-	-	-	-	-	-	-
161637175	5346	G	A	Homozygous	1|1	A	A	385536	-	-	-	-	-	-	-
161637192	5363	A	T	Homozygous	1|1	T	T	385553	-	-	-	-	-	-	-
161637290	5461	C	T	Homozygous	1|1	T	T	385651	-	-	-	-	-	-	-
161638043	6214	C	T	Heterozygous	1|0	T	C	386404	C	T	1|1|0|0	T	T	C	C
161638268	6439	A	T	Heterozygous	0|1	A	T	386629	A	T	1|0|0|0	T	A	A	A
161638411	6582	T	A	Heterozygous	1|0	A	T	386772	T	A	1|0|0|0	A	T	T	T

4.5 Discussion

In this chapter, the variants from the Ion semiconductor sequenced samples of *FCGR2B* locus were successfully detected. Then, FreeBayes was used to call variants from the mapping analyses for the same samples. The variants from the PCR source were compared to the variants from the mapping analyses. The expectation was the variant predicted from PCR source of *FCGR2B* should also be predicted by variants of the BWA-MEM mapping on ARR. However, not all the variants from the PCR source were predicted with the variants calls from the mapping analyses that may be due to poor quality of the reads even they are filtered by sequence quality, the loss of the data during the quality filtering (the low percentages of the data left trimming) , relative low coverage after the mapping and BAM refinement as well as the masked reduced reference regions. Furthermore, the differences between the variant's calls can also be because the NGS platforms perform sequencing in different ways.

The haplotype construction of the diploid samples was effectively achieved as seen in the pedigree data (Figure 4.4) by BEAGLE. However, construction of the haplotypes of the mapped samples was not successfully constructed by HapCompass (Table 4.4). HapCompass can be used for polyploid genomes to create accurate pairwise SNP phasing. This software has been applied to 1000 Genomes data simulations to highlight the type of data needed to supplement existing 1000 Genomes Project data to completely phase a chromosome. It has been compared to other two well-known haplotype assembly software packages that can also process arbitrary input sequence data: HapCut and the GATK's read-backed phasing algorithm. HapCompass has been shown to be faster and significantly more accurate than HapCut and GATK using a variety of metrics on real and simulated data (Aguiar and Istrail, 2012). As mentioned previously, the reduced references were masked by repeats which also harbour variants. Thus, lack of full list of variants of a region can cause inaccurate haplotype phasing. As conclusion, mapping short reads to reduced reference is limited.

Overall, the variants were successfully called by using Ion Torrent sequencing and haplotype construction was built by BEAGLE on PCR product of *FCGR2B.* However, not

all the variants from the mapping analyses could predict the variants from PCR source and haplotype construction on the mapped samples were unsuccessful. In future chapters, the results from the PCR products of *FCGR2B* will be used.

CHAPTER 5 Prediction of the functional consequences of variation of the *FCGR2B* gene

5.1 Introduction and study rationale

FcgRs are important for receptor signalling as a family of glycoproteins as a part of the immunoglobulin superfamily (IgSF). These receptors have an IgG binding α-subunit that binds to the Fc domain of IgG which triggers different type of signalling pathway (Rosales and Uribe-Querol, 2013). Linking of the these receptors with the Fc fragment of IgG antibodies activates various functions in the immune system cells such as phagocytosis, cell degranulation, production of various cytokines and chemokines, and activation of genes depending on the cell type (Hargraves et al., 2015).

Three classes of Fcg receptors have been recognized, FcgRI, FcgRII, FcgRIII, in human, and a fourth class, FcgRIV, in mice (Nimmerjahn and Ravetch, 2008). FcgRI (CD64), FcgRIIa (CD32A), FcgRIIb (CD32B), FcgRIIc (CD32C), FcgRIIIa (CD16A), FcgRIIIb (CD16B) are the six Fcg receptors expressed by humans (Bruhns et al., 2009; Hargreaves et al., 2015). These receptors are found on various cells of the immune system and recognize different subclasses of IgG antibodies because they have different affinity for the antibody IgG Fc-fragment. While FcgRI is the only high-affinity receptor, binding monomeric IgG molecules, and FcgRIIa, FcgRIIb, FcgRIIc, FcgRIIIa, and FcgRIIIb are the low-affinity receptors, binding multimeric immune complexes (Hargraves et al., 2015).

FcgRs can be divided functionally into two categories; the activatory receptors that use immunoreceptor tyrosine-based activation motifs (ITAMs) and inhibitor receptor that uses immunoreceptor tyrosine-based inhibition motifs (ITIMs) for creating transmembrane signalling. The FcgRIIa, FcgRIIc, FcgRIIIa are activatory receptors using ITAMs. FcgRIIIb is an activating receptor, bonding to the cell membrane via a glycosylphosphatidylinositol (GPI) anchor instead of using ITAM because of an arginine-to-stop mutation in the *FCGR2B* gene (Ravetch and Perussia, 1989; Machado et al., 2012). FcgRIIb is the only known inhibitory FcgR, containing ITIM. (Ravetch and Lanier, 2000; Nimmerjahn and Ravetch, 2006; Getahun and Cambiesr, 2015). The inhibitory

FcgRIIb prevents antibody production in B cells by down regulating the activation signals from the B cell antigen receptors (BCR) which also have ITAM motifs (Stefanescu et al., 2004). FcgRIIb, containing ITIM, signals by activating the phosphates contrary to the activating receptors which engage kinases. FcgRIIb also creates a threshold for the cell activation to start the cell functions in the phagocytic leukocytes and dendritic cells by working together with activating Fcg receptors (Boruchov et al., 2005; Nimmerjahn and Ravetch, 2006; Rosales and Uribe-Querol, 2013). FcgRIIb inhibits signaling of ITAM-associated receptors by recruitment of phosphotyrosine phosphatases that dephosphorylate the tyrosine residues in several ITAM-induced activation pathway effectors. The inhibitory FcgRIIb exists in two splice variants, namely, FcgRIIb1 (which is exclusively expressed by B cells) and FcgRIIb2 (expressed by B cells, macrophages, and dendritic cells) (Ravetch and Lanier, 2000, van der Heijden, et al., 2012).

The function of FcgRs are crucial for effective control of inflammation and response to infections. In addition to regulating native immune responses, they are thought to be key to the therapeutic monoclonal antibodies (mAbs) (Nimmerjahn and Ravetch, 2015; Nimmerjahn and Gordan, 2015). These receptors are implicated in the genetic susceptibility to autoimmune diseases (Vogelpoel et al., 2015; Hargreaves et al., 2015). Differences in receptor expression levels on the cell surface can alter the balance between activating and inhibitory signals and, therefore, change the cellular response to IgG. Genetic variation within the low-affinity receptors has been associated with different diseases (Breunis et al., 2008). Both sequences and CNV of these genes have been shown to affect function and have been associated with autoimmune disease (Fanciulli et al., 2008; Niederer et al., 2010; Machado et al., 2012). Several SNPs were identified in the promoter and coding regions in the case of polymorphism of *FCGR2B* gene. Previous studies have shown that SNPs are found to be clinically important for malaria and SLE diseases (Blank et al, 2015). Therefore, it is crucial to investigate genetic variation in this gene cluster because that may contribute for the discovery of disease risk factors, therapeutic efficacy of biological agents, and the development of new treatment strategies.

To understanding the variation of the FCGR region, Rahbari, et al. (2016) investigated the role of gene conversion and identified new gene conversion alleles by reanalysing fosmid sequences previously generated for eight individuals using 454 sequencing technology (Kidd et al, 2008; Machado et al., 2012; Mueller et al 2012; Nagelkerke et al., 2015). In the study of Rahbari, et al. (2016), these fosmid sequences were used for *de novo* assembly to generate contigs for each individual. Paralogue A (chr1:159815745-159831746) and paralogue B (chr1:159897573-159913518) sequences were extracted from the human reference genome hg18. Each contig was then mapped to these paralogue regions by using BWA-MEM mapping (alignments spanning less than 400 bases on the reference were discarded). The fraction of matching aligned bases (sequence identity) between the individual and reference was compute using the aligned contigs which was then assigned to the paralog of the highest sequence identity. Then, this sequence identity was used to obtain a single sequence representing the fosmid sequences for each individual. This single sequence was used to investigate potential switches between A-like and B-like sequences to determine the gene conversion events. Using this technique, three different alleles were initially found: an A to B gene conversion of about 9.1 kb, a reciprocal B to A gene conversion of 9.1 kb and a smaller B to A gene conversion of 2.4 kb (Figure 5.1, only shows B to A gene conversions are shown in detail). All three gene conversions occur between FCGR3A/B and FCGR2B/C in the region previously identified as a deletion breakpoint region (Machado et al., 2012; Mueller et al., 2012; Nagelkerke et al., 2015). As seen in Figure 5.1B, the consensus fosmid sequence is aligned to paralogue A and B based on hg18 genome assembly coordinates. The numbers above and below of the midline between paralogues (A and B) indicates the number of mismatches to A and B, respectively. For example, the sample NA18956 has a gene conversion of about 2.4kb. In this B to A region there are 8 and 34 mismatches for paralogue A and paralogue B, respectively. This means the consensus fosmid sequence is more similar to paralogue B rather than A. B. Therefore, this is concluded as B to A gene conversion.

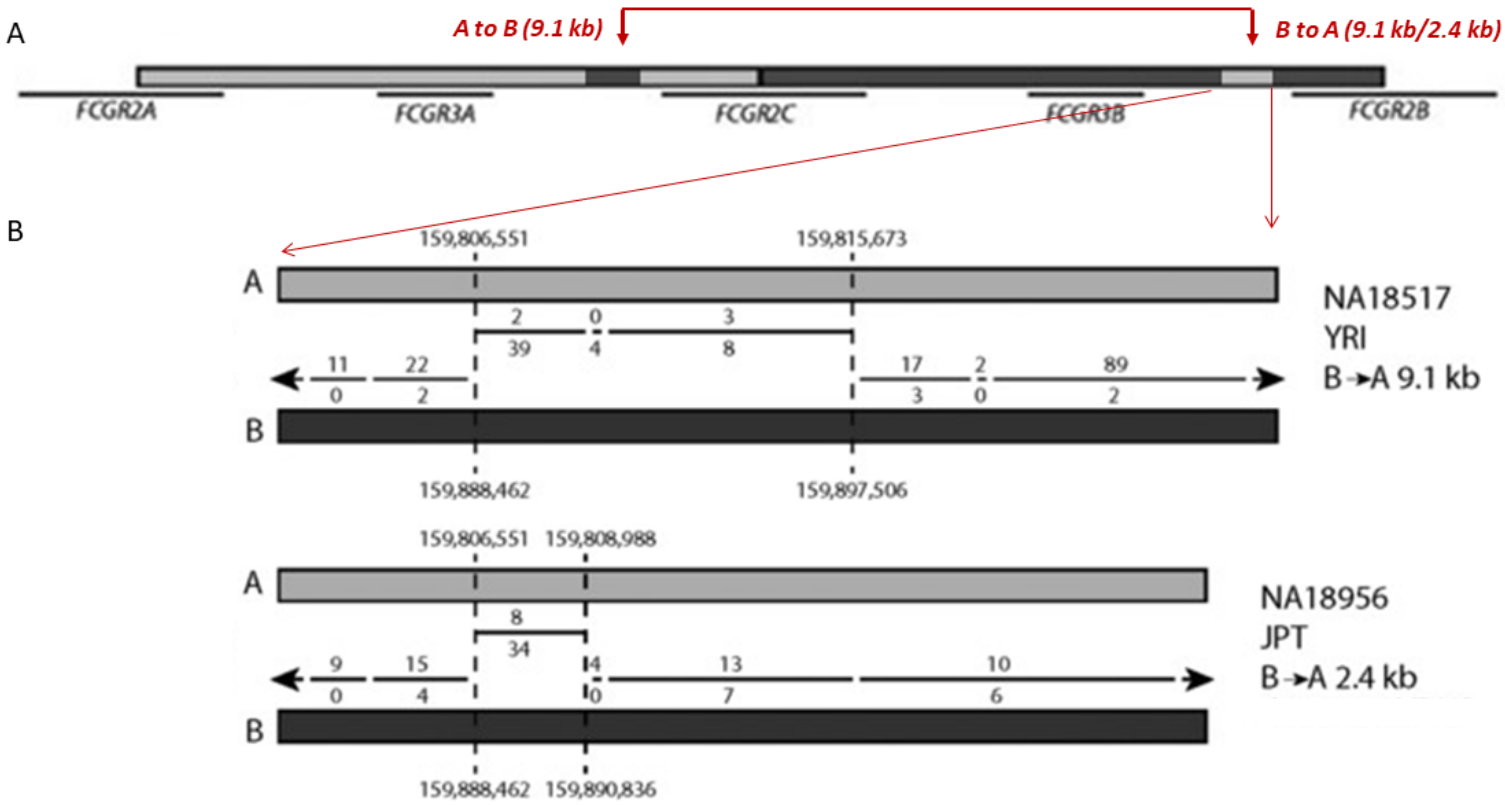


Figure 5. 1 Previously identified gene conversion events. A: the approximal location of the gene conversion events in the paralogue A and B regions are shown. B: shows the reciprocal events; a 9.1-kb and a 2.4 kb gene conversion on fosmid sequence of two different samples based on hg18 genome assembly. A and B are paralogous/references of FCGRs locus. The middle line, with numbers above and below, between paralogues (A and B) indicates a single sequence representing the fosmid sequence that is aligned to both A and B reference sequences. The number of mismatches to A and B shown above and below the line, respectively. Adapted from Rahbari et al., 2016.

This chapter has two parts separately. In the first part, I aim to create a good indication of variant list for the *FCGR2B* gene, to inspect the consequences of these variants, to looked for any variants which variants might explain any GWAS hits. Despite the diverse functions and cellular expression of the different FcgRs, there is high sequence similarity between the genes. The presence of high sequence homology along with the existence of known segmental duplication and extensive single-nucleotide variation within and between the duplicated paralogs, and copy-number variation can make the FcgRs locus and FCGR2B gene challenging to genotype and can cause misinterpretation of the consequences. Therefore, it is particularly challenging but interesting to examine the variants which are specifically associated with to *FCGR2B* which encodes the only inhibitory receptor of FcgRs.

In the second part, I aimed to design a PCR-based assay to confirm the recently identified gene conversions in the upstream of the *FCGR2B* gene, and to investigate the role of gene conversion in this region, and to differentiate the alleles for the gene conversion events. I also aimed to to look for any association between gene conversion and GWAS SNPs, and to analyse whether the gene conversion influences the expression of the *FCGR2B* gene. This is important because gene conversions may contribute to variation in gene expression at the transcript and protein level from functional perspective. In response, altered FcgRs levels can change the functional capacity in response to activation by IgG complexes for a given cell.

5.2 Predicted functional consequences of *FCGR2B*

5.2.1 Analysis of predicted functional consequences using Variant Effect Predictor

The Ensembl Variant Effect Predictor (VEP) is a web-based tool that annotates variants using a wide range of reference data, including transcripts, regulatory regions, frequencies from previously observed variants, citations, clinical significance information, and predictions of biophysical consequences of variants (McLaren et al., 2016). VEP shows the consequences of the variants such as the genomic location of the novel/existing variants, identifiers for the variants, the consequence types of the

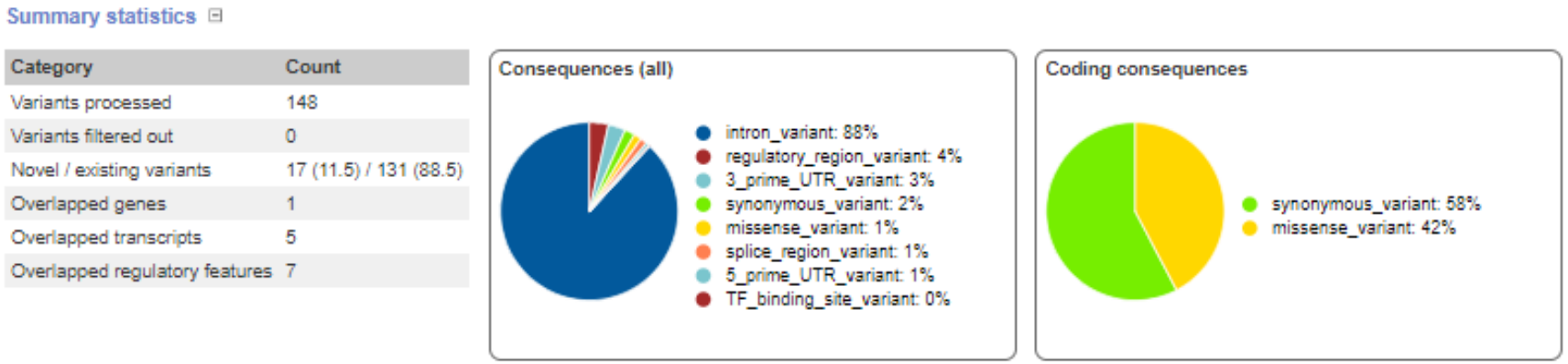
variants, the allele frequency (AF) data for existing variants from several major genotyping projects, the 1000 Genomes Project, the NHLBI-ESP (NHLBI GO Exome Sequencing Project) and gnomAD (Genome Aggregation Database). VEP searches several databases such as the Ensembl Variation databases that contain a large catalogue of freely available germline and somatic variation data in vertebrates (Chen et al., 2010b; Rios et al., 2010). Additional human data include mutations from COSMIC (Forbes et al., 2011) and the Human Gene Mutation Database (Stenson et al., 2012) and structural variants by VEP and copy number variants from the Database of Genomic Variants archive (Lappalainen et al., 2013) are also included in the VEP. Therefore, the VEP can reference millions of variants to identify those previously described (McLaren et al., 2016).

For each variant that is mapped to the reference genome, all the overlapping Essembl transcripts are identified. A set of consequence terms and given impact, defined by Sequence Ontology, are assigned to each combination of allele and transcript. VEP provides two different impact scores PolyPhen (polymorphism phenotyping) and SIFT (sorting intolerant from tolerant) that are used to predict whether an amino acid substitution affects protein function. These scores use the same range, 0.0 to 1.0, but have contrary meanings. While a variant with a PolyPhen score of 0.0 is predicted to be benign, a variant with a SIFT score of 0.0 is predicted to be deleterious. The output of VEP starts with a pie chart that shows the proposition of consequences type called across all the variants. It also includes a result table that shows one row per transcript and variant. The column titles can be location, allele, consequence, impact, gene, feature, feature type, biotype, existing variation, distance, strand, HGVS identifiers, with any chosen extras (Figure 5.2). The result page can be navigated, and a combination of filters can be applied, and new filters can be added manually.

All the variants found for *FCGR2B* region (Chapter 4.2, Appendix 13) from the Ion Torrent sequencing experiment were downloaded to the interface of the VEP. According to the summary statistics, there are 148 variants processed (Figure 5.2A). 17 (11.5%) novel variants were not found and 131 (88.5%) existing variants were found

(Appendix 12 and 13). Most of the variants were found in introns (88%). For the variants of the coding regions, the synonymous variants (58%) were found more than the missense variants (42%). An impact of a variant can be high (disruptive impact in the protein, protein truncation or loss of function), moderate (non-disruptive variant that might change protein effectiveness), low (harmless, unlikely to change protein behaviour), or modifier (usually non-coding variants). While 142 of the total variants have modifier impact, four variants have low impacts. These are rs6665610, rs2298022, rs182968886 as synonymous variants, and rs2125684 as a splice region variant. No variants were found with high impact. Only three variants were found to have moderate impact as missense variants; rs1050501, rs148534844, and NC_000001.10:g.161645052C>T. The rs1050501, and NC_000001.10:g.161645052C>T were predicted by SIFT as tolerated (>0.8) and PolyPhen as benign (<0.25). None of the variants were found to be 'deleterious' for SIFT score nor 'probably damaging' for PolyPhen score.

A



B

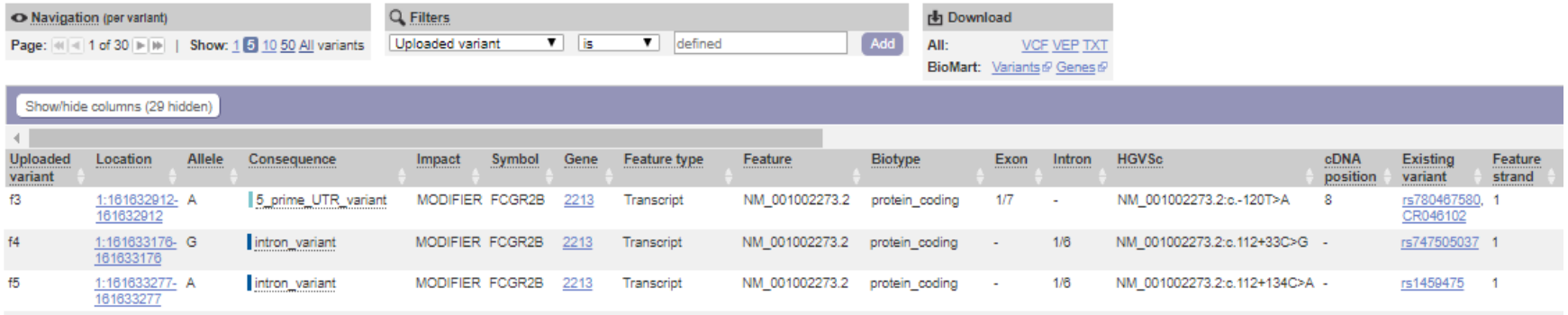


Figure 5. 2 The predicted effects of the variants in *FCGR2B* gene. A: A screenshot of the VEP results summary statistics. B: The preview of the results table with navigation, filtering, and download options.

The SNP rs1050501 is the only variant predicted to be as a risk factor or leading to a protective function as clinical significance states assigned by ClinVar (Landrum et al., 2014). The substitution happens at the position c.695T>C in the *FCGR2B* exon 5 due to a nonsynonymous change, an isoleucine (I) to threonine (T) substitution (FcgRIIBT232) in the transmembrane domain which may modify the function of FcgRIIb in B cells (Li et al., 2009). This single amino acid substitution disturbs the inhibitory function of the receptor on B cells. Given the crucial roles of lipid rafts in integrating BCR signalling, reduced association of FcgRIIBT232 could decrease the inhibitory potential towards BCR signalling. Accordingly, the FcgRIIBT232 substitution may affect the localization and function of FcgRIIB, and the molecular mechanism may be associated with the polymorphism and susceptibility to SLE a multisystem autoimmune disease that can affect many organs, including the skin, joints, the central nervous system and the kidneys (Kono, et al., 2005; Floto et al., 2005). The FcgRIIBT232 allele of the inhibitory receptor FcgRIIB is found to be high frequency in African and Asian populations, mostly corresponding to areas where malaria is endemic. Malaria is a life-threatening disease caused by protozoan parasites of the Plasmodium genus which are most commonly spread to people or animal through infected female Anopheles mosquitoes (Snow et al., 2005; Langhorne et al., 2008). The FcgRIIBT232 was also reported as a risk factor in other disease such as rheumatoid arthritis (RA) and anti-glomerular basement membrane (anti-GBM) disease (Radstake et al., 2006; Willcocks et al., 2010; Zhou et al., 2010; Bonatti et al., 2017).

5.2.2 Linkage disequilibrium of the GWAS SNPs of *FCGR2B*

GWAS (Genome-wide association study) is a study of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a disease and/ or other phenotypes, thus, aims to find the genetic variation that contributes or explains complex diseases. It focuses on associations between SNPs and traits. The goal of GWAS is to use genetic risk factors to make predictions about who is at risk and to identify the biological foundations of disease susceptibility for developing new prevention and treatment strategies (Bush and Moore, 2012). There are eight SNPS listed in the GWAS catalogue for the *FCGR2B* gene. Five SNPs were found in the dataset used in this study (rs17413015, rs6665610, rs72480273, rs6427615, rs182968886) (Table 5.1).

Table 5. 1 The list of GWAS SNPs with associations.

GWAS Association	Allele	GWAS SNP ID	RAF*	Study accession
Blood protein measurement	A	rs6665610	0.205	GCST005806
Low affinity immunoglobulin gamma Fc region receptor II-a/b measurement	?		NR	GCST004365
Monocyte percentage of white cells	A	rs182968886	0.116	GCST004609
Low affinity immunoglobulin gamma Fc region receptor II-a/b measurement	T	rs17413015	0.147	GCST004365
Birth weight	C	rs72480273	0.17	GCST005146
Blood protein measurement	?	rs6427615	NR	GCST005989
*Risk Allele Frequency				

These five SNPs were first used to investigate the LD among them. The consequence and predicted impact results from the VEP analysis were also included. According to Table 5.2, the rs17413015 and rs72480273, both were found to be intronic variants with modifier effect, have a complete LD as also seen on Figure 5.3 and Figure 5.4 where total number of 52 SNP haplotypes was detected (Chapter 6.2.1). The rs6665610, synonymous variant with a low impact, has relatively strong LD with the rs17413015 and rs72480273. The other GWAS SNPs seem to have very weak LD among each other.

Table 5. 2 LD between GWAS SNPs used in this study.

Position 1	Consequence	Predicted Impact	Position 2	R^2	|D'|
rs6665610	Synonymous variant	Low	rs182968886	0.01	1
			rs17413015	0.712	0.844
			rs72480273	0.712	0.84
			rs6427615	0.077	1
rs182968886	Synonymous variant	Low	rs17413015	0.01	1
			rs72480273	0.01	1
			rs6427615	0.089	1
rs17413015	Intron variant	Modifier	**rs72480273**	**1**	**1**
			rs6427615	0.077	1
rs72480273	Intron variant	Modifier	rs6427615	0.077	1
rs6427615	Intron variant	Modifier			

The GWAS SNPs were also examined with the other *FCGR2B* locus variants found in this study. Several SNPs were found to be relatively in strong LD with GWAS SNPs as listed in Table 5.3. The rs17413015 has a complete LD with other two SNPs rs12116547, rs1214584. The rs2480273, associated with birth weight, is also in complete LD with the same SNPs rs12116547, rs12145843. The rs17413015 has also complete LD with the variant 161644663.C>T. Thus, the rs17413015 and rs2480273 are in complete LD with the same SNPs (Figure 5.3).

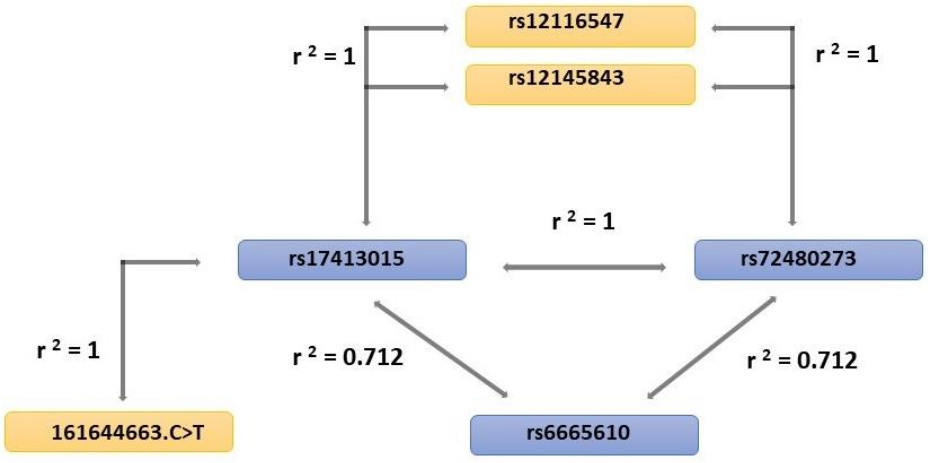


Figure 5. 3 The relationship of the GWAS SNPs and other SNPs used in this study by LD statistics. Blue rectangles: GWAS SNPs; Orange rectangles: SNPS in this study.

Table 5. 3 The list of GWAS SNPs and LD with other SNPs found in the study. (R^2 is over 0.7)

GWAS SNP	LD statistics in this study				
	SNP of *FCGR2B*	**Consequence**	**Predicted Impact**	**R^2**	**|D'|**
rs6665610	rs12116547	Intron variant	Modifier	0.712	0.844
	rs12145843			0.712	0.844
	rs12145988			0.846	1
	rs7519636			0.846	1
	rs12117530			0.846	1
	161644663.C>T			0.712	0.844
rs182968886	rs10917643	Intron variant	Modifier	0.730	1
	rs6661332			0.742	0.861
	rs6658657			0.730	1
	rs12145090			0.730	1
	rs1344438491			0.730	1
rs17413015	**rs12116547**	Intron variant	Modifier	**1**	**1**
	rs12145843			**1**	**1**
	rs12145988			0.846	1
	rs7519636			0.846	1
	rs12117530			0.846	1
	161644663.C>T			**1**	**1**
Rs72480273	**rs12116547**	Intron variant	Modifier	**1**	**1**
	rs12145843			**1**	**1**
	rs12145988			0.846	1
	rs7519636			0.846	1
	rs12117530			0.846	1
rs6427615	rs10465555	Intron variant	Modifier	0.863	1
	rs60081850			0.858	0.949
	rs7539744			0.858	0.949
	rs1984769			0.73	0.897
	rs2045571			0.859	0.95
	rs1050501	**Missense variant**	**Moderate**	**0.859**	**0.95**
	rs7532925	Intron variant	Modifier	0.861	1
	rs13376485			0.952	1
	rs3767641			0.952	1
	rs3767640			0.952	1

The clinically important SNP rs1050501 is found to be in strong LD with only rs6427615 among GWAS SNPs (Table 5.3). The rs1050501 has also in relatively high LD with the SNPs listed for the rs6427615 (Table 5.4) Therefore, all the SNPs in this group can be in strong LD.

Table 5. 4 The linkage disequilibrium of SNP rs1050501 with other SNPs of *FCGR2B*.

SNP	LD statistics			
	SNP of *FCGR2B*	**Consequence and Predicted Impact**	**R^2**	**|D'|**
rs1050501	rs10465555	Intron variant, Modifier impact	0.730	0.897
	rs60081850	"	0.906	1
(Missense Variant, Moderate impact)	rs7539744	"	0.906	1
	rs1984769	"	0.770	0.899
	rs2045571	"	1	1
	rs7532925	"	0.820	1
	rs6427615*	"	0.859	0.950
	rs13376485	"	0.904	0.950
	rs3767641	"	0.904	0.950
	rs3767640	"	0.904	0.950
(R^2 is over 0.7) *(GWAS SNP)				

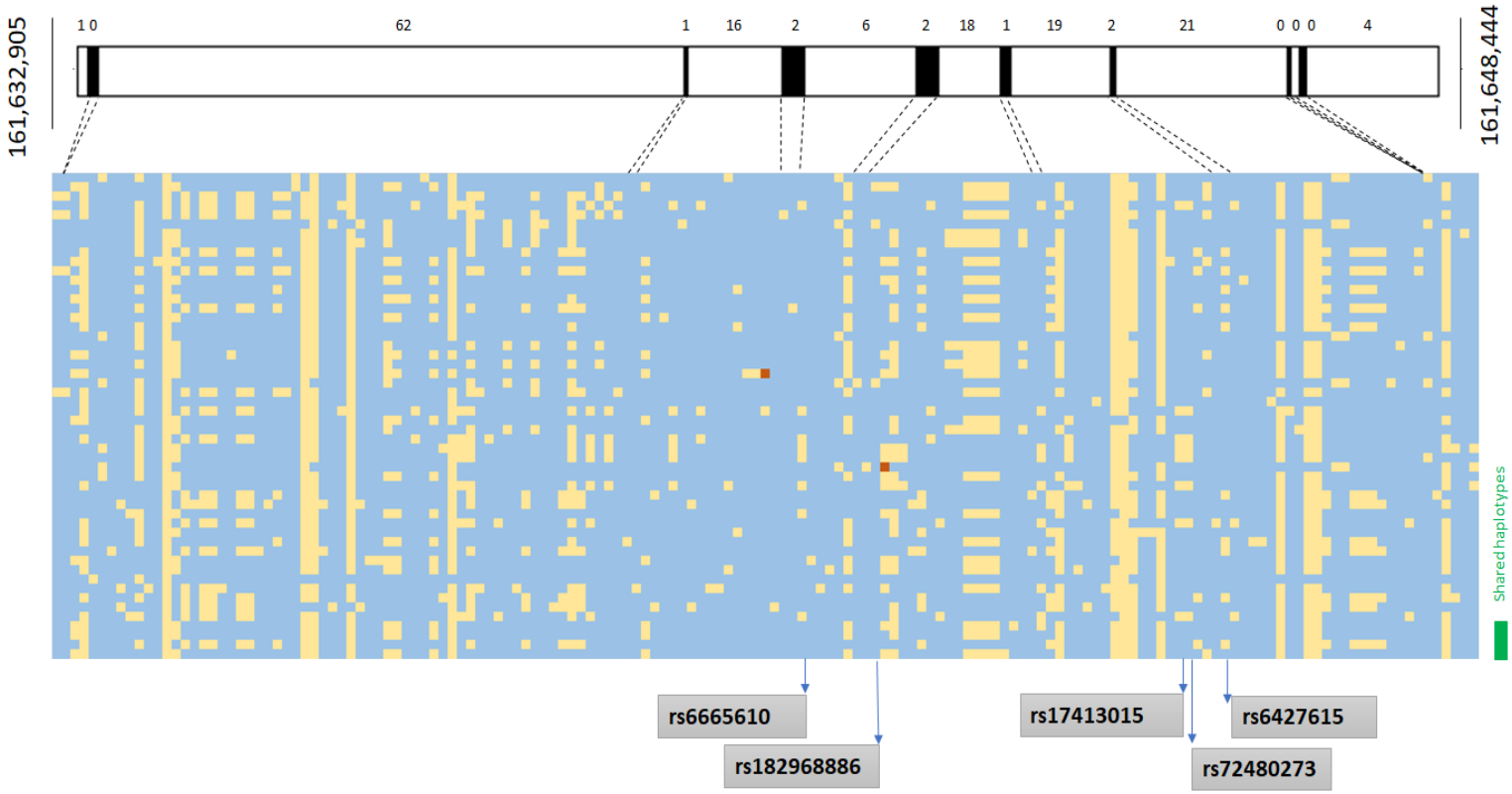


Figure 5. 4 Constructed SNP haplotypes for the *FCGR2B* locus. The colour blue indicates the same variants as the reference. The colour yellow shows the alterative variant and the colour red shows a second variant for the same position. The horizontal bars above is a relative scaling of *FCGR2B* depicting together with summary count of the number of variants. All haplotypes are unique to one individual, except those indicated by the green box on the left bottom which are found in more than one individual. Below, the GWAS SNPs are listed based on their relative position on the *FCGR2B* gene. Coordinates are according to GRCh37/ hg19.

5.3 Development of a PCR assay for genotyping the gene conversion of *FCGR2B*

To investigate whether the previously defined upstream gene conversions influence the expression of the *FCGR2B* gene, first a PCR assay was designed in order to differentiate between the normal alleles, 9.1kb and 2.4kb gene conversion events. Second, with the use of SNPs present in the last two exons and the 3' UTR of the *FCGR2B* gene, the expression levels of alleles were examined for the samples that have gene conversions in comparison to samples without gene conversions. For the confirmation of the gene conversions, the positive control samples were needed. For that, the eight lymphoblastoid cell lines (NA18517, NA18507, NA18956, NA19240, NA18555, NA12878 and NA12156) were used and extract DNA/RNA were extracted by MSc student Poonam Thakkar as a part of her Masters' thesis (Table 5.5).

Table 5. 5 The summary of the samples used for detecting gene conversion.

Sample ID	Population	Gene conversion	Copy number (n)
NA18517	YRI	Allele 1: B->A 9.1kb Allele 2: B->A 2.4 kb	3
NA18507	YRI	Allele 1: no gene conversion Allele 2: unknown	4
NA19240	YRI	Allele 1: no gene conversion Allele 2: no gene conversion	4
NA18956	JPT	Allele 1: B->A 2.4kb Allele 2:no gene conversion	5
NA18555	CHB	Allele 1: B->A ~800bp Allele 2: unknown	5
NA12878	CEU	Allele 1: no gene conversion Allele 2: unknown	4
NA12156	CEU	Allele 1: no gene conversion Allele 2: unknown	4
(Adapted from Rahbari et al, 2017). NA19129 was not used. It has a copy number of 4. One of the alleles was found to have 9.1 b gene conversion.			

The copy number of these cell lines had already been studied. For NA185107, NA12878 and NA12156, only one allele was sequenced, and no gene conversion events were noted. In each case, the other allele was still unknown and hence potentially may carry a gene conversion event. NA18555 carries a B→A gene conversion event which is approximately 800bp which is much smaller than the 9.1kb and 2.4kb gene conversion events. Therefore, it was unknown whether this was an actual gene conversion event or an artefact and so it was excluded from Rahbari et al., (2016).

5.3.1 Primer design of the 9.1kb/ 4.5 kb gene conversions

The primers were designed by MSc student Poonam Thakkar. The fosmid sequences for each sample were mapped to the A and B paralogous sequences and assigned to one of them based on variant identity. The A-like and B-like switches of the variants on fosmid sequences were used to design primers for both left-side and right-side assays of 9.1 kb and 2.4 kb gene conversions. The left-side assay was designed for detecting the beginning of the gene conversions whereas the right-side assay was designed for detecting the end of the gene conversions (Figure 5.5B). The forward primers were designed as A paralogue specific in sequence whereas the reverse primers were designed as B paralogue specific in sequence. An alternative primer was designed as B specific in sequence in the beginning of the gene conversion so it can detect the absence gene conversion event (Figure 5.5A). For the samples NA18956, B specific primer (red coloured arrow) was designed on paralogue B so that it will bind to the paralogue B. Two other primers in the gene conversion region were designed as A specific or B specific. While A specific primer (blue arrow) ends with base 'C', B specific primer (orange arrow) ends with base 'T'. If the individual has base 'C', the beginning of the gene conversion was detected with A specific primer. If the individual has base 'T', no gene conversion was detected with B specific primer. A sample can be homozygous for the gene conversion or no gene conversion. If both primers work, it can be concluded that the sample is heterozygous for the gene conversion.

As both the 9.1kb and the 2.4kb gene conversion begin at the same location, that both sets of primers that were designed at the beginning of the gene conversion events detected the same region so that left-side assay detect for the present or absence of a gene conversion event regardless of length. The primers designed for the right-side assays were non-specifically amplified all the samples so that it will not be mentioned further.

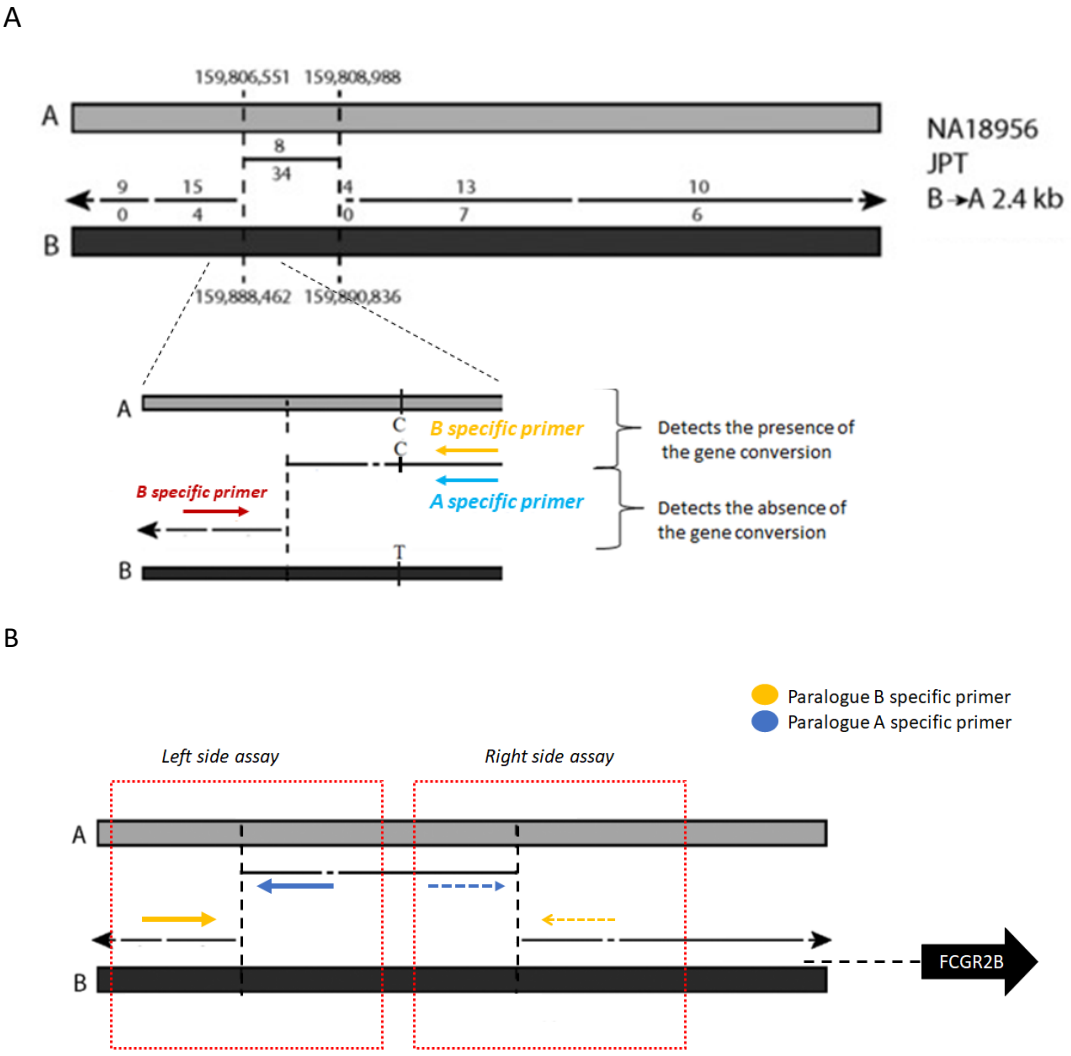
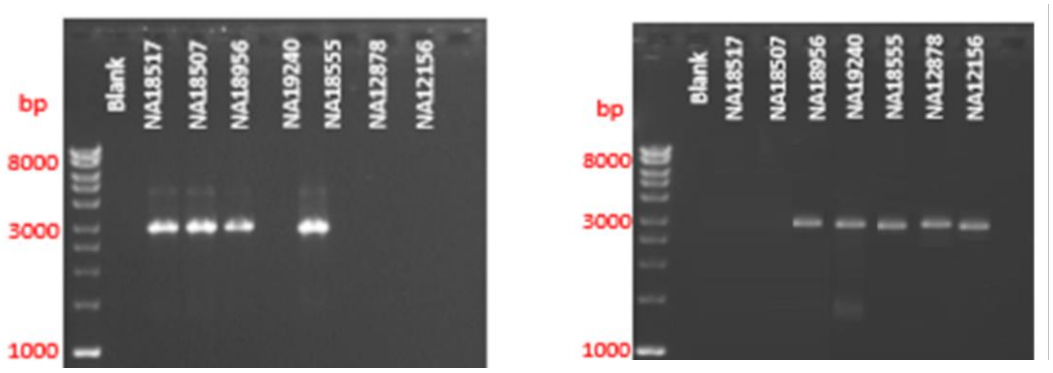


Figure 5. 5 The PCR approach to detect the gene conversion. A: An illustration of the detection of the normal and the gene conversion allele with the PCR primers ending with different bases so that the absence/presence of the start of the gene conversion can be detected. B: The orientation of the primers is shown on paralog A and B to detect the 9.1kb and 2.4kb gene conversions. The left side assay aims to detect the beginning of the gene conversion event whereas the right-side assay aims to detect the end of the gene conversion event. The dash lined primers on the right-side assay were not used for further studies because of nonspecific amplification.

5.3.2 Genotype confirmation of the 9.1kb gene conversion

The seven samples were used to confirm the presence/absence of the gene conversion. Two different PCR reactions were used; first confirms the presence of gene conversion second confirms the absence of the gene conversion (Figure 5.6). The sample NA18517 was used as a positive control NA19240 was used as negative control the left side assays to detect the presence of the gene conversion (refer Table 5.5). With the alternative primer to detect the absence of the gene conversion, the NA19240 was used as positive control, NA18517 was used as negative control. For NA19129, the amplification of the sample failed so no further analysis was performed for this sample to detect any gene conversion event. After the confirmation of the genotypes of these seven samples, the assay was applied to the 24 unrelated samples and the 17 pedigree samples to identify the sample's genotypes. The samples genotyped for the gene conversion assay with control samples is listed in appendix 14.



Presence of gene conversion *Absence of gene conversion*

Figure 5. 6 Gel electrophoresis of the samples for the confirmation of the gene conversion. The DNA was loaded on a 0.8% w/v agarose gel and the ladder used was HyperLadder™ 1kb. assay. NA18517 and NA18507 are homozygous for the gene conversion. NA18956 and NA18555 are heterozygous for the gene conversion. NA19240, NA12878, and NA12156 are homozygous for no gene conversion.

Linkage disequilibrium analyses was performed to investigate the association between the gene conversion and the GWAS SNPs found in our data (Table 5.6). A weak linkage disequilibrium was noted between the gene conversion and the disease associated SNPs.

Table 5. 6 The association of the GWAS SNPs and the gene conversion of FcgR locus.

		LD statistics					
		The samples with copy number 3, 4, 5 and 6			The samples with copy number 3 and 4		
Gene Conversion	GWAS SNP ID	D'	r2	X2	D'	r^2	X^2
9.1kb/2.4kb	rs6665610	-1	0.1234	3.21	-1	0.1071	2.68
	rs182968886	-1	0.0456	1.28	-1	0.0502	1.26
	rs17413015	-1	0.1152	3.23	-1	0.1071	2.68
	rs72480273	-1	0.1152	3.23	-1	0.1071	2.68
	rs6427615	-0.76	0.1864	5.22	-0.749	0.1889	4.72

5.3.3 Allelic imbalance in gene expression.

The RNA of the eight cell lines were analysed in order to measure the *FCGR2B* relative transcript levels and to detect whether any differences were related to the gene conversion events. Alternative splicing of FcgRIIB transcripts results in Fc*γ*RIIB1 and FcgRIIB2 isoforms. FcgRIIB1 is has a longer cytoplasmic tail than FcgRIIB2. The longer cytoplasmic tail enables the prevention of endocytosis of the receptor when engagement of the receptor with IgG immune complexes takes place. In comparison, FcɣRIIB2 is expressed mainly in myeloid cells. mAbs are mainly internalised when FcɣRIIB2 isoform is present. FcgRIIB1 is predominantly expressed on B cells, whereas myeloid cells express dominantly the FcgRIIB2 isoform (Lehmann et al., 2012; Roghanian; et al, 2018). The difference between the two transcripts is because FcgRIIB2 does not contain exon 6 whereas FcgRIIB1 does (Amigorena et al., 1992). On UCSC Genome Browser, GRCh38/hg38 assembly, another transcript is also present for the *FCGR2B* gene which does not contain exon 2 (Figure 5.7).

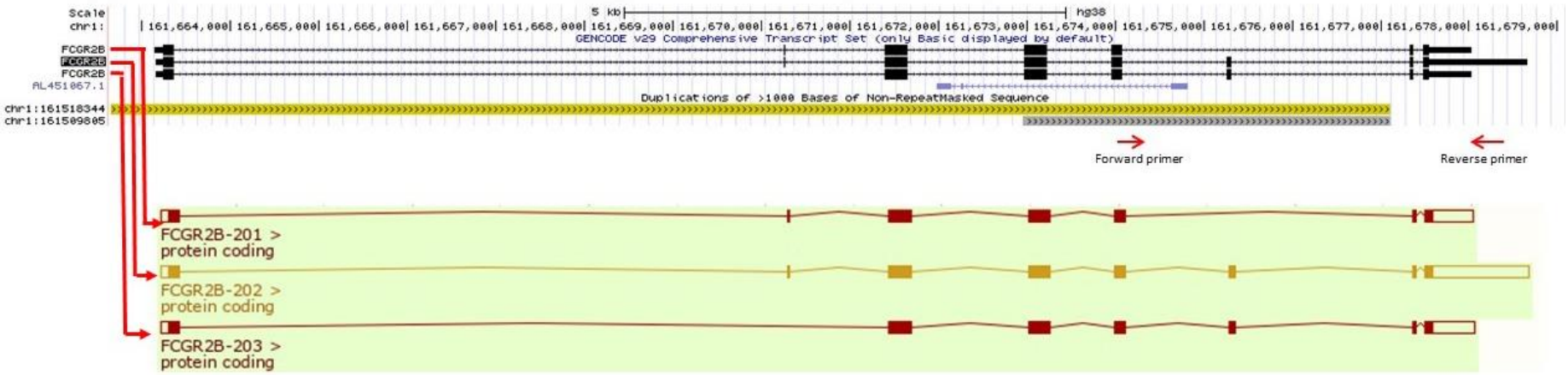


Figure 5. 7 The isoforms of FcgRIIB and primer design for measuring the *FCGR2B* relative transcript levels. The three isoforms of FcgRIIB are shown. The first transcript is the FcgRIIB2, no exon 6, Refseq accession number is NM_001002274.2. The second transcript is the FcgRIIB1, contains both exon 2 and exon 6, RefSeq accession number is NM_004001.4. he third transcript is the third isoform of FcgRIIB, no exon 2, RefSeq accession number is NM_001190828.1. The red arrows show where the primers were designed for amplifying the region of two exons and the 3'UTR region of *FCGR2B*.

The last two exons and the 3'UTR are shared by FcgRIIB1 and FcgRIIB2 and the third isoform. To differentiate between the two alleles from each sample, SNPs were identified in the last two exons and the 3'UTRs. There is no evidence of complete monoallelic expression, and any expression differences are subtle. Therefore, the fosmid sequences were used to check which gene conversion allele was on the same haplotype as the SNP allele used to measure allelic imbalance. Only two SNPs were chosen to be analysed: rs844 and rs60519172 as these were the only two SNPs where heterozygotes existed in the samples used. Only four of the samples were found to be heterozygous for the SNPs. PCR amplification of cDNA samples were performed using the same primer pair so that the product can cover the area which has these two SNPs. The primer pair is specific to *FCGR2B* as one of the primers is in the non-duplicated region (Figure 5.8). cDNA samples of NA18517, NA19240, NA18555 and NA19129 were amplified and Sanger sequenced. Sanger sequencing of the PCR products allows quantification of the different alleles in heterozygotes using QSV analyser (Chapter 2.16). The raw averages of the variants for each sample is shown in Figure 5.8.

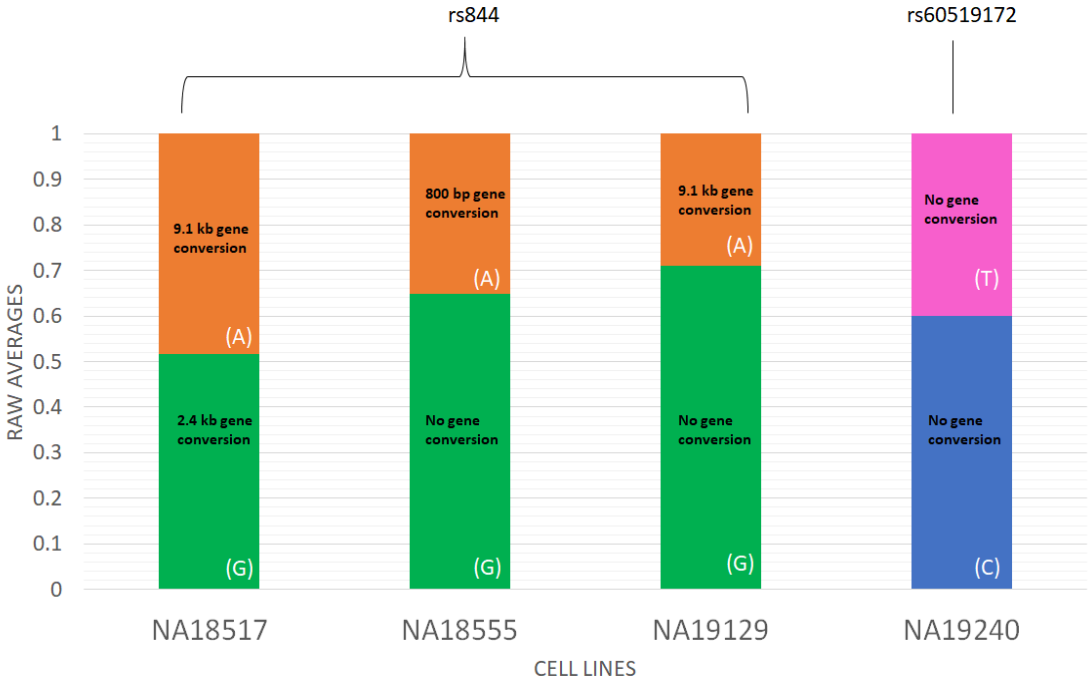


Figure 5. 8 The raw averages of the variants for each cell line.

At the SNP rs844, NA18517 has no significant difference between the expression of allele G and allele A, however, for the NA18555 and NA19129, the expression of the allele with the G is higher than the alleles with A at the same SNP position. There is a lower expression of the A alleles where the gene conversion is present. While there is approximately equal expression of both alleles in NA18517 as they both carry gene conversion events, G allele is expressed higher in NA18507 and NA19129. There may be slightly higher expression of G allele when there is no gene conversion. For NA19240, there is a higher expression of the allele with a C compared to the allele with T at the rs60519172 SNP position. As only one sample has rs60519172 variation, it is not possible to link this variant to any gene conversion as none of the alleles carry any gene conversion. There may be another factor influences the expression levels of this variant.

5.4 Discussion

One of the aims of this chapter was performing deeper analysis of the variants found specifically for the *FCGR2B* gene after high-throughput sequencing. The novel 17 variant was discovered within 148 variants for specifically *FCGR2B* locus. rs1050501, associated with malaria and SLE, was found to be in relatively strong linkage disequilibrium with the rs6427615 which has been associated with blood protein measurement. The GWAS SNPs rs17413015 and rs72480273 were also found to be in strong linkage disequilibrium with the SNPs rs12116547 and rs12145843.

Another aim of this chapter was to design an assay to confirm gene conversion in the upstream region of the *FCGR2B* gene and analyse whether the gene conversion influences the expression of the gene. A PCR-based assay amplified the left-side but not the right-side of the gene conversion so the length of the gene conversion could not be identified, however, the gene conversion was confirmed with the left-side assay (either 9.1kb or 2.4kb). It is possible that the mismatches may not be at the same position and the primers designed for the right-side assay were not able to amplify all the samples.

The samples with gene conversion are searched for any association between GWAS SNPs, however, no high linkage disequilibrium was found. To investigate whether gene conversions

have an impact on the levels of *FCGR2B* expression, the region covering two SNPs was sequenced and the expression of the variants were analysed. There is a much higher expression of the allele with a G at the position rs844 for the NA18555 and NA19129 compared to NA18517 which both alleles carry gene conversion events and have equal expression levels. It might be suggested that expression of G is slightly higher when there is no gene conversion. As only one sample was inspected for the rs60519172 SNP, it was not possible to draw any robust conclusions from the data.

CHAPTER 6 Population genetics and evolutionary analysis of *FCGR2B* locus

6.1 Introduction and study rationale

Humans show the most complex system for the low-affinity FcgRs region and the gene duplications were probably accompanied by intergenic recombination. The low-affinity Fc receptors recently underwent species-specific duplication and recombination that leads to different activatory receptor genes. However, the inhibitory receptor FcgRIIB remained as a single inhibitory gene in all mammal species in contrast to the activatory receptors (Fanciulli et al., 2008). There have been several SNPs identified in human *FCGR2B*, but only one of these SNPs has been stated as clinical significance and occurs at notable frequency which is rs1050501 as previously mentioned (Chapter 5.2.1). The substitution of a threonine for isoleucine due to the non-synonymous T-to-C transition in exon 5 (rs1050501) results in decreased expression of FcgRIIB (Smith and Clatworthy, 2010). The frequency of the FcgRIIBT232 variant in different populations varies significantly (Figure 6.1). The frequency of the homozygous genotype of FcgRIIBT232 is lower in Caucasians (1%) than Africans (8–11%) and Southeast Asians (5–7%) (Kyogoku et al., 2002; Siriboonrit et al., 2003). The FcgRIIBT232 polymorphism in humans is also associated with susceptibility to SLE in Southeast Asian populations and in Caucasians. The increased frequency of FcgRIIBT232 in individuals of Southeast Asian and African descent may contribute to the increased prevalence and/or severity of SLE. The FcgRIIBT232 allele is common in the areas where malaria is endemic; therefore, this may suggest that decreased FcgRIIB function provides a survival advantage against this disease and thus could explain the higher frequency of FcgRIIBT232 in Africans and Southeast Asians. (Kwiatkowski, 2005; Smith and Clatworthy, 2010). It has been suggested that the fragile balance between proinflammatory and anti-inflammatory mediators required to survive repeated malarial infections modulates the immune system and protects against autoimmune disease. This could explain why SLE is less common in malaria-endemic (Africa Butcher et al., 2008; Willcocks et al., 2010).

Since *FCGR2B* is located on the distal position of the segmentally duplicated region of the low affinity FcgRs locus without showing CNV, specific amplification of the gene was successful

(Chapter 4.2). Consequently, it became possible accurately genotype this region and discriminate the DNA sequence polymorphism from other FCGR genes. Since the sequence variation of *FCGR2B* can accurately be analysed without the issues regarding the copy number or duplications, it is possible to analyse the population genetic pressures on *FCGR2B* in humans. This is particularly important given the role of *FCGR2B* in malaria, SLE or other diseases mentioned before. The rs1050501 variants have been shown to affect the susceptibility to diseases in certain populations which creates a strong selective force on *FCGR2B* gene.

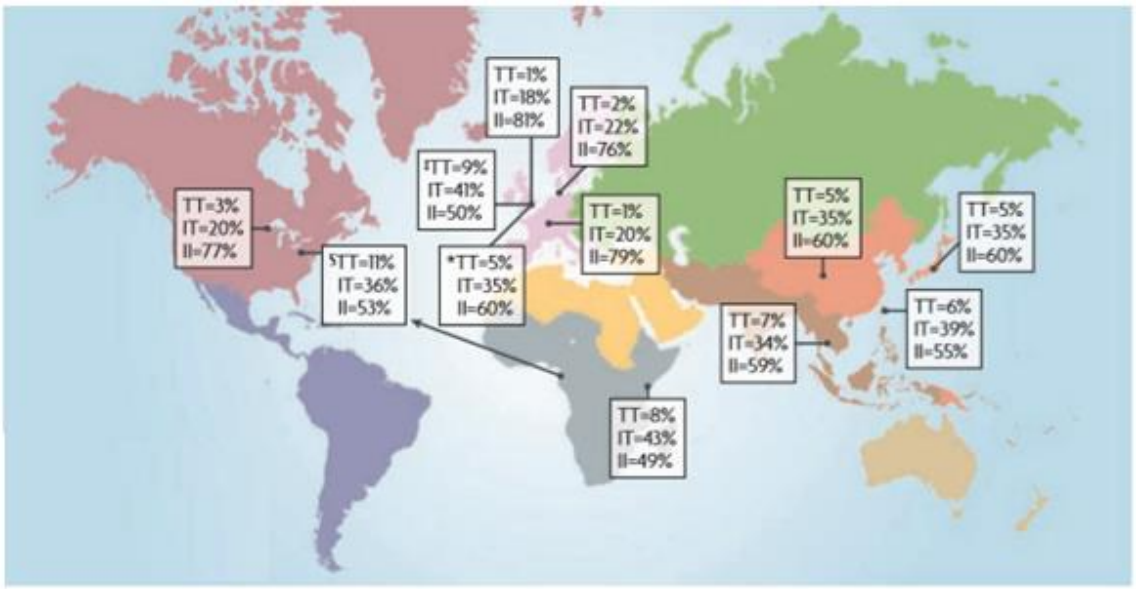


Figure 6. 1 Frequency of FcgRIIBT232 polymorphism (rs1050501). The frequency of the FcγRIIBT232 variant in different populations varies substantially. It is uncommon in Caucasians (1% homozygosity) but more common in populations found in areas of malarial endemicity such as Africa and Southeast Asia (5–11% homozygosity) suggesting that decreased FcγRIIB function may provide a survival advantage against this disease. Obtained from Smith and Calthworty, 2010.

The analysis of DNA sequence polymorphisms and measuring genetic variation can provide an understanding into the evolutionary forces acting on this gene across populations and species. The aim of this chapter was to look at the human *FCGR2B* from the perspective of genetic diversity, population genetics and inspecting it for the signature of selection.

6.2.1 Population genetic statistics

A total number of 67 sequences were obtained from the 24 of 1000 Genomes Project, the four founders of the platinum data plus the fosmid sequences for six samples from another

study (Rahbari et al., 2016). The distribution of the populations used is shown in Figure 6.3. As previously described (Chapter 2.12, Chapter 5.2.2), a total number of 52 haplotypes was detected for the *FCGR2B* from the sequence of 34 samples (Appendix 17). There are 155 total number of variable (polymorphic) sites with 157 total number of mutations and 52 singleton variable sites (Figure 5.4, at two different position, there are second alternative variants). A variable site contains at least two types of nucleotides. Figure 6.2 shows the location of the variable sites for each population and groups used in this study. Most of the populations have the nucleotide change in the same positions. Most variable sites are shared across groups. Among the populations, Asian populations together (EAS and SAS) have the lowest number of polymorphic sites. African populations have the highest number of variable sites. AMR and AFR populations also have the highest number of singletons. The numbers are indicating the total number of variable sites and singletons respectively for each population and group. EAS:51/17, SAS: 64/20, EUR: 94/28, AFR: 110/37, AMR: 76/37. Asian: 69/13, Non-Asian: 145/43, Non-European: 145/52, Non-African:112/30, Non-American: 149/50. Most of the variation is observed in the intron regions compared to exon regions.

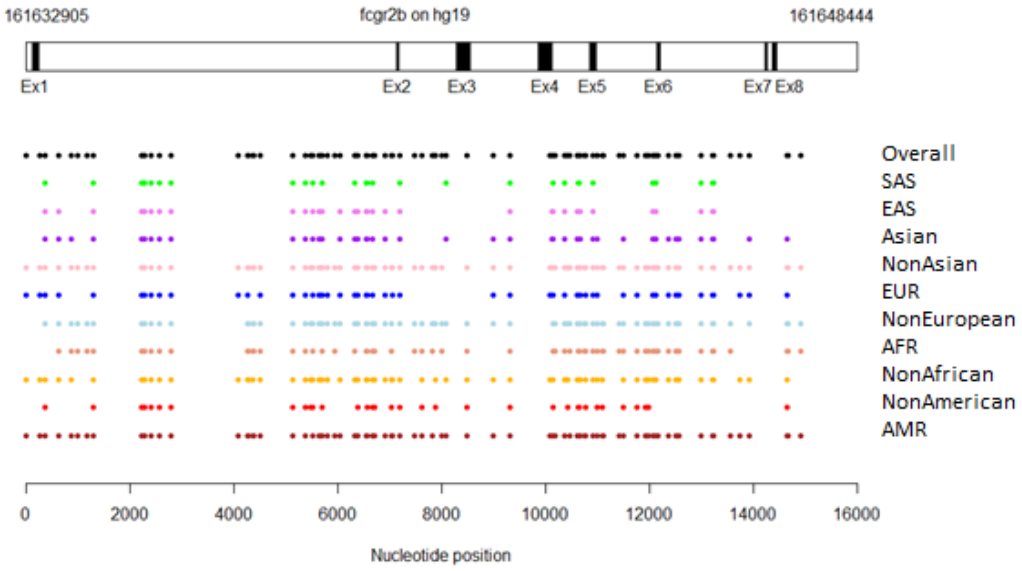


Figure 6. 2 Comparison of location of the polymorphic sites of *FCGR2B*. The top rectangle represents the *FCGR2B* gene, the exons also indicated as black lines. Based on populations/groups, each dot corresponds approximate location (whether the polymorphic site is in the intron or exon region) on *FCGR2B* gene. Coordinates are according to GRCh37/ hg19.

Figure 6. 3 The distribution of the populations used for the population genetics and evolutionary analyses of *FCGR2B*. Yellow circles: African; Red circles: American; Green circles: East Asian; Blue circles: European; Purple circles: South Asian.

The average haplotype diversity was found to be high in human *FCGR2B* (Table 6.1). Especially AFR and AMR populations seems to have higher diversity while Asian populations have the lowest diversity. Average number of nucleotide differences is the highest in African followed by EUR and the lowest in Asian populations.

Table 6. 1 Comparison of population genetic statistics of *FCGR2B* locus.

Population/group	n	S	S (singleton)	h	Hd	pi
EAS	13	51	17	9	0.91	0.0012
SAS	10	64	20	7	0.911	0.0016
EUR	16	94	28	13	0.95	0.0019
AFR	20	110	37	20	1	0.0019
AMR	8	76	37	8	1	0.0018
Asian	23	69	13	15	0.933	0.0014
Non-Asian	44	145	43	38	0.986	0.0019
Non-European	51	145	52	41	0.978	0.0017
Non-African	47	112	30	33	0.957	0.0016
Non-American	59	149	50	46	0.974	0.0017
Overall	67	155	52	52	0.975	0.0017

Sample sizes (n) is the number of chromosomes, the number of haplotypes (h), total number of variable sites (S), haplotype diversity (Hd), nucleotide diversity (π-pi) **sample sizes refer to the total number of chromosome sequences, there is only one chromosome sequence obtained for some samples. AFR: African; AMR: Ad Mixed American; EAS: East Asian; EUR: European; SAS: South Asian; Asian: EAS, SAS; Non-Asian: EUR, AFR, AMR; Non-European: EAS, AFR, AMR, SAS; Non-African: EAS, EUR, AMR, SAS; Non-American: EAS, EUR, AFR, SAS.

Nucleotide diversity is a measure of genetic variation. The average nucleotide diversity was found 0.002 for FCGR2B gene. Analysis with the sliding window plot revealed that nucleotide diversity is not uniform. Maximum nucleotide diversity was found in the regions where exon 5 is located (Figure 6.4).

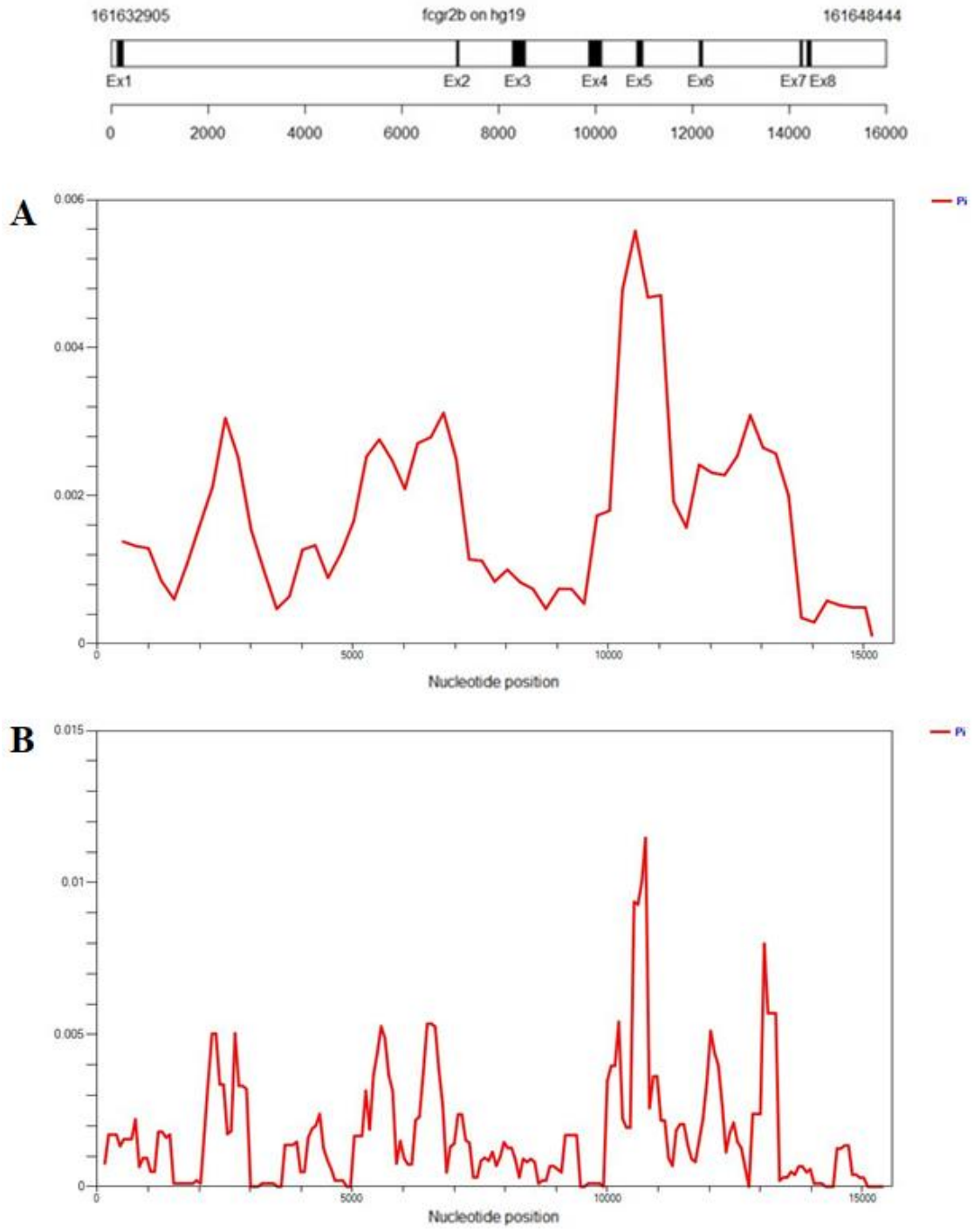


Figure 6. 4 Sliding window plot of nucleotide diversity (pi) of the *FCGR2B* locus. A: the graph for window length 1000bp with step size 250bp. B: the graph for window length 300bp with step size 75bp. The Y-axis is the nucleotide diversity. The X-axis is the nucleotide position. The red line shows the value of nucleotide diversity. A schematic representation of *FCGR2B* is given on the top based on the coordinates of GRCh37/hg19.

Nucleotide diversity for the exons of *FCGR2B* is shown in Figure 6.5 and Table 6.2. The average nucleotide diversity of the exons is equal to the overall value for this locus. While exon 5 has the highest nucleotide diversity among the exons of the *FCGR2B* region, exon 1, 7 and 8 do not show any nucleotide diversity. The SNP rs1050501 is located in the exon 5 and found in many samples as heterozygous. The reason why nucleotide diversity is higher for along the *FCGR2B* gene and its exons could be because the variants of rs1050501 has been associated with SLE and malaria so that higher nucleotide diversity can be expected for the exon 5 in the populations.

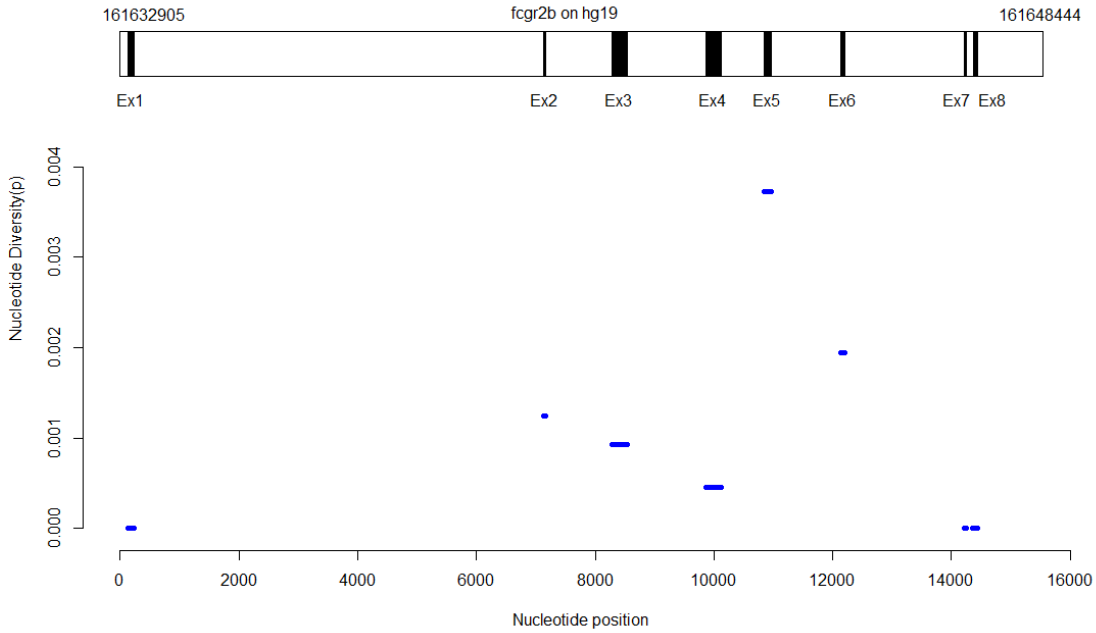


Figure 6. 5 The nucleotide diversity for the exons of *FCGR2B*. The figure shows the comparison of the nucleotide diversity for the exons of *FCGR2B*. A schematic representation of *FCGR2B* is given on the top based on the coordinates of GRCh37/hg19.

Table 6. 2 Genetic diversity in the exons of *FCGR2B*.

Domain	Region	Size (bp)	S (segregating sites)	π (nucleotide diversity)	SNP
exon1	128-239	112	0	0.0000	-
exon 2	7138-7159	21	1	0.0014	rs1478203045
exon 3	8278-8535	259	2	0.0009	rs6665610 rs1487501626
exon 4	9861-10115	255	2	0.0005	rs2298022 rs182968886
exon 5	10846-10959	114	1	0.0038	rs1050501
exon 6	12143-12199	57	2	0.0020	rs2865183 rs148534844
exon 7	14214-14251	38	0	0.0000	-
exon8	14362-14439	78	0	0.0000	-
Total	1-15540	15540	8/155	0.0017	

6.2.2 Population differentiation analysis

Genetic diversity can be measured by different parameters. To test if rs1050501 contributes to genetic differentiation between populations, Gst scores for each pairwise population and group comparison were calculated (Table 6.3). Gst measures the proportion of gene diversity that is distributed among populations and it is defined as the coefficient of gene differentiation. It is a common index used for the interpopulation differentiation for several loci. When there are two populations and two alleles, Gst ranges from 0.0 to 1.0. 0 representing no differences in allele frequencies between two populations and populations are undifferentiated genetically Gst 1.0 indicates that the two populations are fixed for alternate alleles, differentiation is at its maximum and Gst exhibits the intuitive value of one. Gst necessarily approaches zero when gene diversity is high. The Gst values of the populations and groups used in this study were found to be very low (less than 0.025) indicating minimum or zero population differentiation at *FCGR2B* locus.

Table 6. 3 Pairwise Gst values among populations and groups for *FCGR2B*.

Population/Group 1	Population/Group 2	Gst
EAS	SAS	0.023
EAS	EUR	0.017
EAS	AFR	0.019
EAS	AMR	0.015
SAS	EUR	-0.004
SAS	AFR	0.017
SAS	AMR	0.004
EUR	AFR	0.007
EUR	AMR	-0.001
AFR	AMR	0.005
Asian	Non-Asian	0.009
EUR	Non-European	0.005
African	Non-African	0.008
AMR	Non-American	0.018

6.2.3 Neutrality tests on human *FCGR2B*

The Tajima's D, Fu and Li's D* and F* statistics test the hypothesis that all mutations are selectively neutral and there are no deviations in the distribution of allele frequencies. These tests were run on the data set to estimate the deviation from neutrality, asking if the values differ from 0. The Tajima's D test uses the total number of segregating sites (S) and the average number of mutations between pairs (π) in a randomly sampled populations (Tajima, 1989). Tajima's D is expected to be 0 under the neutral model. Negative values indicate high frequency of rare haplotypes which may be because of population expansion under positive selection. Positive values may indicate balancing selection because of lack of the rare alleles in a population. Fu and Li's D* statistic compares the number of singleton mutations and the mean pairwise difference between sequences while Fu and Li's F* statistic uses the number of singleton mutations and the total number of nucleotide variants (Fu and Li, 1993). Negative values for both tests show an excess of singletons and indicate excess number of alleles which can be expected from a recent population expansion and genetic drift. Positive values also indicate allele deficiency in the populations which may be because of a bottleneck.

In this study, the Tajima's D (-0.633) and Fu and Li's D* (-1.408) and F* (-1.317) statistics, all have p values of >0.10 (Table 6.4). These statistics therefore failed to detect a significant departure from neutral expectations in this dataset.

Table 6. 4 Neutrality tests on human *FCGR2B.*

Populations/groups	Tajima's D	Fu and Li's D*	Fu and Li's F*
EAS	0.588	0.074	0.243
SAS	0.512	0.335	0.429
EUR	0.125	0.126	0.145
AFR	-0.230	-0.293	-0.320
AMR	-0.190	-0.173	-0.198
Asian	0.492	0.600	0.664
Non-Asian	-0.492	-0.610	-0.677
Non-European	-0.696	-1.330	-1.304
Non-African	-0.062	-0.332	-0.277
Non-American	-0.618	-1.265	-1.214
Human	-0.633	-1.408	-1.317
	P > 0.10	P > 0.10	P > 0.10

The McDonald-Kreitman is another neutrality test (McDonald and Kreitman, 1991) based on a comparison of intraspecific polymorphism to interspecific divergence to infer the impact of natural selection in the human lineage since the split with other species such as chimpanzee or gorilla. This test requires enough time for an adequate number of adaptive substitutions to accumulate in the divergence of a gene between two species. The assumption of the test is that the ratio of nonsynonymous (Pn) to synonymous (Ps) variation within a species is to equal the ratio of nonsynonymous (Dn) to synonymous (Ds) variation between species. The neutrality index (NI) estimates the direction and degree of departure from neutrality. It is calculated by using this formula: (Pn/Ps) / (Dn/Ds) (Stoletzki and Eyre-Walker, 2010). Assuming the silent mutations are neutral, a positive NI is the result against neutrality and may indicate that negative selection is at work or a balanced selection. A negative NI value indicates positive selection because of excess non-silent divergence. High NI values also indicate higher selective restrictions during the human lineage evolution.

The comparison of the human and chimp, human and gorilla *FCGR2B* sequences shows a positive neutrality index for both coding regions as well as coding and noncoding regions. At low value, the neutrality index indicates there may be a negative selection, however, the p-values are not significant, we can conclude that the silent mutations are neutral on the *FCGR2B* (Table 6.5).

Table 6. 5 The McDonald-Kreitman tests on *FCGR2B.*

Populations	Polymorphic changes within species		Fixed differences between species		Neutrality index (NI)	*p value
	Coding regions					
	Synonymous (Ps)	Nonsynonymous (Pn)	Synonymous (Ds)	Nonsynonymous (Dn)		
Human vs Chimp	2	7	2	7	1	1
Human vs Gorilla	2	7	2	8	0.875	1
	Coding and noncoding regions					
	Synonymous	Nonsynonymous	Synonymous	Nonsynonymous		
Human vs Chimp	150	7	187	7	1.247	0.786
Human vs Gorilla	148	7	227	8	1.342	0.599
***Fisher's exact test: not significant** **Chimpanzee: Pan troglodytes (chimpanzee) Assembly access code: GCF_002880755.1** **Gorilla: Gorilla gorilla gorilla (western lowland gorilla) Assembly access code: GCA_900006654.3**						

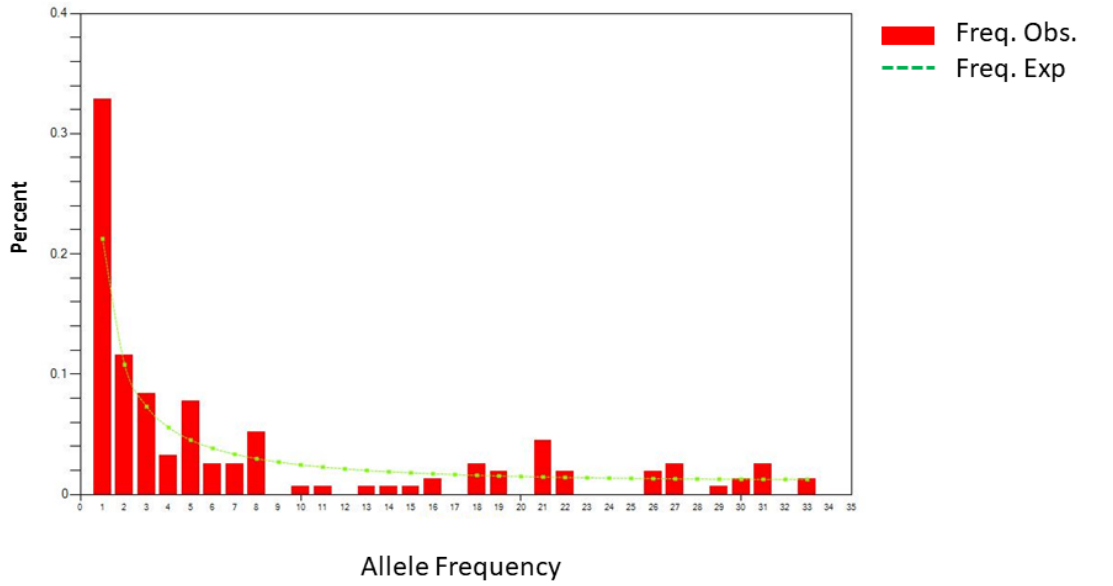
6.2.4 Scan of signature of selection on *FCGR2B*

The allele frequency spectrum (AFS) is the distribution of the allele frequencies of a given set of loci in a sample. The shape of the allele frequency spectrum is affected by population genetic processes such as population size changes and natural selection (Han et al., 2014). When there is a directional selection, mutations are expected to occur at low frequency, skewing the spectrum of allele frequencies to left. Rapid growth in populations can result in more low-frequency alleles than expected under neutrality because rare alleles tend to stay in the growing population. When there is balancing selection, there is an excess of intermediate-frequency alleles, skewing the spectrum of allele frequencies to right. The distribution of the allele frequencies of *FCGR2B* for this dataset is shown in Figure 6.6A. Most of the observed allele's frequencies are under the

expected line. The shape of the spectrum is relatively skewed to the left and indicate some rare haplotypes and there are clear no intermediate observed allele frequencies to indicate balancing selection. Although the rare haplotypes may indicate a population growth under directional selection, Tajima's D, Fu and Li's D* and F* statistics also did not detect a significant departure from neutral expectations.

The mismatch distribution is a frequency graph of pairwise nucleotide differences between pairs of individuals. It can be used to estimate the population dynamics and can be a graphical way of visualizing the signature of population expansion (Rogers and Harpending, 1992). While a unimodal distribution may show a recent population growth, a multimodal distribution may indicate demographic stability. Figure 6.6B shows the mismatch distribution under the constant population size for the *FCGR2B*. The position of the peak reflects the time of the population growth. The observed frequencies show some waves but overall, it is a unimodal because one clear peak with higher frequency indicates many haplotypes are diverse from each other. Tajima's D, Fu and Li's D* and F* tests were not significant, we reject the presence of a population expansion or selection in the analysed dataset.

A



B

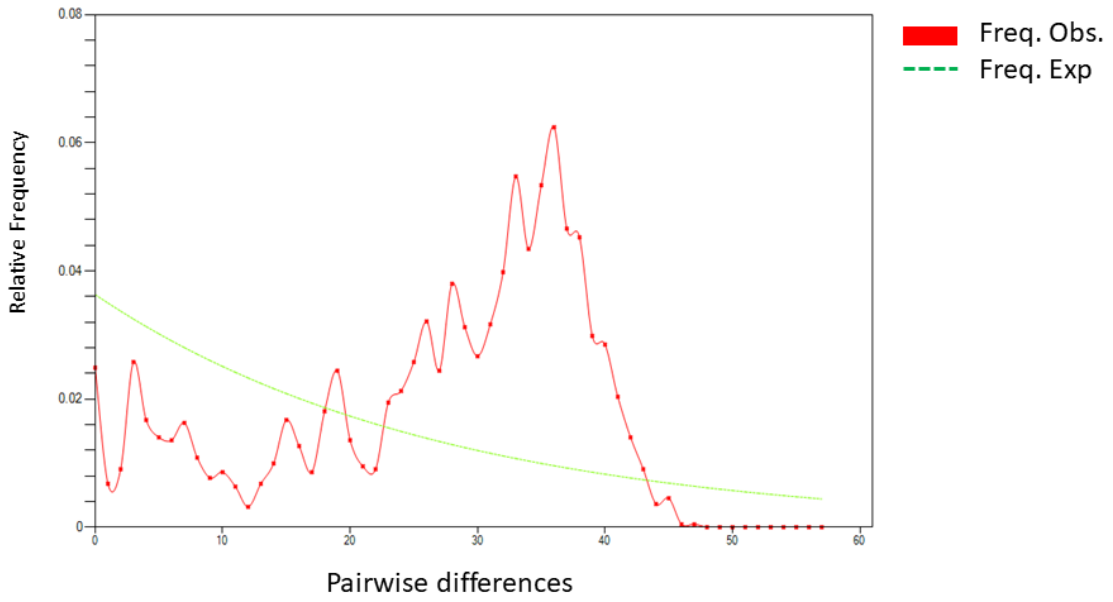


Figure 6. 6The allele frequency spectrum and mismatch distribution. A: Allele Frequency Spectrum (AFS): The plots illustrate the expected (solid green line) and observed (red bars bars) frequencies of alleles for the allele frequency spectrum. B: Mismacth Distribution: The plot shows Exp: Expected Obs: Observed alleles frequencies for the mismatch distribution.

6.2.5 Haplotype network of *FCGR2B*

Haplotype networks are connected graphs with cycles to explore DNA sequence variation at the intraspecific level which is a very informative way to illustrate relationships among the sampled haplotypes (Mardulyn, 2012). To investigate whether there is a selection pressure on a particular haplotype and to explore any particular population structure among observed haplotypes in this dataset, a haplotype network was constructed using Median-joining (MJ) algorithm. While drawing MJ networks, the parameter epsilon can be used in different values. This parameter states a weighted genetic distance to the known sequences in the data set, within which potential median vectors may be constructed. If epsilon is less than the greatest weighted genetic distance within the data set, then theoretically the MJ network will not contain all possible shortest trees. If epsilon is set equal to (or greater than) the greatest weighted genetic distance, the MJ algorithm is guaranteed to produce a full median network. A better representation of network will be produced. Frequently epsilon=10 is found to be sufficient (Bandelt et al. 1999).

In this work, epsilon=0 (default) and epsilon=10 were used to explore whether and how the network changes in this dataset (Figure 6.7A and 6.8B). The *Pan troglodytes* (chimpanzee)(GCF_002880755.1) and Gorilla: *Gorilla gorilla gorilla* (western lowland gorilla) (GCA_900006654) are also used to show outgroups.

The network constructed with epsilon 10 shows a high lozenge-shaped reticulation among median vectors indicating ancient recombination events compared to the network with epsilon value 0. It also shows that there has been some historical recombination since the separation among human, chimpanzee and gorilla for the *FCGR2B* locus. There may be ongoing recombination events for the current populations. Both figures do not show any rational separation or grouping based on any criteria. As indicated in the mismatch distribution the haplotypes are very diverse and there is no clear clustering of haplotypes. rs1050501 does not seem to be responsible for selection and no evidence for selection at *FCGR2B.* It can be suggested that there is no existence of selection pressure on a particular haplotype in a particular population.

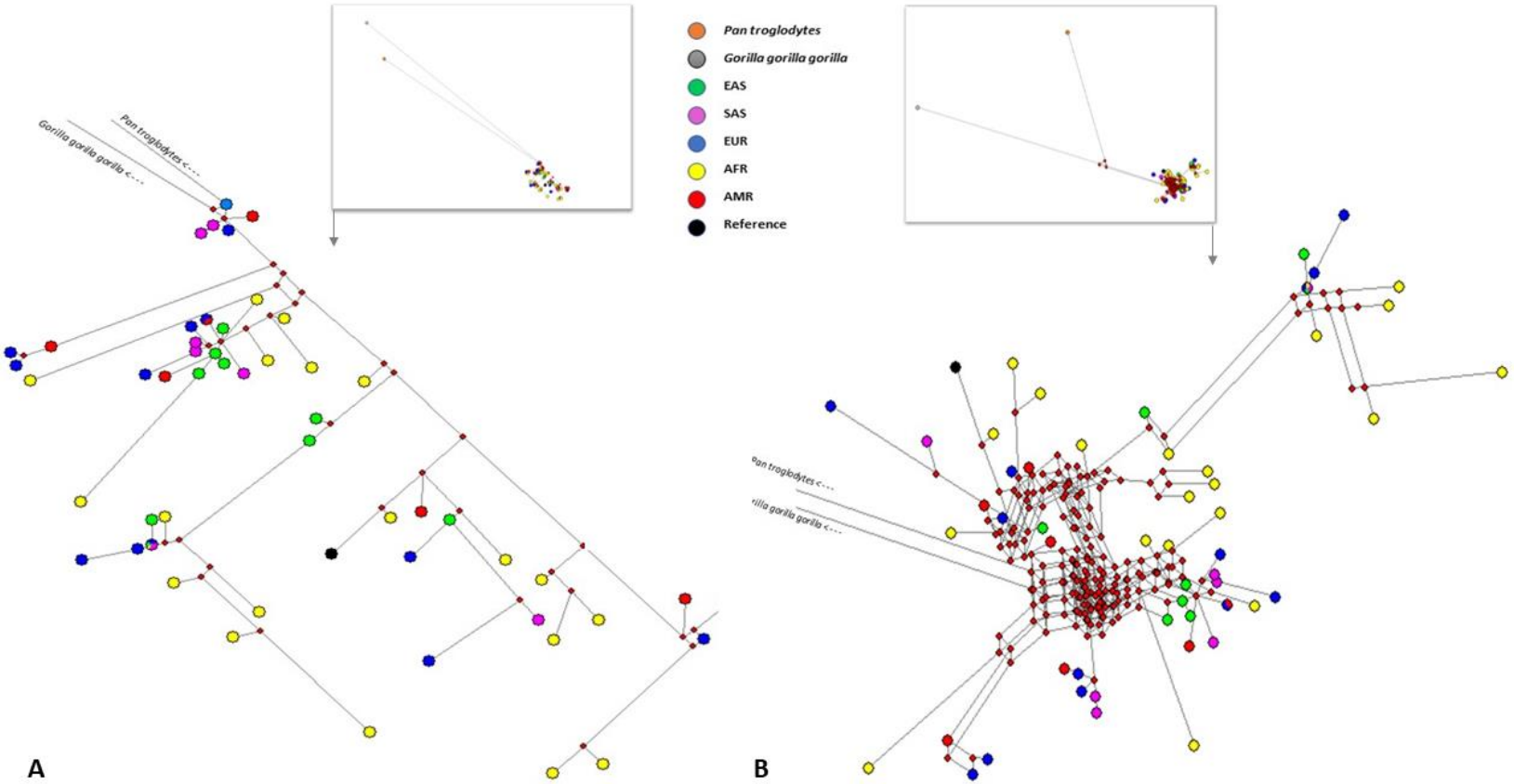


Figure 6. 7 The haplotype network of human FCGR2B with epsilon value 0 (A) and with epsilon 10 (B). The haplotypes are represented by circles/nodes. The red diamond shapes are hypothetical ancestors (median vectors). The branch length between nodes is proportional to the number of mutations and the branches between nodes indicate recombination. Both figures do not show a clear clustering of haplotypes based on any criteria. rs1050501 does not seem to be responsible for selection and no evidence for selection at *FCGR2B.*

6.3 Discussion

The genetic diversity and population genetics of the human *FCGR2B* were investigated based on sequencing of total 34 samples. Each individual used in this study revealed two haplotypes mostly because of singletons. The genetic diversity of this locus revealed high haplotype diversity (52 haplotypes, Hd:0.975). The high haplotype diversity and low nucleotide diversity values suggest small differences between haplotypes. Despite the wide distributional range of the populations, the estimation of interpopulation comparison (Gst) support low level of genetic differentiation between these populations. This was also demonstrated by the haplotype network, which represents mostly single nucleotide differences between majority of haplotypes. The haplotype network shows ancient recombination events and existing recombination for current populations. The combination of high haplotype and low nucleotide diversity can be a signature of population expansion. The Tajima's D, Fu and Li's D* and F*, neutrality tests did not find statistical significance so that fail to say the gene follow the neutral evolution model. The McDonald-Kreitman test was also applied to human, gorilla, and chimpanzee *FCGR2B* sequences to compare intraspecific polymorphism and interspecific divergence. No statistical significance was found to reject that mutations are selectively neutral. rs1050501 does not seem to be responsible for selection and no evidence for selection at *FCGR2B* in humans in contrast to what might be predicted given its suggested role in malaria.

Once the population genetics and evolutionary analyses of *FCGR2B* was completed, the results were attempted to be compared to metrics provided by the PopHuman online genome browser. (http://pophuman.uab.cat). This database is a population genomics-oriented genome browser providing a broad catalogue of population genetics metrics from the 1000 Genomes project phase 3 (Casillas, 2018). Several selection tests are accessible (Fu and Li's D* and F*, Tajima's D, NI or alfa from Macdonald-Kreitman test) and metapopulations or single population can be chosen for variation statistics. However, the PopHuman genome browser did not have statistical results for the test mentioned above for *FCGR2B* as well as the whole low-affinity FcgR locus. No

comparison was performed between this study and the PopHuman genome browser statistics. These statistical analyses might have been excluded from the PopHuman genome browser for the regions/genes that contain high sequence homology with the presence of copy number variable regions such happens in the low-affinity FcgR locus. locus.

CHAPTER 7 DISCUSSION

The human genome includes numerous blocks of duplicated sequences which have a high percentage sequence similarity. Locating and characterizing of these regions has been always a great interest because of their great contribution to species divergence, the association between these regions and chromosomal instability or evolutionary rearrangement, the contribution to the genetic variation in human and their fully or partially association to the genetic diseases and disorders (Bailey et al., 2002; Sharp et al., 2006; Alkan et al., 2011; Numanagić et al., 2018). However, accurate identification and detection of sequence variation within these regions using high throughput sequencing can be challenging due to the limitations of short read sequencing.

This PhD thesis designed a computational pipeline in an attempt to resolve the ambiguity in the read mapping for the duplicated DNA regions in the human genome by using reduced references. It performed mapping of short sequence reads against reduced references by using different computational mapping programs and compared the performance of these programs. To investigate the sequence variation, the copy number of low-affinity FcgR locus were investigated. The copy number of previously studied samples for this locus was confirmed and more samples were genotyped for copy number. For the confirmation for the sequence variation (variant calling), *FCGR2B* gene was sequenced, and haplotypes were constructed for the samples used in this study. A PCR-based assay was developed to confirm the 9.1kb/2.4kb gene conversion in the upstream of F*CGR2B* and effects of gene conversion were investigated on the expression profile of *FCGR2B*. *FCGR2B* was also studied from the perspective of genetic diversity, population genetics and inspected for the signature of selection regarding the predicted importance of rs1050501.

7.1 Nucleotide variant calling is not successful by reduced reference mapping.

To solve the read mapping ambiguity in short read sequencing and to investigate the sequence variation in the repeated DNA regions of the genome, two different reduced references were constructed using GRCh37/hg19 human reference genome coordinates with masked repeats. The previously defined three segmentally duplicated

regions in the human genome were included; The Rh Blood group genes (*RHCE/RHD*), the low-affinity FC gamma receptor, and the Beta-defensin repeat regions. There is a high sequence similarity within these regions. The breakpoints of these regions were previously studied, all containing copy number variable functionally important genes.

24 of 1000 Genome Project samples and the three-generation pedigree from Illumina's Platinum Genomes Project were chosen for mapping because the data was high coverage paired-end short reads and DNA samples were available for the wet lab validation of the computational pipeline. The short reads first were mapped against the reduced reference (RR) sequence using two different mapping tools (mrsFAST-Ultra and BWA-MEM). Then, the performances of the mapping tools were compared to each other. A correlation was inspected with the generated ratios of mapping results and previously found copy number of the gene clusters. According to Figure 3.5, 3.6, ad 3.7, the plots show that there is a consistency between the known copy numbers and the generated ratios for the BWA-MEM mapping compared to mrsFAST-Ultra mapping for both filtered and unfiltered version of data. However, it is shown that there is no clear clustering of copy numbers based on generated ratios for the Beta-defensin samples. For RHD and FcgRs region, the copy numbers were clustered better and a clear separation between copy number clusters can be observed. According to these results *RHD/RHCE* and Beta-Defensins repeat regions were not investigated further. As a result, BWA-MEM mapping results show a better performance compared to mrsFAST-Ultra even though mrsFAST-Ultra is designed to map short reads generated with the Illumina platform to reference genome assemblies. According to the results, the same mapping process was repeated with alternative reduced reference (ARR) by just using BWA-MEM mapping but mrsFAST-Ultra.

The robust PRT assay was applied to measure copy number variations of the low-affinity FcgR region for all the samples. The estimated copy numbers were found to be the same as previously found copy numbers for some samples from the previous studies (Hollox et al., 2008a; Handsaker et al. 2015). The copy numbers were also determined from the RR and ARR mapping. The copy numbers were needed because the sequence variation

was performed based on the copy numbers found for the low-affinity FcgR region. The copy number of most of the samples were predicted the same from the PRT assay, RR and ARR mapping. As displayed (Table 3.2 and 3.3), there are some disagreements between the mapping approaches and PRT assay for the samples HG01583, HG01879, NA12880, NA12882, NA12883. As a result, there is not enough evidence and nor samples to decide which mapping approach is better than the other. Trying to force all reads of a sample to map to a reduced reference does not work too well, at least for this region with a similarity of >97%, for some samples. It can be suggested that mapping should be performed for both paralogues of the segmentally duplicated regions and a wet lab validation method should be applied to estimate the copy number of a sample as final decision.

In order to validate the variants from the mapping analysis, the complete *FCGR2B* gene amplified and sequenced as a gold standard method. The variant calls were obtained from TVC and haplotype estimation for this samples were performed by BEAGLE. Especially pedigree samples were used for haplotyping because it increases the accuracy of imputation. FreeBayes was tested to see if it could identify the variants found by TVC on PCR samples. There were some differences between two variant callers regarding to the total number of variants and the way it is called. First, FreeBayes can successfully call the variants but it also detects the homopolymer calls as variants with low quality. It could not detect the AAAAT deletion in any sample (which is rs200504085 on dbSNP151 (rs3039548 on dbSNP152, October 11, 2018)). TVC can handle the low quality homopolymer calls due to its homopolymer error catcher. Even though TVC and FreeBayes predict the same variants without homopolymer calls, there are some differences regarding how a variant is called (Table 4.1). Homopolymer variant calls or the quality of the calling variants needs to be visually inspected to avoid any false positive variant calling. These variants should be carefully filtered before further analysis for FreeBayes and the indel-realigner of FreeBayes can be improved. If homopolymer calls are excluded, TVC and FreeBayes variant calls the same which indicates that FreeBayes might be a good candidate variant caller for a region that is copy number variable for the next analyses step.

FreeBayes was also used to call the variants of *FCGR2B* from the ARR mapping. The samples were treated as polyploid based on the copy number of the whole low-affinity FcgR region from the PRT assay. Not fully but partially the variants were predicted by FreeBayes. The reasons why some variants not detected by FreeBayes might be diverse. It could be due to the loss of the data during the quality filtering of the samples, low coverage after the mapping as well as the masked reduced reference regions. Besides, the differences between the variants' calls can also be because the NGS platforms perform sequencing in different ways.

The haplotype phasing of the diploid samples was effectively accomplished as seen in the pedigree data (Figure 4.4). However, construction of the haplotypes of the mapped samples (i.e. phasing of the polyploidy sequence data) was not successfully called by HapCompass (Table 4.4). HapCompass has been shown to be faster and significantly more accurate than HapCut and GATK using a variety of metrics on real and simulated data (Aguiar and Istrail, 2012). However, the reduced references were masked by repeats which result in lack of variant calls and therefore inaccurate haplotype phasing. Therefore, mapping short reads to a reduced reference is a limited approach.

Aa result, reduced reference mapping did not properly work for nucleotide variant calling for the duplicated regions used in this study. However, the PCR results of *FCGR2B* revealed more variants for this region compared to the 1000 genomes project.

7.2 A deeper analysis of the variants of FCGR2B

The FcgR receptors have very diverse functions and cellular expression. There is a presence of high sequence homology along with the existence of known segmental duplication in the low-affinity FcgR locus. Moreover, extensive single-nucleotide variation within and between the duplicated paralogs, and copy-number variation in this region make this region challenging to genotype and prevents making a clear prediction about the consequences. A deeper analysis as performed for the variants found specifically for the *FCGR2B* gene which encodes the only inhibitory receptor of FcgRs in order to inspect the possible effects of these variants.

According to VEP results, there are 17 variants that are novel within the 148 variants for specifically *FCGR2B* locus (Appendix 12 & 13). Most of the variants were found in introns (88%) for all the predicted consequences of the variants.While the 142 of the total variants have modifier impact (usually non-coding variants), four variants have lower impacts (harmless, unlikely to change protein behaviour). These are rs6665610, rs2298022, rs182968886 as synonymous variants, and rs2125684 as a splice region variant). No variants were found with high impact (disruptive impact in the protein, protein truncation or loss of function). Only three variants were found to have moderate impact as missense variants (non-disruptive variant that might change protein effectiveness); rs1050501, rs148534844, and the variant at the position NC_000001.10:g.161645052C>T. The rs1050501, and NC_000001.10:g.161645052C>T which are the only variants showing SIFT as tolerated (>0.8) and PolyPhen as benign (<0.25). None of the variants were found to be 'deleterious' for SIFT score and 'probably damaging' for PolyPhen score. One clinical significance state was seen on SNP rs1050501 that is the only variant showed as a risk factor/ leading to a protective function.

The five of GWAS SNPs were found in the list of predicted variants of *FCGR2B* gene. These GWAS SNPs showed high LD among themselves for the samples used in this study (Table 5.2). They were also investigated with the *FCGR2B v*ariants for any association. Several SNPs were found to be relatively in strong LD with GWAS SNPs as listed in Table 5.3 and Figure 5.3. The rs17413015 has a complete LD with other two SNPs rs12116547, rs1214584. The rs2480273, associated with birth weight, is also in complete LD with the same SNPs rs12116547, rs12145843. The rs17413015 has also complete LD with the variant 161644663.C>T. Thus, the rs17413015 and rs2480273 are in complete LD with the same SNPs rs12145843 and rs12116547 (Figure 5.3). rs1050501, (its polymorphism is associated with Malaria and SLE) was found to be in relatively strong LD with the rs6427615 which has been associated with total serum protein level. GWAS SNPs rs17413015 and rs72480273 were also found to be in strong LD with the SNPs rs12116547 and rs12145843 as both intron variant with modifier effect.

7.3 No clear relationship between *FCGR2B* gene conversion and expression.

The low affinity FCGR genes are subject to both single nucleotide polymorphisms and copy number variation (Breunis et al., 2009). In fact, several of these polymorphisms lead to an altered function of FcγRs (Blank et al, 2005). The rs3219018 in found on the promoter region of *FCGR2B*. The homozygosity for this -343C polymorphism reduces the transcription of *FCGR2B* in activated B cells and represents a susceptibility factor for the development of SLE (Blank et al, 2005). The surface expression of FcγRIIb receptors was significantly reduced in activated B cells from (–343 C/C) SLE patients. These findings suggest that genetic defects may lead to deregulated expression of the *FCGR2B* gene in –343 C/C homozygous subjects and may play a role in the pathogenesis of human SLE. The FcγRIIBT232 polymorphism of rs1050501 in humans is also associated with susceptibility to SLE. The increased frequency of FcγRIIBT232 in individuals of Southeast Asian and African descent may contribute to the increased prevalence and/or severity of SLE. The decreased FcγRIIB function provides a survival advantage against this disease and thus could explain the higher frequency of FcγRIIBT232 in Africans and Southeast Asians (Kwiatkowski, 2005; Smith and Clatworthy, 2010).

A PCR- based assay was designed to confirm gene conversion in the upstream region of the *FCGR2B* gene and it is analysed whether the gene conversion influences the expression of the gene. Different PCR assay were designed to detect the 9.1kb and 2.4kb gene conversion events. The gene conversion event was confirmed regardless of its size for the samples. The association between gene conversion and GWAS SNPS were investigated but very low LD found among them. There was no evidence complete monoallelic expression. Therefore, Using the SNPs that are present in the last two exons and the 3’ UTR of the *FCGR2B* gene, differentiation of alleles can be possible which means that it is possible to link the SNPs to the specific gene conversion events and measure allelic imbalance. The region covering the two SNPs (rs60519172, rs844) was sequenced and the expression of the variants were analysed for some samples. rs60519172 SNP was heterozygous on one sample so that It was not possible to draw

any robust conclusions. There is a higher expression of the allele with a G of rs844 for the NA18555 and NA19129 compared to NA18517 which both alleles carry gene conversion events and have equal expression levels. It might be suggested that expression of G is slightly higher when there is no gene conversion. Overall, no clear relationship between *FCGR2B* gene conversion and expression was found. This can also be because of on small sample size and limitations of lymphoblastoid cell lines.

7.4 No evidence for selection on *FCGR2B*

The analysis of DNA sequence polymorphisms and measuring genetic variation can provide an understanding into the evolutionary forces acting on *FCGR2B* gene across populations and species. The human *FCGR2B* locus was inspected from the perspective of genetic diversity, population genetics and signature of selection regarding the importance of SNP rs1050501 which variants has been found to be associated with susceptibility to SLE and protective function against malaria. Becuase the high mortality from malaria has resulted in the strongest known force for evolutionary selection in the recent history of the human genome (Willcocks et al., 2010).

The genetic diversity of this locus revealed high haplotype diversity (52 haplotypes, Hd:0.975) in five populations (EAS, SAS, EUR, AFR, AMR). The combination of high haplotype and low nucleotide diversity can suggest a signature of population expansion. However, The Tajima's D (-0.633) and Fu and Li's D* (-1.408) and F* (-1.317) statistics have p values being not significant P> 0.10 for the *FCGR2B* locus. These statistics failed to detect a significant departure from neutral expectations in this dataset. The McDonald-Kreitman test also failed to o reject that mutations are selectively neutral.

Gst values demonstrates low to level of genetic differentiation between the populations. The haplotype network shows that the haplotypes are very diverse and there is no clear clustering of haplotypes. Thus, rs1050501 does not seem to be responsible for selection and there is no evidence for selection at *FCGR2B* in humans in contrast to what might be predicted given its suggested role in malaria (Willcocks et al., 2010).

In addition, PopHuman genome browser statistics were not available for *FCGR2B* and the other genes of the low-affinity FcgRs. FCGR genes are characterized by high sequence similarity. There is a 92%–96% sequence homology among FCGR2A, FCGR2B and FCGR2C genes. Therefore, it is also suggested that the human genome reference and SNPs databases may contain inaccuracies for the FCGR region which have high homology in sequence (Hargreaves et al., 2015). The presence of high sequence homology in the FcgRs loci along with the existence of known segmental duplication and structural variations (CNV) may prevent the identification of specific SNPs in the FcgR gene complex as well as population statistic. Therefore, the population genetics analyses might also have been excluded for these regions/genes on the PopHuman genome browser.

There have been other studies in the literature that analysed repeated DNA regions in the human genome. Both Sudmant et al., (2010) and Forni et al., (2015) used short-read mapping depth approach and contributed to the sequence variation of copy number variable regions. In both study, samples from 1000 Genomes Project were used. The copy number prediction was performed in this study with the same method as done in Sudmant et al., (2010). Similar mapping tools (different versions of mrsFAST, in both study) and haplotype-based variant detector tools (FreeBayes in Forni et al., 2015) were also used in this study. The human reduced references in this study were repeat masked such happened in Sudmant et al., (2015). Lyons et al., (2016) also used a 'deduced short consensus region' with BWA mapping on short read sequencing and similar helper bioinformatics tools to investigate copy number variable *TPSAB1* gene. As happens in these other studies, thesis used a computational pipeline attempted to solve the read mapping ambiguity in short read sequencing and to investigate the sequence variation in the repeated DNA regions of the genome by using similar methods. However, the reduced reference mapping did not properly work for nucleotide variant calling for the duplicated regions. On the other hand, the PCR results of *FCGR2B* revealed more variants for this region compared to the 1000 genomes project. Both the reduced reference read approaches and the 1000 genomes variant calls did not call all variants found by the Ion torrent sequencing variant calls, with the 1000 Genomes variant calls

significantly underestimating and mis-genotyping samples. Several variants in *FCGR2B* were found to be in strong LD with variants previously associated with complex traits by GWAS. However, these GWAS variants were in weak LD with a gene conversion variant upstream of *FCGR2B*. The variation data *FCGR2B* was interrogated for signature of selection across global populations, and the genetic diversity of this locus revealed high haplotype diversity with 52 haplotypes. However, the population genetic statistics showed no evidence of natural selection at *FCGR2B*.

BIBLIOGRAPHY

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

Aguiar, D., & Istrail, S. (2013). Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics*, 29(13), 352-360.

Ahmed, S., Usman, M., Ferzund, J., Atif, M., Rehman, A., & Mehmood, A. (2017). Modern Data Formats for Big Bioinformatics Data Analytics. *International Journal of Advanced Computer Science and Applications*, 8(4), 366-377.

Aigner, J., Villatoro, S., Rabionet, R., Roquer, J., Jiménez-Conde, J., Martí, E., & Estivill, X. (2013). A common 56-kilobase deletion in a primate-specific segmental duplication creates a novel butyrophilin-like protein. *BMC Genetics*, 14(61).

Aitman, T. J., Dong, R., Vyse, T. J., Norsworthy, P. J., Johnson, M. D., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A. J., Petretto, E., Hodges, M. D., Bhangal, G., Patel, S. G., Sheehan-Rooney, K., Duda, M., Cook, P. R., Evans, D. J., Domin, J., Flint, J., … Cook, H. T. (2006). Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, 439(7078), 851–855.

Aldhous, M. C., Bakar, S. A., Prescott, N. J., Palla, R., Soo, K., Mansfield, J. C., Mathew, C. G., Satsangi, J., & Armour, J. A. L. (2010). Measurement methods and accuracy in copy number variation: Failure to replicate associations of beta-defensin copy number with Crohn's disease. *Human Molecular Genetics*, 19(24), 4930–8.

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. In *Nature Reviews Genetics*. 12(5), 363–376.

Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E. B., & Müller-Myhsok, B. (2012). A beginner's guide to SNP calling from high-Throughput DNA-sequencing data. In *Human Genetics*, 131(10), 1541-1554.

Ambardar, S., Gupta, R., Trakroo, D., Lal, R., & Vakhlu, J. (2016). High Throughput Sequencing: An Overview of Sequencing Chemistry. In *Indian Journal of Microbiology*, 56 (4), 394-404.

Andrews, S. (1973). Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. In *Soil*, 5(1), 47-81.

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology*, 25(4), 195-203.

Arlt, M. F., Wilson, T. E., & Glover, T. W. (2012). Replication stress and mechanisms of CNV formation. In *Current Opinion in Genetics and Development*, 22(3), 204-10.

Armour, J. A. L., Palla, R., Zeeuwen, P. L. J. M., Heijer, M. Den, Schalkwijk, J., & Hollox, E. J. (2007). Accurate, high throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Research*, 35(3), e19.

Avent, N. D., & Reid, M. E. (2000). The Rh blood group system: A review. In *Blood*, 95(2), 375-387.

Bailey, J. A., Liu, G., & Eichler, E. E. (2003). An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications. *American Journal of Human Genetics*, 73(4), 823-834.

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., & Eichler, E. E. (2002). Recent segmental duplications in the human genome. *Science*, 297(5583), 1003-1007.

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

Bandelt, H. J., Forster, P., & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16(1), 37-48.

Barber, J. C. K. (2005). Directly transmitted unbalanced chromosome abnormalities and euchromatic variants. In *Journal of Medical Genetic*, 42(8), 609–629.

Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D., & Hurles, M. E. (2008). A robust statistical method for case-control association testing with copy number variation. *Nature Genetics*, 40(10), 1245-1252.

Bateman, M. S., Mehta, S. G., Willatt, L., Selkirk, E., Bedwell, C., Zwolinski, S., Sparnon, L., Simonic, I., Abbott, K., & Barber, J. C. K. (2010). A de novo 4q34 interstitial deletion of at least 9.3 Mb with no discernible phenotypic effect. *American Journal of Medical Genetics, Part A*. 152(7), 1764–1769.

Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L., & Trees, E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. In *Clinical Microbiology and Infection*, 24(4), 335-341.

Blank, M. C., Stefanescu, R. N., Masuda, E., Marti, F., King, P. D., Redecha, P. B., Wurzburger, R. J., Peterson, M. G. E., Tanaka, S., & Pricop, L. (2005). Decreased transcription of the human FCGR2B gene mediated by the -343 G/C promoter polymorphism and association with systemic lupus erythematosus. *Human Genetics*, 117(2-3), 220-227.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120.

Bonatti, F., Adorni, A., Percesepe, A., & Martorana, D. (2018). Discrimination of FCGR2B polymorphism without coamplification of FCGR2A and FCGR2C genes. *International Journal of Immunogenetics*, 45(1), 22-25.

Boruchov, A. M., Heller, G., Veri, M. C., Bonvini, E., Ravetch, J. V., & Young, J. W. (2005). Activating and inhibitory IgG Fc receptors on human DCs mediate opposing functions. *Journal of Clinical Investigation*, 115(10), 2914-2923.

Breunis, W. B., Van Mirre, E., Geissler, J., Laddach, N., Wolbink, G., Van Schoot, E. Der, De Haas, M., De Boer, M., Roos, D., & Kuijpers, T. W. (2009). Copy number variation at the FCGR locus includes FCGR3A, FCGR2C and FCGR3B but not FCGR2A and FCGR2B. *Human Mutation*, 30(5), E640-650.

Brouwers, N., Van Cauwenberghe, C., Engelborghs, S., Lambert, J. C., Bettens, K., Le Bastard, N., Pasquier, F., Montoya, A. G., Peeters, K., Mattheijssens, M., Vandenberghe, R., De Deyn, P. P., Cruts, M., Amouyel, P., Sleegers, K., & Van Broeckhoven, C. (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.

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: Existing methods and new developments. In *Nature Reviews Genetics*, 12(10), 703–714.

Browning, B. L., & Browning, S. R. (2008). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, 84(2), 210–223.

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81(5), 1084-1097.

Bruhns, P., Iannascoli, B., England, P., Mancardi, D. A., Fernandez, N., Jorieux, S., & Daëron, M. (2009). Specificity and affinity of human Fcγ receptors and their polymorphic variants for human IgG subclasses. *Blood*, 113(16), 3716-3125.

Butcher, G. (2008) Autoimmunity and malaria. *Trends Parasitol*, 24(7), 291-292.

Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1842 (10), 1932-1941.

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12), e1002822.

Buysse, K., Delle Chiaie, B., Van Coster, R., Loeys, B., De Paepe, A., Mortier, G., Speleman, F., & Menten, B. (2009). Challenges for CNV interpretation in clinical molecular karyotyping: Lessons learned from a 1001 sample experience. *European Journal of Medical Genetics*, 52(6):398–403.

Cantsilieris, S., Baird, P. N., & White, S. J. (2013). Molecular methods for genotyping complex copy number polymorphisms. In *Genomics*, 101(2), 86–93.

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

Carr, I. M., Robinson, J. I., Dimitriou, R., Markham, A. F., Morgan, A. W., & Bonthron, D. T. (2009). Inferring relative proportions of DNA variants from sequencing electropherograms. *Bioinformatics*, 25(24), 3244-3250.

Carreto, L., Eiriz, M. F., Gomes, A. C., Pereira, P. M., Schuller, D., & Santos, M. A. S. (2008). Comparative genomics of wild type yeast strains unveils important genome diversity. *BMC Genomics*, 9, 524.

Casillas, S., Mulet, R., Villegas-Mirón, P., Hervas, S., Sanz, E., Velasco, D., Bertranpetit, J., Laayouni, H., & Barbadilla, A. (2018). PopHuman: The human population genomics browser. *Nucleic Acids Research*, 46(D1), D1003-D1010.

Chen, J. M., Cooper, D. N., Férec, C., Kehrer-Sawatzki, H., & Patrinos, G. P. (2010a). Genomic rearrangements in inherited disease and cancer. *Seminars in Cancer Biology*, 20(4), 222-223.

Chen, Y., Cunningham, F., Rios, D., McLaren, W. M., Smith, J., Pritchard, B., Spudich, G. M., Brent, S., Kulesha, E., Marin-Garcia, P., Smedley, D., Birney, E., & Flicek, P. (2010b). Ensembl variation resources. *BMC Genomics*, 11(1), 293.

Chen, W. K., Swartz, J. D., Rush, L. J., & Alvarez, C. E. (2009). Mapping DNA structural variation in dogs. *Genome Research*, 19(3), 500–509.

Cheung, J., Estivill, X., Khaja, R., MacDonald, J. R., Lau, K., Tsui, L. C., & Scherer, S. W. (2003). Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biology*, 4(4), R25.

Cho, Y., Lee, S., Hong, J. H., Kim, B. J., Hong, W. Y., Jung, J., Lee, H. B., Sung, J., Kim, H. N., Kim, H. L., & Jung, J. (2018). Development of the variant calling algorithm, ADIScan, and its use to estimate discordant sequences between monozygotic twins. *Nucleic Acids Research*, 46(15), e92.

Coe, B. P., Girirajan, S., & Eichler, E. E. (2012). The genetic variability and commonality of neurodevelopmental disease. *American Journal of Medical Genetics, Part C: Seminars in Medical Genetics*, 160(2), 118–129.

Colin, Y., Cherif-Zahar, B., Le Van Kim, C., Raynal, V., Van Huffel, V., & Cartron, J. P. (1991). Genetic basis of the RhD-positive and RhD-negative blood group polymorphism as determined by southern analysis. *Blood*, 78(10), 2747–2752.

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

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E., & Pritchard, J. K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nature, Genetics*, 38(1), 75–81.

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., Abdel-Hamid, H., Bader, P., McCracken, E., Niyazov, D., Leppig, K., Thiese, H., Hummel, M., Alexander, N., Gorski, J., Kussmann, J., Shashi,

V., Johnson, K., Rehder, C., Ballif, B.C., Shaffer, L,G,, Eichler, E. E. (2011). A copy number variation morbidity map of developmental delay. *Nature Genetics*, 43(9): 838–846.

De Cid, R., Riveira-Munoz, E., Zeeuwen, P. L. J. M., Robarge, J., Liao, W., Dannhauser, E. N., Giardina, E., Stuart, P. E., Nair, R., Helms, C., Escaramís, G., Ballana, E., Martín-Ezquerra, G., Heijer, M. Den, Kamsteeg, M., Joosten, I., Eichler, E. E., Lázaro, C., Pujol, R. M., … Estivill, X. (2009). Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nature Genetics*, 41(2), 211-215.

Diskin, S. J., Hou, C., Glessner, J. T., Attiyeh, E. F., Laudenslager, M., Bosse, K., Cole, K., Mossé, Y. P., Wood, A., Lynch, J. E., Pecor, K., Diamond, M., Winter, C., Wang, K., Kim, C., Geiger, E. A., McGrady, P. W., Blakemore, A. I. F., London, W. B., … Maris, J. M. (2009). Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*, 459(7249), 987–991.

Dopman, E. B., & Hartl, D. L. (2007). A portrait of copy-number polymorphism in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50), 19920–19925.

Ebbert, M. T. W., Jensen, T. D., Jansen-West, K., Sens, J. P., Reddy, J. S., Ridge, P. G., Kauwe, J. S. K., Belzil, V., Pregent, L., Carrasquillo, M. M., Keene, D., Larson, E., Crane, P., Asmann, Y. W., Ertekin-Taner, N., Younkin, S. G., Ross, O. A., Rademakers, R., Petrucelli, L., & Fryer, J. D. (2019). Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biology*, 20(1), 97.

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., Duce, J., Kauwe, J. S. K., & Ridge, P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, *17*(Suppl 7), 239.

Ebert, G., Steininger, A., Weißmann, R., Boldt, V., Lind-Thomsen, A., Grune, J., Badelt, S., Heßler, M., Peiser, M., Hitzler, M., Jensen, L. R., Müller, I., Hu, H., Arndt, P. F., Kuss, A. W., Tebel, K., & Ullmann, R. (2014). Distribution of segmental duplications in the context of higher order chromatin organisation of human chromosome 7. *BMC Genomics*, 15(1), 537.

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: Description, history, and methods to detect structural variation. *Briefings in Functional Genomics*, 14(5), 305-314.

Fanciulli, M., Vyse, T. J., & Aitman, T. J. (2009). Copy number variation of Fc gamma receptor genes and disease predisposition. In *Cytogenetic and Genome Research*, 123(1-4):161-8.

Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J. M., Gough, S. C. L., De Smith, A., Blakemore, A. I. F., Froguel, P., Owen, C. J., Pearce, S. H. S., Teixeira, L., Guillevin, L., Graham, D. S. C., Pusey, C. D., Cook, H. T., Vyse, T. J., & Aitman, T. J. (2007). FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature Genetics*, 39(6), 721–723.

Fellermann, K., Stange, D. E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C. L., Reinisch, W., Teml, A., Schwab, M., Lichter, P., Radlwimmer, B., & Stange, E. F. (2006).

A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *American Journal of Human Genetics*, 79(3), 439–448.

Fernando, M. M. A., Boteva, L., Morris, D. L., Zhou, B., Wu, Y. L., Lokki, M. L., Yu, C. Y., Rioux, J. D., Hollox, E. J., & Vyse, T. J. (2010). Assessment of complement C4 gene copy number using the paralog ratio test. *Human Mutation*, 31(7), 866-874.

Ferragina, P., & Manzini, G. (2000). Opportunistic data structures with applications. *Annual Symposium on Foundations of Computer Science - Proceedings*. 390-398.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. In *Nature Reviews Genetics*, 7(2), 85–97.

Fiegler, H., Redon, R., Andrews, D., Scott, C., Andrews, R., Carder, C., Clark, R., Dovey, O., Ellis, P., Feuk, L., French, L., Hunt, P., Kalaitzopoulos, D., Larkin, J., Montgomery, L., Perry, G. H., Plumb, B. W., Porter, K., Rigby, R. E., … Carter, N. P. (2006). Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Research*, 16(12), 1566–1574.

Field, S. F., Howson, J. M. M., Maier, L. M., Walker, S., Walker, N. M., Smyth, D. J., Armour, J. A. L., Clayton, D. G., & Todd, J. A. (2009). Experimental aspects of copy number variant assays at CCL3L1. In *Nature Medicine*, 15(10), 1115–1117.

Filges, I., Röthlisberger, B., Noppen, C., Boesch, N., Wenzel, F., Necker, J., Binkert, F., Huber, A. R., Heinimann, K., & Miny, P. (2009). Familial 14.5 Mb interstitial deletion 13q21.1-13q21.33: Clinical and array-CGH study of a benign phenotype in a three-generation family. *American Journal of Medical Genetics, Part A*. 149(2), 237–241.

Flegel, W. A. (2011). Molecular genetics and clinical applications for RH. *Transfusion and Apheresis Science*, 44(1), 81–91.

Flicek, P., & Birney, E. (2009). Sense from sequence reads: Methods for alignment and assembly. *Nature Methods*, 6(11S), S6.

Floto, R. A., Clatworthy, M. R., Heilbronn, K. R., Rosner, D. R., MacAry, P. A., Rankin, A., Lehner, P. J., Ouwehand, W. H., Allen, J. M., Watkins, N. A., & Smith, K. G. C. (2005). Loss of function of a lupus associated FcγRIIb polymorphism through exclusion from lipid rafts. *Nature Medicine*, 11(10), 1056-1058.

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

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R., & Futreal, P. A. (2011). COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 39, D945–950.

Forni, D., Martin, D., Abujaber, R., Sharp, A. J., Sironi, M., & Hollox, E. J. (2015). Determining multiallelic complex copy number and sequence variation from high coverage exome sequencing data. *BMC Genomics*, 16(1), 891.

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

Freeman, P. J., Hart, R. K., Gretton, L. J., Brookes, A. J., & Dalgleish, R. (2018). VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions. *Human Mutation*, 39(1), 61-68.

Fu, Y. X., & Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133(3), 693-709.

Fuchs, J., Nilsson, C., Kachergus, J., Munz, M., Larsson, E. M., Schüle, B., Langston, J. W., Middleton, F. A., Ross, O. A., Hulihan, M., Gasser, T., & Farrer, M. J. (2007). Phenotypic variation in a large Swedish pedigree due to SNCA duplication and triplication. *Neurology*, 68(12), 916–922.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing -- Free bayes -- Variant Calling -- Longranger. *ArXiv Preprint ArXiv:1207.3907*.

Getahun, A., & Cambier, J. C. (2015). Of ITIMs, ITAMs, and ITAMis: Revisiting immunoglobulin Fc receptor signaling. In *Immunological Reviews*, 268(1), 66-73.

Girirajan, S., Campbell, C. D., & Eichler, E. E. (2011). Human Copy Number Variation and Complex Genetic Disease. *Annual Review of Genetics*, 45(1), 203-26.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J., Murthy, K. K., Rovin, B. H., Bradley, W., Clark, R. A., Anderson, S. A., O'Connell, R. J., Agan, B. K., Ahuja, S. S., Bologna, R., … Ahuja, S. K. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, 307(5714), 1434–1440.

Graubert, T. A., Cahan, P., Edwin, D., Selzer, R. R., Richmond, T. A., Eis, P. S., Shannon, W. D., Li, X., McLeod, H. L., Cheverud, J. M., & Ley, T. J. (2007). A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genetics*, 3(1), 21-29.

Griffin, D. K., Robertson, L. B., Tempest, H. G., Vignal, A., Fillon, V., Crooijmans, R. P. M. A., Groenen, M. A. M., Deryusheva, S., Gaginskaya, E., Carré, W., Waddington, D., Talbot, R., Völker, M., Masabanda, J. S., & Burt, D. W. (2008). Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution. *BMC Genomics*, 9, 168.

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

Gu, W., Zhang, F., & Lupski, J. R. (2008). Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(4).

Guo, T., Mcdonald-Mcginn, D., Blonska, A., Shanske, A., Bassett, A. S., Chow, E., Bowser, M., Sheridan, M., Beemer, F., Devriendt, K., Swillen, A., Breckpot, J., Digilio, M. C., Marino, B., Dallapiccola, B., Carpenter, C., Zheng, X., Johnson, J., Chung, J., … Morrow, B. (2011). Genotype and cardiovascular phenotype correlations with TBX1 in 1,022 velo-cardio-facial/digeorge/22q11.2 deletion syndrome patients. *Human Mutation*, 32(11), 1278-89.

Hach, F., Sarrafi, I., Hormozdiari, F., Alkan, C., Eichler, E. E., & Sahinalp, S. C. (2014). MrsFAST-Ultra: A compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Research*, 42(W1), W494-W500.

Han, E., Sinsheimer, J. S., & Novembre, J. (2015). Fast and accurate site frequency spectrum estimation from low coverage sequence data. *Bioinformatics*, 31(5), 720–727.

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., & Mccarroll, S. A. (2015). Large multiallelic copy number variations in humans. *Nature Genetics*, 47(3), 296–303.

Hargreaves, C. E., Rose-Zerilli, M. J. J., Machado, L. R., Iriyama, C., Hollox, E. J., Cragg, M. S., & Strefford, J. C. (2015). Fcγ receptors: Genetic variation, function, and disease. In *Immunological Reviews*, 268(1), 6-24.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009) Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8), 551–64.

Haynes, E., Jimenez, E., Pardo, M. A., & Helyar, S. J. (2019). The future of NGS (Next Generation Sequencing) analysis in testing food authenticity. *Food Control*, 101, 134-143.

Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14(1), 184.

Head, S. R., Kiyomi Komori, H., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, 56(3), 61-77.

Hollox, E. J., Barber, J. C. K., Brookes, A. J., & Armour, J. A. L. (2008a). Defensins and the dynamic genome: What we can learn from structural variation at human chromosome band 8p23.1. In *Genome Research*, 18(11), 1686-1697.

Hollox, E. J., Huffmeier, U., Zeeuwen, P. L. J. M., Palla, R., Lascorz, J., Rodijk-Olthuis, D., Van De Kerkhof, P. C. M., Traupe, H., De Jongh, G., Heijer, M. Den, Reis, A., Armour, J. A. L., & Schalkwijk, J. (2008b). Psoriasis is associated with increased β-defensin genomic copy number. *Nature Genetics*, 40(1), 23–25.

Hollox, E. J., Detering, J. C., & Dehnugara, T. (2009). An integrated approach for measuring copy number variation at the FCGR3 (CD16) locus. *Human Mutation*, 30(3), 477-484.

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

Hosseini, M., Pratas, D., & Pinho, A. J. (2016). A survey on data compression methods for biological sequences. *Information*, 7(4), 56.

Itsara, A., Cooper, G. M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R. M., Myers, R. M., Ridker, P. M., Chasman, D. I., Mefford, H., Ying, P., Nickerson, D. A., & Eichler, E. E. (2008). Population analysis of large copy number variants and hotspots of human genetic disease. *American Journal of Human Genetics*, 84(2), 148–161.

Iyer, J., & Girirajan, S. (2015). Gene discovery and functional assessment of rare copy-number variants in neurodevelopmental disorders. *Briefings in Functional Genomics*, 14(5):315–328.

Kauppi, L., May, C. A., & Jeffreys, A. J. (2009). Analysis of meiotic recombination products from human sperm. *Methods in Molecular Biology*, 557, 323-355.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tüzün, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K.A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J.D., Korn, J.M., McCarroll, S.A., Altshuler, D.A., Peiffer, D.A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D.A., Mullikin, J.C., Wilson, R.K., Bruhn, L., Olson, M.V., Kaul, R., Smith, D.R., Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191), 56-64.

Kim, P. M., Lam, H. Y. K., Urban, A. E., Korbel, J. O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., & Gerstein, M. B. (2008). Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Research*, *18*(12), 1865–1874.

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing - Concepts and limitations. In *BioEssays*, 32(6), 524-536.

Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K. T., Jonasdottir, A., Frigge, M. L., Gylfason, A., Olason, P. I., Gudjonsson, S. A., Sverrisson, S., Stacey, S. N., Sigurgeirsson, B., Benediktsdottir, K. R., Sigurdsson, H., Benediktsdottir, K.R., Sigurdsson, H., Jonsson, T., Benediktsson, R., Olafsson, J.H., Johannsson, O.T., Hreidarsson, A.B., Sigurdsson, G., DIAGRAM Consortium, Ferguson-Smith, A.C., Gudbjartsson, D.F., Thorsteinsdottir, U., Stefansson, K. (2009). Parental origin of sequence variants associated with complex diseases. *Nature*, 462(7275), 868-74.

Kono, H., Kyogoku, C., Suzuki, T., Tsuchiya, N., Honda, H., Yamamoto, K., Tokunaga, K., & Honda, Z. I. (2005). FcγRIIB Ile232Thr transmembrane polymorphism associated with human systemic lupus erythematosus decreases affinity to lipid rafts and attenuates inhibitory effects on B cell receptor signalling. *Human Molecular Genetics*, 14(19), 2881-2892.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C. E., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T., Gerstein, M.,

Egholm, M., Snyder, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849), 420–426.

Kwiatkowski, D. P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. In *American Journal of Human Genetics*, 77(2), 171-192.

Kyogoku, C., Dijstelbloem, H. M., Tsuchiya, N., Hatta, Y., Kato, H., Yamaguchi, A., Fukazawa, T., Jansen, M. D., Hashimoto, H., Van De Winkel, J. G. J., Kallenberg, C. G. M., & Tokunaga, K. (2002). Fcγ receptor gene polymorphisms in Japanese patients with systemic lupus erythematosus: Contribution of FCGR2B to genetic susceptibility. *Arthritis and Rheumatism*, 46(5), 1242-1254.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1), D980–985.

Langhorne, J., Ndungu, F. M., Sponaas, A. M., & Marsh, K. (2008). Immunity to malaria: More questions than answers. In *Nature Immunology*, 9(7), 725–732.

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., Paschall, J., Ananiev, V., Flicek, P., & Church, D. M. (2013). DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Research*, 41(D1), D936-41.

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*., 15(6), R84.

Lee, J. A., Carvalho, C. M. B., & Lupski, J. R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, 131(7), 1235-47.

Lee, J. H., & Jeon, J. T. (2009). Methods to detect and analyze copy number variations at the genome-wide and locus-specific levels. In *Cytogenetic and Genome Research*, 123(14), 333-342.

Lehmann, B., Schwab, I., Bhm, S., Lux, A., Biburger, M., & Nimmerjahn, F. (2012). FcγRIIB: A modulator of cell activation and humoral tolerance. In *Expert Review of Clinical Immunology*, 8(3), 243–254.

Li, H. (2013). [Heng Li - Compares BWA to other long read aligners like CUSHAW2] Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Preprint ArXiv*.1303.3997. Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA (00) 00:1-3.

Li, W., & Olivier, M. (2013). Current analysis platforms and methods for detecting copy number variation. *Physiological Genomics*, 45(1), 1–16.

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Research*, 21(6), 940–951.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5), 589-595.

Lieber, M. R. (2008). The mechanism of human nonhomologous DNA End joining. *Journal of Biological Chemistry*, *283*(1), 1–5.

Linardopoulou, E. V., Williams, E. M., Fan, Y., Friedman, C., Young, J. M., & Trask, B. J. (2005). Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*, 437(7055), 94–100.

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. In *Genomics, Proteomics and Bioinformatics*, 14(5), 265-279.

Ma, H., & DiFazio, S. (2008). An efficient method for purification of PCR products for sequencing. *BioTechniques*, 44(7), 921-3.

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42(Database issue), D986–D992.

Machado, L. R., & Ottolini, B. (2015). An evolutionary history of defensins: A role for copy number variation in maximizing host innate and adaptive immune responses. In *Frontiers in Immunology*, 6(115).

MacHado, L. R., Hardwick, R. J., Bowdrey, J., Bogle, H., Knowles, T. J., Sironi, M., & Hollox, E. J. (2012). Evolutionary history of copy-number-variable locus for the low-affinity Fcγ receptor: Mutation rate, autoimmune disease, and the legacy of helminth infection. *American Journal of Human Genetics*, 90(6), 973–985.

Magi, A., Tattini, L., Pippucci, T., Torricelli, F., & Benelli, M. (2012). Read count approach for DNA copy number variants detection. *Bioinformatics*, 28(4), 470-478.

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7), 906–913.

Mardulyn, P. (2012). Trees and/or networks to display intraspecific DNA sequence variation. *Molecular ecology*, 21(14), 3385-90.

Maydan, J. S., Lorch, A., Edgley, M. L., Flibotte, S., & Moerman, D. G. (2010). Copy number variation in the genomes of twelve natural isolates of Caenorhabditis elegans. *BMC Genomics*, 11(1), 62.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., De Bakker, P. I. W., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K.W., Rava, R., Daly, M.J., Gabriel, S.B., Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, 40(10), 1166-1174.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328), 652-654.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1), 122.

Medvedev, P., Fiume, M., Dzamba, M., Smith, T., & Brudno, M. (2010). Detecting copy number variation with mated short reads. *Genome Research*, 20(11), 1613–1622.

Miller, D. W., Hague, S. M., Clarimon, J., Baptista, M., Gwinn-Hardy, K., Cookson, M. R., & Singleton, A. B. (2004). α-synuclein in blood and brain from familial Parkinson disease with SNCA locus triplication. *Neurology*, 62(10), 1835–1838.

Montavon, T., Thevenet, L., & Duboule, D. (2012). Impact of copy number variations (CNVs) on long-range gene regulation at the HoxD locus. *Proceedings of the National Academy of Sciences of the United States of America*, 109(50), 20204-20211.

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., Lee, J., Chu, C., Lin, C., Dzakula, Z., Cao, H., Schlebusch, S. A., Giorda, K., Schnall-Levin, M., Wall, J. D., & Kwok, P. Y. (2016). A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods*, 13(7), 587-590.

Mueller, M., Barros, P., Witherden, A. S., Roberts, A. L., Zhang, Z., Schaschl, H., Yu, C. Y., Hurles, M. E., Schaffner, C., Floto, R. A., Game, L., Steinberg, K. M., Wilson, R. K., Graves, T. A., Eichler, E. E., Cook, H. T., Vyse, T. J., & Aitman, T. J. (2013). Genomic pathology of sle-associated copy-number variation at the FCGR2C/FCGR3B/FCGR2B locus. *American Journal of Human Genetics*, 92(1), 28-40.

Nagelkerke, S. Q., Schmidt, D. E., de Haas, M., & Kuijpers, T. W. (2019). Genetic Variation in Low-To-Medium-Affinity Fcγ Receptors: Functional Consequences, Disease Associations, and Opportunities for Personalized Medicine. In *Frontiers in Immunology*, 10(2237).

Nicholas, T. J., Cheng, Z., Ventura, M., Mealey, K., Eichler, E. E., & Akey, J. M. (2009). The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Research*, 19(3), 491–499.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. In *Nature Reviews Genetics*. 12(6), 443-451.

Niederer, H. A., Willcocks, L. C., Rayner, T. F., Yang, W., Lau, Y. L., Williams, T. N., Scott, J. A. G., Urban, B. C., Peshu, N., Dunstan, S. J., Hien, T. T., Phu, N. H., Padyukov, L., Gunnarsson, I., Svenungsson, E., Savage, C. O., Watts, R. A., Lyons, P. A., Clayton, D. G., & Smith, K. G. C. (2010). Copy number, linkage disequilibrium and disease association in the FCGR locus. *Human Molecular Genetics*, 19(16), 3282–3294.

Nimmerjahn, F., Gordan, S., & Lux, A. (2015). FcγR dependent mechanisms of cytotoxic, agonistic, and neutralizing antibody activities. In *Trends in Immunology*, 36(6):325-36.

Nimmerjahn, F., & Ravetch, J. V. (2008). Fcγ receptors as regulators of immune responses. In *Nature Reviews Immunology*, 8(1):34-47.

Nimmerjahn, F., & Ravetch, J. V. (2006). Fcγ receptors: Old friends and new family members. In *Immunity*, 24(1), 19-28.

Nowakowska, B. (2017). Clinical interpretation of copy number variants in the human genome. *Journal of Applied genetics*, *58*(4), 449–457.

Nuttle, X., Huddleston, J., O'roak, B. J., Antonacci, F., Fichera, M., Romano, C., Shendure, J., & Eichler, E. E. (2013). Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. *Nature Methods*, 10(9), 909-909.

Numanagić, I., Gökkaya, A. S., Zhang, L., Berger, B., Alkan, C., & Hach, F. (2018). Fast characterization of segmental duplications in genome assemblies. *Bioinformatics*, 34(17), i706-i714.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E., Wei, Z., Wang, K., & Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. *Genome Medicine*, 5(3), 28.

Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., & Weigel, D. (2008). Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Research*. 18(12), 2024–2033.

Ottolini, B., Hornsby, M. J., Abujaber, R., MacArthur, J. A. L., Badge, R. M., Schwarzacher, T., Albertson, D. G., Bevins, C. L., Solnick, J. V., & Hollox, E. J. (2014). Evidence of convergent evolution in humans and macaques supports an adaptive role for copy number variation of the β-defensin-2 gene. *Genome Biology and Evolution*, 6(11):3025-3038.

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L., & Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. In *Genome Biology*, 11(5), R52.

Park, L. (2012). Linkage Disequilibrium Decay and Past Population History in the Human Genome. *PLoS ONE*, 7(10): e46603.

Park, S. S., Stankiewicz, P., Bi, W., Shaw, C., Lehoczky, J., Dewar, K., Birren, B., & Lupski, J. R. (2002). Structure and evolution of the Smith-Magenis syndrome repeat gene clusters, SMS-REPs. *Genome Research*, 12(5), 729–738.

Patel, R. K., & Jain, M. (2012). NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE*, 7(2).

Pazgier, M., Hoover, D. M., Yang, D., Lu, W., & Lubkowski, J. (2006). Human β-defensins. In *Cellular and Molecular Life Sciences*, 63(11):1294–313.

Perne, A., Zhang, X. H., Lehmann, L. E., Groth, M., Stuber, F., & Book, M. (2009). Comparison of multiplex ligation-dependent probe amplification and real-time PCR accuracy for gene copy number quantification using the β-defensin locus. *BioTechniques*, 47(6), 1023-1027.

Perry, G. H., Xue, Y., Smith, R. S., Meyer, W. K., Çalışkan, M., Yanez-Cuna, O., Lee, A. S., Gutiérrez-Arcelus, M., Ober, C., Hollox, E. J., Tyler-Smith, C., & Lee, C. (2012). Evolutionary genetics of the human Rh blood group system. In *Human Genetics*, *131*(7), 1205–1216.

Perry, G. H., Tchinda, J., McGrath, S. D., Zhang, J., Picker, S. R., Cáceres, A. M., Iafrate, A. J., Tyler-Smith, C., Scherer, S. W., Eichler, E. E., Stone, A. C., & Lee, C. (2006). Hotspots for copy number variation in chimpanzees and humans. *Proceedings of the National Academy of Sciences of the United States of America*, 103(21), 8006–8011.

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A. C., Thiruvahindrapuram, B., MacDonald, J. R., Mills, R., Prasad, A., Noonan, K., Gribble, S., Prigmore, E., Donahoe, P. K., Smith, R. S., Park, J. H., Hurles, M. E., Carter, N. P., E. Prigmore, P.K. Donahoe, R.S. Smith, J.H. Park, M.E. Hurles, N.P. Carter, C. Lee,S.W. Scherer, Feuk, L. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology*, 29(6), 512–520.

Pirooznia, M., Goes, F., & Zandi, P. P. (2015). Whole-genome CNV analysis: Advances in computational approaches. *Frontiers in Genetics*, 6:138.

Radstake, T. R. D. J., Franke, B., Wenink, M. H., Nabbe, K. C. A. M., Coenen, M. J. H., Welsing, P., Bonvini, E., Koenig, S., Van Den Berg, W. B., Barrera, P., & Van Riel, P. L. C. M. (2006). The functional variant of the inhibitory Fcγ receptor IIb (CD32B) is associated with the rate of radiologic joint damage and dendritic cell function in rheumatoid arthritis. *Arthritis and Rheumatism*, 54(12), 3828–3837.

Rahbari, R., Zuccherato, L. W., Tischler, G., Chihota, B., Ozturk, H., Saleem, S., Tarazona-Santos, E., Machado, L. R., & Hollox, E. J. (2016). Understanding the Genomic Structure of Copy-Number Variation of the Low-Affinity Fcγ Receptor Region Allows Confirmation of the Association of FCGR3B Deletion with Rheumatoid Arthritis. *Human Mutation*, 38 (4): 390–399.

Ravetch, J. V., & Lanier, L. L. (2000). Immune inhibitory receptors. In *Science*, 290(5489), 84–89.

Ravetch, J. V., & Perussia, B. (1989). Alternative membrane forms of FcγRIII(CD16) on human natural killer cells and neutrophils. Cell type-specific expression of two genes that differ in single nucleotide substitutions. *Journal of Experimental Medicine*, 170(2), 481–497.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, 444 (7118), 444-454.

Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of Next-Generation Sequencing Reads. *Annual Review of Genomics and Human Genetics*, 16(1), 133-151.

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*. 13(5), 278-289.

Rios, D., McLaren, W. M., Chen, Y., Birney, E., Stabenau, A., Flicek, P., & Cunningham, F. (2010). A database and API for variation, dense genotyping and resequencing data. *BMC Bioinformatics*, 11, 238.

Rogers, A.R., & Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Moleculer Biology and Evolution*, 9(3), 552-569.

Roghanian, A., Stopforth, R. J., Dahal, L. N., & Cragg, M. S. (2018). New revelations from an old receptor: Immunoregulatory functions of the inhibitory Fc gamma receptor, FcγRIIB (CD32B). In *Journal of Leukocyte Biology*, 103(6), 1077-1088.

Rosales, C., & Uribe-Querol, E. (2013). Fc receptors: Cell activators of antibody functions. *Advances in Bioscience and Biotechnology*, 4(4), 21-33.

Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sanchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution*, 34(12), 3299-3302.

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerrière, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M., Dubas, F., Frebourg, T., & Campion, D. (2006). APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genetics*, 38(1), 24-26.

Ruffalo, M., Laframboise, T., & Koyutürk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20), 2790-2796.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832-837.

Samarakoon, P. S., Sorte, H. S., Kristiansen, B. E., Skodje, T., Sheng, Y., Tjønnfjord, G. E., Stadheim, B., Stray-Pedersen, A., Rødningen, O. K., & Lyle, R. (2014). Identification of copy number variants from exome sequence data. *BMC Genomics*, 15(661).

Samonte, R. V., & Eichler, E. E. (2002). Segmental duplications and the evolution of the primate genome. In *Nature Reviews Genetics*, 3(1), 65–72.

Scheet, P., & Stephens, M. (2008). Linkage disequilibrium-based quality control for large-scale genetic studies. *PLoS Genetics*, 4(8), e1000147.

Schrider, D. R., & Hahn, M. W. (2010). Gene copy-number polymorphism in nature. In *Proceedings of the Royal Society B: Biological Sciences*, 277(1698), 3213-3221.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y. H., Hicks, J., Spence, S.

J., Lee, A. T., Puura, K., Lehtimäki, T., … Wigler, M. (2007). Strong association of de novo copy number mutations with autism. *Science*, 316(5823), 445-449.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Anders Zetterberg, A., Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305(5683), 525-528.

Sekar, A., Bialas, A. R., De Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S. A., Handsaker, R. E., Daly, M. J., Carroll, M. C., Stevens, B., & McCarroll, S. A. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature*, 530(7589), 177-183.

Sharp, A. J., Cheng, Z., & Eichler, E. E. (2006). Structural Variation of the Human Genome. *Annual Review of Genomics and Human Genetics*, 7(1), 407–442.

She, X., Cheng, Z., Zöllner, S., Church, D. M., & Eichler, E. E. (2008). Mouse segmental duplication and copy number variation. *Nature Genetics*, 40(7), 909–914.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10), 1135-1145.

Siriboonrit, U., Tsuchiya, N., Sirikong, M., Kyogoku, C., Bejrachandra, S., Suthipinittharm, P., Luangtrakool, K., Srinak, D., Thongpradit, R., Fujiwara, K., Chandanayingyong, D., & Tokunaga, K. (2003). Association of Fcγ receptor IIb and IIIb polymorphisms with susceptibility to systemic lupus erythematosus in Thais. *Tissue Antigens*, 61(5), 374-383.

Slatkin, M. (2008). Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. In *Nature Reviews Genetics*, 9(6), 477-85.

Smith, K. G. C., & Clatworthy, M. R. (2010). FcγRIIB in autoimmunity and infection: Evolutionary and therapeutic implications. In *Nature Reviews Immunology*, 10(5), 328-343.

Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y., & Hay, S. I. (2005). The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature*, 434(7030), 214-217.

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C. T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A. L., Barbazuk, W. B., Jeddeloh, J. A., Nettleton, D., & Schnable, P. S. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics*, 5(11), e1000734.

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements, and genomic disorders. In *Trends in Genetics*, 18(2), 74–82.

Stankiewicz, P., Park, S. S., Inoue, K., & Lupski, J. R. (2001). The evolutionary chromosome translocation 4;19 in Gorilla gorilla is associated with microduplication of the chromosome fragment syntenic to sequences surrounding the human proximal CMT1A-REP. In *Genome Research*, 11(7),1205–1210.

Stefanescu, R. N., Olferiev, M., Liu, Y. I., & Pricop, L. (2004). Inhibitory Fc gamma receptors: From gene to disease. *Journal of Clinical Immunology*, 24(4), 315-326.

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., & Cooper, D. N. (2012). The human gene mutation database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Current Protocols in Bioinformatics*, 39: 1.13.1-1.13.20.

Stoletzki, N., & Eyre-Walker, A. (2011). Estimation of the neutrality index. *Molecular Biology and Evolution*, 28(1), 63-70.

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., Eichler, E. E., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., … Peterson, J. L. (2010). Diversity of human copy number variation and multicopy genes. *Science*, 330(6004), 641–646.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585–595.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, 30(12), 2725-2729.

Tao, H., Cox, D. R., & Frazer, K. A. (2006). Allele-specific KRT1 expression is a complex trait. *PLoS Genetics*, 2(6), 0848-0858.

Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. In *Frontiers in Bioengineering and Biotechnology*, 3:92.

Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S., & Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. In *Bioinformatics*, 28(21), 2711–2718.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J., & Schork, N. J. (2011). The importance of phase information for human genomics. In *Nature Reviews Genetics*, 12(3), 215–223.

Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonné-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P., Krings, M., Pääbo, S., Watson, E., Risch, N., Jenkins, T., & Kidd, K. K. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, 271(5254), 1380-1387.

Tyler-Smith, C., & Xue, Y. (2012). Sibling rivalry among paralogs promotes evolution of the human brain. In *Cell*, 149(4), 737-739.

Urban, T.J., Weintrob, A.C., Fellay, J., Colombo, S., Shianna, K.V., Gumbs, C., Rotger, M., Pelak, K., Dang, K.K., Detels, R., Martinson, J.J., O'Brien, S.J., Letvin, N.L., McMichael, A.J., Haynes, B.F., Carrington, M., Telenti, A., Michael, N.L., Goldstein, D.B. (2009). CCL3L1 and HIV/AIDS susceptibility. *Nature Medicine*, 15(10), 1110–1112.

Valsesia, A., Macé, A., Jacquemont, S., Beckmann, J. S., & Kutalik, Z. (2013). The growing importance of CNVs: New insights for detection and clinical interpretation. In *Frontiers in Genetics*, 4(92).

van der Heijden, J., Breunis, W. B., Geissler, J., de Boer, M., van den Berg, T. K., & Kuijpers, T. W. (2012). Phenotypic Variation in IgG Receptors by Nonclassical FCGR2C Alleles. *The Journal of Immunology*, 188(3), 1318-1324.

Veal, C. D., Xu, H., Reekie, K., Free, R., Hardwick, R. J., McVey, D., Brookes, A. J., Hollox, E. J., & Talbot, C. J. (2013). Automated design of paralogue ratio test assays for the accurate and rapid typing of copy number variation. *Bioinformatics*, 29(16), 1997-2003.

Vogelpoel, L. T. C., Baeten, D. L. P., de Jong, E. C., & den Dunnen, J. (2015). Control of cytokine production by human Fc gamma receptors: Implications for pathogen defense and autoimmunity. In *Frontiers in Immunology*, 6(79).

Wagner, F. F., & Flegel, W. A. (2000). RHD gene deletion occurred in the Rhesus box. *Blood*, 95(12):3662-3668.

Wain, L. V, Armour, J. a L., Tobin, M. D. (2009). Genomic copy number variation, human health, and disease. *Lancet*, 374(9686), 340–350.

Wellcome Trust Case Control Consortium, Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., Giannoulatou, E., Holmes, C., Marchini, J.L., Stirrups, K., Tobin, M.D., Wain, L.V., Yau, C., Aerts, J., Ahmad, T., Andrews, T.D., Arbury, H., Attwood, A., Auton, A., Ball, S.G., Balmforth, A.J., Barrett, J.C., Barroso, I., Barton, A., Bennett, A.J., Bhaskar, S., Blaszczyk, K., G., Eyre, S., Farmer, A., Ferrier, I.N. and Donnelly, P. (2010) 'Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls', Nature, 464(7289), 713-720.

Willcocks, L. C., Carr, E. J., Niederer, H. A., Rayner, T. F., Williams, T. N., Yang, W., Anthony G Scott, J., Urban, B. C., Peshu, N., Vyse, T. J., Lau, Y. L., Lyons, P. A., & Smith, K. G. C. (2010). A defunctioning polymorphism in FCGR2B is associated with protection against malaria but susceptibility to systemic lupus erythematosus. *Proceedings of the National Academy of Sciences of the United States of America*, 107(17), 7881-7885.

Yang, T. L., Chen, X. D., Guo, Y., Lei, S. F., Wang, J. T., Zhou, Q., Pan, F., Chen, Y., Zhang, Z. X., Dong, S. S., Xu, X. H., Yan, H., Liu, X., Qiu, C., Zhu, X. Z., Chen, T., Li, M., Zhang, H., Zhang, L., … Deng, H. W. (2008). Genome-wide Copy-Number-Variation Study Identified a Susceptibility Gene, UGT2B17, for Osteoporosis. *American Journal of Human Genetics*, 83(6), 663–674.

Yang, Y., Chung, E. K., Wu, Y. L., Savelli, S. L., Nagaraja, H. N., Zhou, B., Hebert, M., Jones, K. N., Shu, Y., Kitzmiller, K., Blanchong, C. A., McBride, K. L., Higgins, G. C., Rennebohm, R. M., Rice, R. R., Hackshaw, K. V., Roubey, R. A., Grossman, J. M., Tsao, B. P., Birmingham, D. J., Rovin, B. H., Hebert, L. A., Yu, C. Y. (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): Low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *American Journal of Human Genetics*, 80(6):1037-1054.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., & Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, 19(9), 1586–1592.

Yohe, S., & Thyagarajan, B. (2017). Review of clinical next-generation sequencing. In *Archives of Pathology and Laboratory Medicine*, 141(11), 1544-1557.

Yu, Z., Garner, C., Ziogas, A., Anton-Culver, H., & Schaid, D. J. (2009). Genotype determination for polymorphisms in linkage disequilibrium. *BMC Bioinformatics*, 10,63.

Zarrei, M., MacDonald, J.R., Merico, D., Scherer, S.W. (2015) A copy number variation map of the human genome. *Nature Reviews Genetics*, 16(3),172-83.

Zhang, H., Li, J., Zhang, X., Wang, Y., Qiu, W., Ye, J., Han, L., Gao, X., & Gu, X. (2011a). Analysis of the IDS gene in 38 patients with Hunter syndrome: The c.879G>A (p.Gln293Gln) synonymous variation in a female create exonic splicing. *PLoS ONE*. 6(8), e22951.

Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011b). The impact of next-generation sequencing on genomics. In *Journal of Genetics and Genomics*, 38(3), 95–109.

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, 10, 451-81.

Zhang, L., Lu, H. H. S., Chung, W. Y., Yang, J., & Li, W. H. (2005). Patterns of segmental duplication in the human genome. *Molecular Biology and Evolution*, 22(1), 135-141.

Zhao, M., Wang, Q., Wang, Q., Jia, P., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. *BMC Bioinformatics*, 14(Suppl 11): S1.

Zhou, X. jie, Lv, J. cheng, Bu, D. fang, Yu, L., Yang, Y. rong, Zhao, J., Cui, Z., Yang, R., Zhao, M. hui, & Zhang, H. (2009). Copy number variation of FCGR3A rather than FCGR3B and FCGR2B is associated with susceptibility to anti-GBM disease. *International Immunology*, 22(1), 45-51.

APPENDICES

Appendix 1. 10x Low dNTP-PCR mix used for PRT Assay				
Reagent	**Stock Conc.**	**Volume(ul)**	**Conc. In 10x**	**Conc.in PCR mix**
Tris-HCl(pH8.8)	2M	500	500mM	50mM
Ammonium sulphate	1M	250	125mM	12.5mM
MgCl2	1M	28	14mM	1.4mM
2-mercaptoethanol	100%	10.5	75mM	7.5mM
dATP*	100mM	40	2mM	0.2mM
dCTP	100mM	40	2mM	0.2mM
dCTP	100mM	40	2mM	0.2mM
dCTP	100mM	40	2mM	0.2mM
BSA**	10mg/ml	250	1.25mg/ml	125µg/ml
DNAase free H2O		800		
Total		2000		
*Promega is used for dNTPs **NOT acetylated				

Appendix 2. 11.1x PCR mix for long-range PCR			
Reagent	**Stock Conc.**	**Volume(ul)**	**Final Conc. In Reaction**
Tris-HCl(pH8.8)	2M	167	45mM
Ammonium sulphate	1M	83	11mM
MgCl2	1M	33.5	4.5mM
2-mercaptoethanol	100%	3.6	6.7mM
EDTA(pH8.0)	10mM	3.4	4.4µM
dATP*	100mM	75	1mM
dCTP	100mM	75	1mM
dCTP	100mM	75	1mM
dCTP	100mM	75	1mM
BSA**	10mg/ml	17	113µg/ml
DNAase free H2O		68	
Total		675.5	
*Promega is used for dNTPs **NOT acetylated			

Appendix 3. The percentage of short reads remained after trimming by quality.					
	Sampe ID	**mrsFast-Ultra, min length is 250**		**BWA, min length is 150**	
		Set 1	**Set 2**	**Set 1**	**Set 2**
1	HG00096	29	35	57	56
2	HG00268	40	43	59	64
3	HG00419	37	31	54	43
4	HG00759	32	31	41	45
5	HG01051	19	17	26	28
6	HG01112	20	25	34	35
7	HG01500	42	46	56	53
8	HG01565	28	12	30	28
9	HG01583	3	10	18	20
10	HG01595	17	17	26	30
11	HG01879	41	30	49	48
12	HG02568	10	15	35	19
13	HG02922	18	14	18	16
14	HG03006	20	25	45	42
15	HG03052	30	12	46	46
16	HG03642	7	15	17	24
17	HG03742	30	33	21	18
18	NA18525	44	43	64	65
19	NA18939	42	44	57	57
20	NA19017	43	33	52	53
21	NA19625	9	16	33	35
22	NA19648	28	23	47	45
23	NA20502	36	14	55	53
24	NA20845	20	19	16	17

Appendix 4. The average coverage calculated for the samples mapped by mrsFast-Ultra							
		Filtered			Non-filtered		
	Sample ID	**94%**	**92%**	**90%**	**94%**	**92%**	**90%**
1	HG00096	7.3	7.9	8.6	21.6	24.1	26.6
2	HG00268	10.0	10.9	11.6	23.4	25.9	28.3
3	HG00419	7.5	8.2	8.9	19.3	21.5	23.7
4	HG00759	8.0	8.7	9.2	20.4	22.6	24.7
5	HG01051	4.7	5.1	5.5	18.5	21.0	23.2
6	HG01112	4.6	5.0	5.3	14.8	16.5	18.0
7	HG01500	11.7	12.7	13.6	22.8	25.3	27.6
8	HG01565	4.9	5.4	5.8	16.5	18.7	20.7
9	HG01583	0.9	1.1	1.2	11.8	13.5	15.2
10	HG01595	3.8	4.1	4.5	16.5	18.7	20.7
11	HG01879	10.8	11.6	12.4	26.5	29.3	31.8
12	HG02568	1.5	1.7	1.8	9.6	11.0	12.4
13	HG02922	2.9	3.4	3.8	15.8	18.2	20.5
14	HG03006	9.8	10.6	11.3	23.4	26.0	28.3
15	HG03052	4.5	5.0	5.4	19.1	21.8	24.5
16	HG03642	1.8	2.0	2.2	13.5	15.7	17.7
17	HG03742	6.9	7.5	8.1	18.0	20.1	22.1
18	NA18525	10.0	10.9	11.8	25.8	29.0	31.9
19	NA18939	10.2	11.1	11.9	15.9	18.5	21.0
20	NA19625	2.3	2.5	2.7	21.9	24.3	26.8
21	NA19648	5.1	5.6	6.0	21.6	23.9	26.1
22	NA19017	10.7	11.7	12.6	17.8	20.0	22.2
23	NA20502	4.9	5.3	5.7	18.7	20.9	23.1
24	NA20845	3.5	4.0	4.4	10.0	11.8	13.7

Appendix 5. The average coverage calculated for the samples mapped by BWA-MEM for RR*

		Filtered				Non-filtered		
	Sample ID	**94%**	**92%**	**90%**	**92%** for ARR**	**94%**	**92%**	**90%**
1	HG00096	28.0	28.8	28.79	28.2	81.21	84.24	84.24
2	HG00268	30.5	31.5	31.55	30.9	83.52	87.52	87.52
3	HG00419	23.7	24.7	24.72	20.0	79.68	84.21	84.21
4	HG00759	21.3	21.9	21.93	21.6	78.21	67.53	67.53
5	HG01051	15.3	15.7	15.69	15.4	84.69	87.39	98.85
6	HG01112	15.8	16.4	16.42	15.9	69.08	72.13	72.13
7	HG01500	27.8	28.8	28.84	28.2	82.99	87.10	87.10
8	HG01565	15.7	16.4	16.41	15.9	84.13	87.64	87.64
9	HG01583	7.7	7.9	7.88	7.8	65.37	67.14	67.14
10	HG01595	14.1	14.7	14.67	14.4	63.75	68.00	68.00
11	HG01879	28.1	29.1	29.06	28.5	94.92	98.87	98.87
12	HG02568	8.3	8.7	8.73	8.4	45.74	47.50	47.50
13	HG02922	8.5	8.7	8.70	8.5	82.93	85.79	85.43
14	HG03006	23.9	24.7	24.75	24.2	88.56	92.57	92.57
15	HG03052	23.1	23.7	23.75	23.3	82.93	85.79	85.79
16	HG03642	8.9	9.2	9.21	8.9	65.72	67.84	67.84
17	HG03742	22.5	23.4	23.39	22.9	76.96	80.73	80.73
18	NA18525	29.1	29.8	29.78	29.2	92.59	95.43	95.43
19	NA18939	15.8	16.2	16.15	15.9	75.34	77.61	77.61
20	NA19017	19.2	19.6	19.64	19.4	92.59	71.35	71.35
21	NA19625	31.1	32.2	32.20	31.7	79.50	83.05	83.05
22	NA19648	28.3	29.3	29.32	28.7	88.24	91.88	91.88
23	NA20502	23.1	23.7	23.72	23.3	69.93	72.36	72.36
24	NA20845	7.9	8.2	8.17	7.9	73.06	75.33	75.33

***RR: Reduced reference**

****ARR: Alternative reduced reference**

Appendix 6. The average coverage calculated for the pedigree samples mapped by BWA-MEM for RR and ARR			
		Coverage for RR	**Coverage for ARR**
1	NA12877	72.7	72.3
2	NA12878	69.5	69.0
3	NA12879	65.6	66.2
4	NA12880	62.8	62.0
5	NA12881	51.7	51.7
6	NA12882	95.3	94.2
7	NA12883	61.8	61.8
8	NA12884	57.4	57.2
9	NA12885	58.6	58.6
10	NA12886	70.3	70.5
11	NA12887	56.8	56.6
12	NA12888	64.6	64.6
13	NA12889	75.3	75.3
14	NA12890	55.1	55.0
15	NA12891	53.4	53.4
16	NA12892	67.4	67.4
17	NA12893	66.6	66.4

Appendix 7. The mapped ratios of BWA-MEM mapping with different mismatch ratios between sequence reads and the RR.				
		94%	**92%**	**90%**
1	**HG00096**	0.819	0.812	0.812
2	**HG00268**	0.829	0.827	0.827
3	**HG00419**	0.977	0.960	0.960
4	**HG00759**	0.864	0.859	0.859
5	**HG01051**	0.831	0.818	0.818
6	**HG01112**	0.832	0.822	0.822
7	**HG01500**	0.847	0.841	0.841
8	**HG01565**	0.994	0.979	0.979
9	**HG01583**	1.062	1.050	1.050
10	**HG01595**	0.678	0.682	0.682
11	**HG01879**	0.822	0.817	0.817
12	**HG02568**	0.687	0.686	0.686
13	**HG02922**	0.672	0.666	0.666
14	**HG03006**	0.825	0.818	0.818
15	**HG03052**	0.851	0.838	0.838
16	**HG03642**	0.848	0.833	0.833
17	**HG03742**	0.841	0.833	0.833
18	**NA18525**	0.815	0.813	0.813
19	**NA18939**	0.848	0.843	0.843
20	**NA19017**	0.818	0.810	0.810
21	**NA19625**	0.845	0.839	0.839
22	**NA19648**	0.860	0.851	0.851
23	**NA20502**	0.832	0.826	0.826
24	**NA20845**	0.840	0.841	0.841

Appendix 8. Bioinformatics Analysis: Example Scripts used this study.	
Trimming the short reads by quality	
Trimmomatic	`java -jar trimmomatic-0.35.jar PE -threads 0 -trimlog logInput12 input_1.fastq.gz input_2.fastq.gz input_1_paired.fastq input_1_unpaired.fastq input_2_paired.fastq input_2_unpaired.fastq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:250 (MINLEN:150 for BWA)`
Indexing the reduced reference	
mrsFast-Ultra	`mrsfast --index reference.fa --ws 14`
SAMtools	`Samtools faidx reference.fa`
BWA	`bwa index -a is reference.fa`
Picard	`java -jar /cm/shared/apps/picard/2.6.0/picard.jar CreateSequenceDictionary.jar R= reference.fa O=reference.dict`
Mapping: mrsfast-ultra (Error threshold 8: 92 % similarity between reference and the short reads.)	
mrsFast-Ultra	`mrsfast --search reference.fa --pe --seq1 paired_end1.fastq --seq2 paired_end2.fastq -e 8 threads 0 --best --disable-nohits -o ALN.sam`
BWA-MEM	`bwa mem -M reference.fa -A 1 -B 3.7174 paired_end1.fastq paired_end2.fastq >ALN .sam`
Converting *.SAM file to *.BAM file	
SAMtools	`samtools view -S -b ALN.sam -o ALN.bam`
Sorting *.BAM file	
SAMtools	`samtools sort ALN.bam -o ALN_sorted.bam`
Indexing *.BAM file	
SAMtools	`samtools index ALN_sorted.bam`
Extracting the only mapped reads from the* .BAM file	
SAMtools	`samtools view -h ALN_sorted.bam -o ALN_extracted.bam reference`
Adding groups to the* .BAM file	
Picard	`java -jar /cm/shared/apps/picard/2.6.0/picard.jar AddOrReplaceReadGroups I= ALN_extracted.bam O=ALN_RG.bam RGID= ALN RGLB=TBNF RGPL=illumina RGPU=unit1 rgsm= ALN VALIDATION_STRINGENCY=LENIENT`
Converting mapping quality score for only mrsfast-ultra	

GATK	`gatk -T PrintReads -R reference.fa -I ALN.bam -o ALN_RG_sorted.bam -rfReassignOneMappingQuality -RMQF 255 -RMQT 60`
Local Realignment	
GATK	`gatk -T PrintReads -R reference.fa -I ALN.bam -o ALN_RG_sorted.bam –rf ReassignOneMappingQuality -RMQF 255 -RMQT 60`
Marking PCR duplicates	
Picard	`java -jar /cm/shared/apps/picard/2.6.0/picard.jar MarkDuplicates I=ALN_LR_sorted.bam O=ALN_MD.bam METRICS_FILE=ALN_dupl_metrics.txt VALIDATION_STRINGENCY=LENIENT`
Obtaining Flagstats	
SAMtools	`samtools flagstat ALN_RG_sorted.bam> ALN_before.txt (Before local realignment)`
SAMtools	`samtools flagstat ALN_MD_sorted.bam> ALN_after.txt (After local realignment and marking duplicates)`
Removing PCR duplicates	
SAMtools	`samtools rmdup ALN_LR_sorted.bam ALN_LR_DR.bam`
Counting the number of mapped reads from the *.BAM file	
Perl	`perl reads_per_region.pl refbed.bed ALN_RG_sorted.bam ALN_OUT.txt (Before local realignment)`
Perl	`perl reads_per_region.pl refbed.bed ALN_LR_DR_sorted.bam ALN_BR_OUT.txt (After local realignment and marking duplicates)`
Obtaining final Flagstats for calculating the Coverage	
samtools	`Samtools flagstat ALN_LR_DR_sorted.bam> ALN_LR_DR_final.txt`
Calling variants	
FreeBayes (on polyploid data)	`freebayes -f reference.fa -p 4 --min-base-quality 30 --min-mapping-quality 50 --min-alternate-count 10 --min-alternate-fraction 0.10 ALN_ LR_DR_sorted.bam> ALN.vcf`
FreeBayes (on diploid data)	`freebayes -f FCG2BLong.fasta --min-base-quality 30 --min-mapping-quality 50 --min-alternate-count 10 --min-alternate-fraction 0.10 X_ALN.bam> X_ALN.vcf`

Appendix 9. The perl script to count the number of mapped reads for a given region.

```
#!/usr/bin/perl
use strict;
use warnings;

# usage is
# perl reads_per_region.pl XXXXXX.bed XXXXXXXXX.bam
XXXXXOUT.txt
# where bed file indicates the regions, bam file is the input
alignment and txt file is the output

my ($bed_file, $bam_file, $out_file) = @ARGV;

open (BED, "<$bed_file") or die "cant open $bed_file - $!";
open (OUT, ">$out_file") or die "cant write to $out_file -
$!";
while( <BED> ){
      my $row = $_;
      chomp ($row);
      my ($csome,$start,$stop) = split (/\s+/,$row);
      my $result = `samtools view -c -F 4 $bam_file
"$csome:$start-$stop"`;
      print OUT "$row $result";
}
close BED;
close OUT;
```

Appendix 10. CNVtools script for analysis of CNs and histograms of the samples used in this study.

```
>library(CNVtools)
# Load the data from the file into a dataframe called
PLeftFCGR<-read.table("Left_FCGR.txt",header=T)
#Create a matrix with just the raw data
#Exclude the 1st and 2nd column and assing to raw.signal as
matrix
raw.signal <- as.matrix(PLeftFCGR[, -c(1, 2)])
#Get PLeftFCGR$Sample column as names
dimnames(raw.signal)[[1]] <- PLeftFCGR$L_Ratios
# Draw histogram
hist(PLeftFCGR[,3], breaks = 50, main = "Mean signal",
cex.lab = 1.3)

ncomp <- 2
batches <- factor(PLeftFCGR[,2])
sample <- factor(PLeftFCGR[,1])
# plots fitted data
fit.mean<-CNVtest.binary(signal = PLeftFCGR[,3], sample =
sample, batch = batches, ncomp = ncomp, n.H0 = 10, n.H1 = 0,
model.var = "~ 1")

In CNVtest.binary(signal = PLeftFCGR[, 3], sample = sample,
print(fit.mean$status.H0)
# expected [1] "C"

cnv.plot(fit.mean$posterior.H0, batch = "FCG_cn", main =
"FCG_cn",breaks = 50, col = "red")
pdf("fitted_Left_FCGR.pdf",width=10, height=5)
cnv.plot(fit.mean$posterior.H0, batch = "FCG_cn", main =
"FCG_cn",breaks = 50, col = "red")
dev.off()
# write data to file
write.csv(fit.mean$posterior.H0,"fitted_Left_FCGR.csv")
OR
write.table(fit.mean$posterior.HO,"fitted_Left_FCGR.txt",
sep="\t")
```

Appendix 11. An example of calculation to generate ratios from the final bam finals.				
Regions for RR	**Reference coordinates**	**Corresponding coordinates for RR**	**Total number of reads**	**Calculations**
R1	chr1:25,584,516-25,594,515	1-10,000	797	=R_C1/ (R1+R2+R3)
R_C1	**chr1:25,594,516-25,655,519**	**10,001-71,004**	**12034**	
R2	chr1:25,655,521-25,688,914	71,005-104,398	9812	=12034/ (797 +9812 +1598)
R3	chr1:25,751,819-25,761,819	104,399-114,400	1598	=0.986
F1	chr1:161,050,791-161,237,982	114,401-301,592	23225	=F_C1/ (F1+F2+F3)
F2	chr1:161,469,969-161,479,969	301,593-311,593	1799	
F_C1	**chr1:161,479,970-161,565,133**	**311,594-396,756**	**38269**	=38269/ (23225+1799+5510)
F3	chr1:161,647,428-161,657,427	396,758-406,757	5510	=1.253
B1	chr8:7,102,590-7,112,589	406,758-416,757	9233	=B_C1/ (B1+B2+B3)
B_C1	**chr8:7,112,590-7,442,590**	**416,758-746,758**	**167691**	
B2	chr8:7,442,591-7,452,590	746,759-756,758	9939	=167691/ (9233+9939+18205)
B3	chr8:21,173,211-21,893,410	756,759-1,476,959	18205	=1.317

Appendix 12. Novel variants found in *FCGR2B* locus described by VariantValidator. These variants were checked on IGV and confirmed that they all exist.				
	GRCh37 (CHR)	GRCh37 position	HGVS_Genomic_Description_GRCh37	Code
1	1	161635076	NC_000001.10:g.161635081del	f13
2		161635614	NC_000001.10:g.161635614G>A	f19
3		161637165	NC_000001.10:g.161637165G>A	f25
4		161638748	NC_000001.10:g.161638748C>T	f44
5		161640787	NC_000001.10:g.161640787T>C	f67
6		161640983	NC_000001.10:g.161640983G>T	f70
7		161640994	NC_000001.10:g.161640994C>T	f71
8		161641003	NC_000001.10:g.161641003C>G	f72
9		161641006	NC_000001.10:g.161641006G>A	f73
10		161641471	NC_000001.10:g.161641471C>T	f76
11		161643327	NC_000001.10:g.161643327C>G	f87
12		161643371	NC_000001.10:g.161643371C>T	f88
13		161644663	NC_000001.10:g.161644663T>C	f109
14		161644713	NC_000001.10:g.161644713G>T	f109h
15		161644998	NC_000001.10:g.161644998G>A	f114
16		161645052	NC_000001.10:g.161645052C>T	f116
17		161646784	NC_000001.10:g.161646784C>A	f138

Appendix 13. The variants found in *FCGR2B* locus described by VariantValidator.			
Code	GRCh37 position	HGVS_Genomic_Description_GRCh37	SNPs
f3	161632912	NC_000001.10:g.161632912T>A	rs780467580 CR046102
f4	161633176	NC_000001.10:g.161633176C>G	rs747505037
f5	161633277	NC_000001.10:g.161633277C>A	rs1459475
f6	161633527	NC_000001.10:g.161633527T>C	rs3767642
f7	161633774	NC_000001.10:g.161633774G>T	rs1832738
f8	161634079	NC_000001.10:g.161634079G>A	rs1347898420
f9	161634189	NC_000001.10:g.161634189A>G	rs1824385
f10	161634758	NC_000001.10:g.161634758G>A	rs4656325
f11	161634945	NC_000001.10:g.161634945T>C	Bonatti et al., 2017917643
f12	161635076	NC_000001.10:g.161635081del	rs529326020
f13	161635076	NC_000001.10:g.161635081del	-
f14	161635125	NC_000001.10:g.161635125T>C	rs6687928
f15	161635196	NC_000001.10:g.161635196A>T	rs796412550
f16	161635221	NC_000001.10:g.161635221C>T	rs1362581961
f17	161635303	NC_000001.10:g.161635303T>C	rs10436883
f18	161635468	NC_000001.10:g.161635468G>A	rs56384835
f19	161635614	NC_000001.10:g.161635614G>A	-

f20	161635689	NC_000001.10:g.161635689G>A	rs12567467
f21	161635694	NC_000001.10:g.161635694A>T	rs12568594
f22	161636708	NC_000001.10:g.161636708T>G	rs10465555
f23	161636917	NC_000001.10:g.161636917C>A	rs941359489
f24	161636974	NC_000001.10:g.161636974C>T	rs55989922
f25	161637165	NC_000001.10:g.161637165G>A	-
f26	161637175	NC_000001.10:g.161637175G>A	rs6658657
f27	161637192	NC_000001.10:g.161637192A>T	rs6661332
f28	161637258	NC_000001.10:g.161637258G>C	rs1302043377
f29	161637265	NC_000001.10:g.161637265C>A	rs6694919
f30	161637273	NC_000001.10:g.161637273G>A	rs1452712765
f31	161637290	NC_000001.10:g.161637290C>T	rs10753599
f32	161637408	NC_000001.10:g.161637408G>A	rs145624007
f33	161638043	NC_000001.10:g.161638043C>T	rs6664704
f34	161638268	NC_000001.10:g.161638268A>T	rs4233376
f35	161638368	NC_000001.10:g.161638368C>T	rs202094230
f36	161638386	NC_000001.10:g.161638386G>A	rs12141072
f37	161638410	NC_000001.10:g.161638410G>T	rs6673373
f38	161638411	NC_000001.10:g.161638411T>A	rs4233377
f39	161638420	NC_000001.10:g.161638420A>G	rs1317922907
f40	161638530	NC_000001.10:g.161638530G>A	rs2007773
f41	161638590	NC_000001.10:g.161638590A>G	rs12116547
f42	161638608	NC_000001.10:g.161638608T>C	rs2007763
f43	161638709	NC_000001.10:g.161638709G>T	rs55858328
f44	161638748	NC_000001.10:g.161638748C>T	-
f45a	161638754	NC_000001.10:g.161638772_161638773del	rs1392591383
f45b	161638754	NC_000001.10:g.161638772_161638773dup	rs1411420147
f46	161638842	NC_000001.10:g.161638842T>C	rs57628745
f47	161638951	NC_000001.10:g.161638951T>C	rs2169051
f48	161639229	NC_000001.10:g.161639229C>T	rs60081850
f49	161639285	NC_000001.10:g.161639285G>A	rs4657086
f50	161639316	NC_000001.10:g.161639316C>T	rs12145090
f51	161639334	NC_000001.10:g.161639337del	rs1383318848
f52	161639364	NC_000001.10:g.161639410_161639414del	rs200504085
f53	161639450	NC_000001.10:g.161639450C>A	rs7539744
f54	161639464	NC_000001.10:g.161639464G>A	rs2333845
f55	161639559	NC_000001.10:g.161639559G>A	rs1984769
f56	161639600	NC_000001.10:g.161639600C>T	rs12145843
f57	161639812	NC_000001.10:g.161639812C>A	rs375037288
f58	161639926	NC_000001.10:g.161639926C>T	rs12145988
f59	161639951	NC_000001.10:g.161639951C>T	rs1282533402

f60	161640027	NC_000001.10:g.161640027C>T	rs1478203045
f61	161640095	NC_000001.10:g.161640095C>T	rs922087
f62	161640375	NC_000001.10:g.161640375A>C	rs1226951241
f63	161640404	NC_000001.10:g.161640404T>C	rs1206566141
f64	161640522	NC_000001.10:g.161640522G>T	rs7519636
f65	161640600	NC_000001.10:g.161640600T>G	rs1344438491
f66	161640723	NC_000001.10:g.161640723C>T	rs557359975
f67	161640787	NC_000001.10:g.161640787T>C	-
f68	161640895	NC_000001.10:g.161640895A>G	rs60388169
f69	161640977	NC_000001.10:g.161640977C>T	rs3754055
f70	161640983	NC_000001.10:g.161640983G>T	-
f71	161640994	NC_000001.10:g.161640994C>T	-
f72		NC_000001.10:g.161641003C>G	-
f73	161641006	NC_000001.10:g.161641006G>A	-
f74	161641148	NC_000001.10:g.161641148C>G	rs141749158
f75	161641384	NC_000001.10:g.161641384G>A	rs6665610
f76		NC_000001.10:g.161641471C>T	-
f77	161641893	NC_000001.10:g.161641893G>A	rs11811662
f78	161642207	NC_000001.10:g.161642207C>T	rs4999919
f79	161642661	NC_000001.10:g.161642661T>G	rs190837656
f80	161642982	NC_000001.10:g.161642982G>A	rs2298022
f81	161642985	NC_000001.10:g.161642985G>A	rs182968886
f82	161643025	NC_000001.10:g.161643025C>A	rs2125684
	161643025	NC_000001.10:g.161643025C>A	-
f83	161643044	NC_000001.10:g.161643044A>G	rs2125685
f84	161643067	NC_000001.10:g.161643067C>T	rs199544489
f85	161643072	NC_000001.10:g.161643072G>T	rs142579257
f86	161643269	NC_000001.10:g.161643269C>T	rs2045571
f87	161643327	NC_000001.10:g.161643327C>G	-
f88	161643371	NC_000001.10:g.161643371C>T	-
f89	161643507	NC_000001.10:g.161643507C>T	rs2793082
f90	161643512	NC_000001.10:g.161643512T>C	rs2793081
f91	161643527	NC_000001.10:g.161643527A>C	rs1674755
f92	161643529	NC_000001.10:g.161643529A>G	rs369224649
f93	161643533	NC_000001.10:g.161643533A>G	rs1674756
f94a	161643546	NC_000001.10:g.161643546_161643547inv	rs1771554
f94b	161643546	NC_000001.10:g.161643546T>C	rs1771554
f95	161643560	NC_000001.10:g.161643560C>T	rs74816838
f96	161643565	NC_000001.10:g.161643565A>G	rs372700655
f97	161643643	NC_000001.10:g.161643643A>G	rs2045572
f98	161643663	NC_000001.10:g.161643663C>T	rs7552498

f99	161643798	NC_000001.10:g.161643798T>C	rs1050501, CM033377
f100	161643889	NC_000001.10:g.161643889G>T	rs6666965
f101	161643984	NC_000001.10:g.161643984G>T	rs12117530
f102	161644002	NC_000001.10:g.161644002C>T	rs570127461
f103	161644003	NC_000001.10:g.161644003G>A	rs537980512
f104	161644258	NC_000001.10:g.161644258G>C	rs75409195
f105	161644307	NC_000001.10:g.161644307_161644308delinsTG	rs1771556
	161644308	NC_000001.10:g.161644308A>G	-
f106	161644387	NC_000001.10:g.161644387A>T	rs1674757
f107	161644408	NC_000001.10:g.161644408T>C	rs1674775
f108	161644422	NC_000001.10:g.161644422G>C	rs1771557
f109	161644663	NC_000001.10:g.161644663T>C	-
f109h	161644713	NC_000001.10:g.161644713G>T	-
f110	161644811	NC_000001.10:g.161644811C>T	rs17413015
f111	161644871	NC_000001.10:g.161644871A>C	rs72480273
f112	161644955	NC_000001.10:g.161644955G>A	rs7532925
f113	161644976	NC_000001.10:g.161644976C>A	rs1674758
f114	161644998	NC_000001.10:g.161644998G>A	-
f115	161645010	NC_000001.10:g.161645010C>A	rs6427615
f116	161645052	NC_000001.10:g.161645052C>T	-
f117	161645058	NC_000001.10:g.161645058T>G	rs148534844
f118	161645170	NC_000001.10:g.161645170G>C	rs12046494
f119	161645187	NC_000001.10:g.161645187A>G	rs12058663
f120	161645232	NC_000001.10:g.161645232T>C	rs6670713
f121	161645259	NC_000001.10:g.161645259T>G	rs1674759
f122	161645329	NC_000001.10:g.161645329A>G	rs180684110
f123	161645368	NC_000001.10:g.161645368G>T	rs571815093
f124	161645409	NC_000001.10:g.161645409C>T	rs1674760
f125	161645471	NC_000001.10:g.161645471A>C	rs1674761
f126	161645877	NC_000001.10:g.161645877G>A	rs13376485
f127	161645894	NC_000001.10:g.161645894_161645895delinsCT	rs386636125
f128	161646112	NC_000001.10:g.161646112C>T	rs3767641
f129	161646116	NC_000001.10:g.161646116T>C	rs3767640
f130a	161646120	NC_000001.10:g.161646120_161646121inv	rs386636127
f130b	161646120	NC_000001.10:g.161646120T>C	rs3767639
f131	161646180	NC_000001.10:g.161646186del	rs968903547
f132	161646194	NC_000001.10:g.161646194C>T	rs1476353150
f133	161646387	NC_000001.10:g.161646387G>A	rs558921632
f134	161646455	NC_000001.10:g.161646455G>T	rs115835689
f135	161646500	NC_000001.10:g.161646500_161646501insC	rs3835613

f137	161646625	NC_000001.10:g.161646625C>G	rs72704062
f138	161646784	NC_000001.10:g.161646784C>A	-
f139	161646787	NC_000001.10:g.161646793del	rs148086886
f140	161646807	NC_000001.10:g.161646807C>T	rs11799952
f141	161646824	NC_000001.10:g.161646824C>A	rs12118043
f142	161647001	NC_000001.10:g.161647001G>C	rs148573502
f143	161647533	NC_000001.10:g.161647533A>G	rs844
f144	161647559	NC_000001.10:g.161647559C>T	rs60519172
f145	161647638	NC_000001.10:g.161647638A>G	rs552605155
f146	161647810	NC_000001.10:g.161647810A>G	rs114873762
f147	161648028	NC_000001.10:g.161648028C>T	rs190526914

Appendix 14. The samples genotyped for the gene conversion assay with control samples

Homozygous no gene conversion		Heterozygous 9.1kb/2.4kb gene conversion		Homozygous 9.1kb/2.4kb gene conversion	
-/-	Copy number	-/+	Copy number	+/+	Copy number
NA19240	4	**NA18956**	5	**NA18517**	3
NA12878	4	**NA18555**	5	**NA18507**	4
NA12156	4	HG00096	4	HG00268	4
HG01500	4	HG00419	5	HG01051	4
HG01595	3	HG00759	4	NA12880	4
HG02568	3	HG01112	4	NA12893	4
HG03006	4	HG01565	5	NA12891	4
NA20845	4	HG01583	4		
NA12881	5	HG01879	3		
NA12882	5	HG02922	3		
NA12884	5	HG03052	4		
NA12886	5	HG03642	4		
NA12892	4	HG03742	4		
		NA19017	4		
		NA19648	4		
		NA20502	4		
		NA18525	4		
		NA18939	4		
		NA19625	4		
		NA12877	5		
		NA12879	5		
		NA12883	4		
		NA12885	5		
		NA12887	5		
		NA12888	4		
		NA12889	4		
		NA12890	6		

Appendix 15. The allele frequencies of the variants that were studied by 1000 Genomes Project and gnomAD projects for the *FCGR2B* locus																	
Existing variation	AF	AFR_AF	AMR_AF	EAS_AF	EUR_AF	SAS_AF	AA_AF	EA_AF	gnomAD _AF	gnomAD _AFR_AF	gnomAD _AMR_AF	gnomAD _ASJ_AF	gnomAD _EAS_AF	gnomAD _FIN_AF	gnomAD _NFE_AF	gnomAD _OTH_AF	gnomAD _SAS_AF
rs747505037	-	-	-	-	-	-	-	-	0.045	0.012	0.038	0.115	0.000	0.085	0.077	0.047	0.028
rs1459475	-	-	-	-	-	-	-	-	0.302	0.169	0.406	0.250	0.351	0.326	0.269	0.298	0.269
rs1832738	0.122	0.173	0.071	0.160	0.094	0.078	-	-	-	-	-	-	-	-	-	-	-
rs529326020	0.060	0.004	0.029	0.122	0.067	0.086	-	-	-	-	-	-	-	-	-	-	-
rs6661332	0.001	0.002	0.000	0.001	0.000	0.000	-	-	-	-	-	-	-	-	-	-	-
rs6694919	0.980	0.930	0.991	0.999	1.000	1.000	-	-	-	-	-	-	-	-	-	-	-
rs6673373	0.027	0.002	0.045	0.010	0.076	0.015	-	-	-	-	-	-	-	-	-	-	-
rs60081850	0.160	0.263	0.082	0.178	0.118	0.100	-	-	-	-	-	-	-	-	-	-	-
rs4657086	0.065	0.006	0.089	0.001	0.174	0.081	-	-	-	-	-	-	-	-	-	-	-
rs200504085	0.373	0.220	0.504	0.443	0.308	0.480	-	-	-	-	-	-	-	-	-	-	-
rs7539744	0.151	0.248	0.079	0.161	0.115	0.096	-	-	-	-	-	-	-	-	-	-	-
rs2333845	0.386	0.486	0.278	0.263	0.504	0.335	-	-	-	-	-	-	-	-	-	-	-
rs1984769	0.223	0.306	0.160	0.160	0.275	0.170	-	-	-	-	-	-	-	-	-	-	-
rs12145843	0.137	0.164	0.108	0.056	0.216	0.124	-	-	-	-	-	-	-	-	-	-	-
rs375037288	0.027	0.002	0.035	0.001	0.087	0.024	-	-	-	-	-	-	-	-	-	-	-
rs1478203045	-	-	-	-	-	-	-	-	0.000	0.000	0.000	-	0.000	-	0.000	0.000	0.000
rs922087	-	-	-	-	-	-	-	-	0.406	-	0.531	-	1.000	-	0.250	0.250	0.227
rs557359975	0.004	0.014	0.001	0.000	0.000	0.000	-	-	-	-	-	-	-	-	-	-	-
rs60388169	0.015	0.048	0.001	0.006	0.002	0.002	-	-	-	-	-	-	-	-	-	-	-
rs3754055	0.037	0.020	0.010	0.069	0.026	0.055	-	-	-	-	-	-	-	-	-	-	-
rs141749158	0.004	0.001	0.000	0.000	0.016	0.002	0.001	0.009	0.012	0.002	0.004	0.012	0.000	0.043	0.013	0.013	0.003
rs6665610	0.138	0.182	0.107	0.060	0.200	0.120	0.162	0.194	0.145	0.173	0.091	0.092	0.060	0.163	0.181	0.156	0.112
rs11811662	0.123	0.198	0.095	0.014	0.096	0.180	-	-	-	-	-	-	-	-	-	-	-
rs190837656	0.007	0.000	0.009	0.000	0.004	0.026	-	-	-	-	-	-	-	-	-	-	-

rs2298022	0.029	0.043	0.022	0.057	0.007	0.011	-	-	0.003	0.020	0.003	0.002	0.009	0.008	0.001	0.003	0.001
rs182968886	0.110	0.169	0.082	0.006	0.094	0.174	0.164	0.117	0.109	0.160	0.059	0.178	0.007	0.088	0.119	0.118	0.156
rs2125684	0.042	0.092	0.032	0.019	0.026	0.022	-	-	0.041	0.082	0.072	0.022	0.054	0.007	0.031	0.055	0.045
rs2125684	-	0.057	0.009	0.000	0.000	0.001	0.011	0.000	0.004	0.042	0.005	0.005	0.000	0.000	0.000	0.003	0.000
rs2125685	0.402	0.261	0.563	0.475	0.397	0.410	0.061	0.115	0.190	0.167	0.387	0.076	0.340	0.166	0.141	0.200	0.163
rs199544489	-	-	-	-	-	-	-	-	0.028	0.043	0.035	0.019	0.006	0.027	0.028	0.026	0.029
rs142579257	0.040	0.142	0.016	0.000	0.001	0.000	-	-	-	-	-	-	-	-	-	-	-
rs2045571	-	-	-	-	-	-	0.263	0.106	0.128	0.243	0.080	0.120	0.230	0.171	0.107	0.126	0.126
rs1674755	0.187	0.066	0.321	0.179	0.230	0.221	-	-	-	-	-	-	-	-	-	-	-
rs369224649	0.185	0.063	0.320	0.175	0.227	0.220	-	-	-	-	-	-	-	-	-	-	-
rs1674756	0.184	0.063	0.320	0.175	0.226	0.219	-	-	-	-	-	-	-	-	-	-	-
rs1771554	0.184	0.064	0.320	0.175	0.225	0.219	-	-	-	-	-	-	-	-	-	-	-
rs74816838	0.064	0.006	0.089	0.001	0.170	0.079	-	-	-	-	-	-	-	-	-	-	-
rs2045572	0.041	0.145	0.017	0.000	0.002	0.000	-	-	-	-	-	-	-	-	-	-	-
rs7552498	0.223	0.300	0.192	0.059	0.337	0.194	-	-	-	-	-	-	-	-	-	-	-
rs1050501	0.186	0.250	0.094	0.254	0.137	0.145	-	-	0.162	0.255	0.099	0.125	0.241	0.248	0.123	0.148	0.143
rs6666965	0.666	0.551	0.745	0.722	0.705	0.669	-	-	0.702	0.568	0.774	0.685	0.727	0.709	0.693	0.711	0.696
rs12117530	0.109	0.095	0.089	0.060	0.178	0.121	-	-	-	-	-	-	-	-	-	-	-
rs570127461	0.013	0.047	0.001	0.001	0.000	0.000	-	-	-	-	-	-	-	-	-	-	-
rs537980512	0.005	0.017	0.001	0.000	0.000	0.000	-	-	-	-	-	-	-	-	-	-	-
rs75409195	0.057	0.002	0.053	0.204	0.002	0.038	-	-	-	-	-	-	-	-	-	-	-
rs1771556	-	0.980	1.000	1.000	1.000	1.000	-	-	-	-	-	-	-	-	-	-	-
rs1674757	-	0.558	0.741	0.724	0.698	0.660	-	-	-	-	-	-	-	-	-	-	-
rs1674775	0.000	0.001	0.000	0.000	0.000	0.000	-	-	-	-	-	-	-	-	-	-	-
rs17413015	0.129	0.165	0.098	0.059	0.180	0.121	-	-	-	-	-	-	-	-	-	-	-
rs7532925	0.101	0.060	0.061	0.187	0.108	0.091	-	-	-	-	-	-	-	-	-	-	-
rs1674758	0.120	0.007	0.272	0.054	0.208	0.145	-	-	-	-	-	-	-	-	-	-	-

rs6427615	0.179	0.244	0.087	0.251	0.123	0.141	0.259	0.100	0.128	0.240	0.076	0.121	0.227	0.179	0.107	0.126	0.124
rs148534844	-	-	-	-	-	-	0.012	0.000	0.001	0.014	0.001	0.000	0.001	0.000	0.000	0.001	0.001
rs12046494	0.006	0.018	0.000	0.002	0.001	0.003	-	-	-	-	-	-	-	-	-	-	-
rs12058663	0.006	0.018	0.001	0.001	0.001	0.001	-	-	-	-	-	-	-	-	-	-	-
rs6670713	0.005	0.000	0.036	0.001	0.000	0.000	-	-	-	-	-	-	-	-	-	-	-
rs1674759	0.793	0.722	0.839	0.781	0.878	0.779	-	-	-	-	-	-	-	-	-	-	-
rs180684110	0.009	0.000	0.065	0.001	0.001	0.000	-	-	-	-	-	-	-	-	-	-	-
rs571815093	0.000	0.001	0.000	0.000	0.001	0.000	-	-	-	-	-	-	-	-	-	-	-
rs1674760	-	0.722	0.837	0.781	0.878	0.779	-	-	-	-	-	-	-	-	-	-	-
rs13376485	0.187	0.266	0.091	0.258	0.125	0.141	-	-	-	-	-	-	-	-	-	-	-
rs3767641	0.183	0.261	0.089	0.251	0.123	0.137	-	-	-	-	-	-	-	-	-	-	-
rs3767640	0.183	0.260	0.089	0.251	0.123	0.137	-	-	-	-	-	-	-	-	-	-	-
rs3767639	0.183	0.260	0.091	0.251	0.123	0.137	-	-	-	-	-	-	-	-	-	-	-
rs558921632	0.000	0.000	0.000	0.000	0.000	0.001	-	-	-	-	-	-	-	-	-	-	-
rs115835689	0.005	0.018	0.001	0.000	0.000	0.000	-	-	-	-	-	-	-	-	-	-	-
rs3835613	0.276	0.390	0.176	0.243	0.290	0.213	-	-	-	-	-	-	-	-	-	-	-
rs72704062	0.012	0.000	0.019	0.000	0.037	0.008	-	-	-	-	-	-	-	-	-	-	-
rs148086886	0.029	0.003	0.035	0.001	0.074	0.044	-	-	-	-	-	-	-	-	-	-	-
rs11799952	0.132	0.206	0.099	0.016	0.120	0.186	-	-	-	-	-	-	-	-	-	-	-
rs12118043	0.091	0.008	0.097	0.061	0.205	0.114	-	-	-	-	-	-	-	-	-	-	-
rs148573502	0.006	0.023	0.000	0.000	0.001	0.000	-	-	-	-	-	-	-	-	-	-	-
rs844	-	0.705	0.757	0.723	0.697	0.653	-	-	-	-	-	-	-	-	-	-	-
rs60519172	0.014	0.054	0.000	0.000	0.000	0.000	-	-	-	-	-	-	-	-	-	-	-
rs552605155	0.000	0.000	0.000	0.000	0.002	0.000	-	-	-	-	-	-	-	-	-	-	-
rs114873762	0.016	0.057	0.006	0.000	0.000	0.000	-	-	-	-	-	-	-	-	-	-	-
rs190526914	0.001	0.004	0.001	0.000	0.000	0.001	-	-	-	-	-	-	-	-	-	-	-

Appendix 16. The constructed haplotypes of the CEPTH/UTAH 1463 from the Ion semiconductor sequenced samples of *FCGRB2B.*

	NA12889		NA12890		NA12891		NA12892	
rs780467580	T	T	T	T	T	A	T	T
rs747505037	C	C	C	C	C	G	C	C
rs1459475	C	C	C	C	A	C	C	C
rs3767642	T	T	T	C	C	C	C	C
rs1824385	G	G	G	G	A	G	G	G
rs4656325	G	G	G	G	A	G	G	G
rs10917643	T	C	C	C	C	C	C	C
rs529326020	C-	CA	CA	CA	CA	CA	C	CA
rs6687928	T	C	C	T	C	T	T	T
rs796412550	A	A	A	T	A	T	T	T
rs10436883	T	T	T	C	T	C	C	C
rs56384835	G	G	G	A	G	A	A	A
rs12567467	G	G	G	A	G	A	A	A
rs12568594	A	A	A	T	A	T	T	T
rs10465555	T	T	T	G	T	G	G	G
rs55989922	C	C	C	C	C	T	C	C
rs6658657	G	A	A	A	A	A	A	A
rs6661332	A	T	T	T	T	T	T	T
rs6694919	A	C	C	C	C	C	C	C
rs10753599	C	T	T	T	T	T	T	T
rs145624007	A	G	A	G	G	G	G	G
rs6664704	T	T	T	C	T	C	C	C
rs4233377	T	T	T	T	A	T	T	T
rs2007773	G	G	G	A	G	A	G	G
rs2007763	C	C	C	C	T	C	C	C
rs55858328	G	G	G	G	G	T	G	G
rs2169051	C	C	C	T	T	T	T	T
rs60081850	C	C	C	T	C	T	T	T
rs4657086	A	A	A	G	G	G	G	G
rs12145090	T	C	C	C	C	C	C	C
rs1383318848	T-	T-	T-	TG	TG	TG	T	TG
rs7539744	C	C	C	A	C	A	A	A
rs2333845	A	A	A	A	G	A	A	A
rs1984769	G	G	G	A	G	A	A	A
rs922087	C	C	C	C	T	C	C	C
rs1344438491	G	T	T	T	T	T	T	T
rs11811662	A	G	G	G	G	G	G	G
rs4999919	C	T	T	C	T	C	C	C
rs182968886	A	G	G	G	G	G	G	G
rs2125684	C	C	C	C	A	C	C	C
rs2125685	A	G	G	A	G	A	A	A
rs2045571	C	C	C	T	C	T	T	T
rs2793082	C	T	T	C	C	C	C	C
rs2793081	T	C	C	T	T	T	T	T
rs1674755	A	C	C	A	C	A	A	A
rs369224649	A	G	G	A	G	A	A	A
rs1674756	A	G	G	A	G	A	A	A
rs1771554	T	C	C	T	C	T	T	T
rs372700655	A	G	G	A	A	A	A	A
rs1050501	T	T	T	C	T	C	C	C
rs6666965	G	T	T	T	T	T	T	T
rs1771556	TG	TG	TG	TG	TG	TG	T	TG
rs1674757	A	T	T	T	T	T	T	T
161644663	C	C	C	C	C	C	C	C
161644713	G	G	G	G	T	G	G	G
rs7532925	G	G	G	A	G	A	A	A
rs1674758	C	C	C	C	A	C	C	C
rs6427615	C	C	C	A	C	A	A	A
rs1674759	T	G	G	G	G	G	G	G
rs1674760	C	T	T	T	T	T	T	T
rs1674761	A	C	C	C	C	C	C	C
rs13376485	G	G	G	A	G	A	A	A
rs386636125	CT	TC	TC	TC	TC	TC	T	TC
rs3767641	C	C	C	T	C	T	T	T
rs3767640	T	T	T	C	T	C	C	C
rs3767639	T	T	T	T	T	T	T	T
rs386636127	TG	TG	TG	CA	TG	CA	C	CA
rs3835613	A-	A-	A-	AC	A-	AC	A	AC
rs72704062	C	C	C	G	C	G	C	C
rs11799952	T	C	C	C	C	C	C	C
rs844	A	G	G	G	G	G	G	G
rs552605155	A	G	A	A	A	A	A	A

	NA1287		NA12878	
rs780467580	T	T	T	A
rs747505037	C	C	C	G
rs1459475	C	C	C	C
rs3767642	T	C	C	C
rs1824385	G	G	G	G
rs4656325	G	G	G	G
rs10917643	T	C	C	C
rs529326020	C-	CA	CA	CA
rs6687928	T	T	T	T
rs796412550	A	T	T	T
rs10436883	T	C	C	C
rs56384835	G	A	A	A
rs12567467	G	A	A	A
rs12568594	A	T	T	T
rs10465555	T	G	G	G
rs55989922	C	C	C	T
rs6658657	G	A	A	A
rs6661332	A	T	T	T
rs6694919	A	C	C	C
rs10753599	C	T	T	T
rs145624007	A	G	G	G
rs6664704	T	C	C	C
rs4233377	T	T	T	T
rs2007773	G	A	G	**A**
rs2007763	C	C	C	C
rs55858328	G	G	G	T
rs2169051	C	T	T	T
rs60081850	C	T	T	T
rs4657086	A	G	G	G
rs12145090	T	C	C	C
rs1383318848	T-	TG	TG	TG
rs7539744	C	A	A	A
rs2333845	A	A	A	A
rs1984769	G	A	A	A
rs922087	C	C	C	C
rs1344438491	G	T	T	T
rs11811662	A	G	G	G
rs4999919	C	C	C	C
rs182968886	A	G	G	G
rs2125684	C	C	C	C
rs2125685	A	A	A	A
rs2045571	C	T	T	T
rs2793082	C	C	C	C
rs2793081	T	T	T	T
rs1674755	A	A	A	A
rs369224649	A	A	A	A
rs1674756	A	A	A	A
rs1771554	T	T	T	T
rs372700655	A	A	A	A
rs1050501	T	C	C	C
rs6666965	G	T	T	T
rs1771556	TG	TG	TG	TG
rs1674757	A	T	T	A
161644663	C	C	C	C
161644713	G	G	G	G
rs7532925	G	A	A	A
rs1674758	C	C	C	C
rs6427615	C	A	A	A
rs1674759	T	G	G	G
rs1674760	C	T	T	T
rs1674761	A	C	C	C
rs13376485	G	A	A	A
rs386636125	CT	TC	TC	TC
rs3767641	C	T	T	T
rs3767640	T	C	C	C
rs3767639	T	T	T	T
rs386636127	TG	CA	CA	CA
rs3835613	A-	AC	AC	AC
rs72704062	C	G	C	G
rs11799952	T	C	C	C
rs844	A	G	G	G
rs552605155	A	A	A	A

	NA12879		NA12880		NA12881		NA12882	
rs780467580	T	T	T	A	T	A	T	A
rs747505037	C	C	C	G	C	G	C	G
rs1459475	C	C	C	C	C	C	C	C
rs3767642	C	C	T	C	C	C	C	C
rs1824385	G	G	G	G	G	G	G	G
rs4656325	G	G	G	G	G	G	G	G
rs10917643	C	C	T	C	C	C	C	C
rs529326020	CA	CA	C-	CA	CA	CA	CA	CA
rs6687928	T	T	T	T	T	T	T	T
rs796412550	T	T	A	T	T	T	T	T
rs10436883	C	C	T	C	C	C	C	C
rs56384835	A	A	G	A	A	A	A	A
rs12567467	A	A	G	A	A	A	A	A
rs12568594	T	T	A	T	T	T	T	T
rs10465555	G	G	T	G	G	G	G	G
rs55989922	C	C	C	T	C	T	C	T
rs6658657	A	A	G	A	A	A	A	A
rs6661332	T	T	A	T	T	T	T	T
rs6694919	C	C	A	C	C	C	C	C
rs10753599	T	T	C	T	T	T	T	T
rs145624007	G	G	A	G	G	G	G	G
rs6664704	C	C	T	C	C	C	C	C
rs4233377	T	T	T	T	T	T	T	T
rs2007773	A	A	G	A	A	A	A	A
rs2007763	C	C	C	C	C	C	C	C
rs55858328	G	G	G	T	G	T	G	T
rs2169051	T	T	C	T	T	T	T	T
rs60081850	T	T	C	T	T	T	T	T
rs4657086	G	G	A	G	G	G	G	G
rs12145090	C	C	T	C	C	C	C	C
rs1383318848	TG	TG	T-	TG	TG	TG	TG	TG
rs7539744	A	A	C	A	A	A	A	A
rs2333845	A	A	A	A	A	A	A	A
rs1984769	A	A	G	A	A	A	A	A
rs922087	C	C	C	C	C	C	C	C
rs1344438491	T	T	G	T	T	T	T	T
rs11811662	G	G	A	G	G	G	G	G
rs4999919	C	C	**C**	C	C	C	C	C
rs182968886	G	G	A	G	G	G	G	G
rs2125684	C	C	C	C	C	C	C	C
rs2125685	A	A	A	A	A	A	A	A
rs2045571	T	T	C	T	T	T	T	T
rs2793082	C	C	C	C	C	C	C	C
rs2793081	T	T	T	T	T	T	T	T
rs1674755	A	A	A	A	A	A	A	A
rs369224649	A	A	A	A	A	A	A	A
rs1674756	A	A	A	A	A	A	A	A
rs1771554	T	T	T	T	T	T	T	T
rs372700655	A	A	A	A	A	A	A	A
rs1050501	C	C	T	C	C	C	C	C
rs6666965	T	T	G	T	T	T	T	T
rs1771556	TG	TG	TG	TG	TG	TG	TG	TG
rs1674757	T	T	A	T	T	T	T	T
161644663	C	C	C	C	C	C	C	C
161644713	G	G	G	G	G	G	G	G
rs7532925	A	A	G	A	A	A	A	A
rs1674758	C	C	C	C	C	C	C	C
rs6427615	A	A	C	A	A	A	A	A
rs1674759	G	G	T	G	G	G	G	G
rs1674760	T	T	C	T	T	T	C	C
rs1674761	C	C	A	C	C	C	A	A
rs13376485	A	A	G	A	A	A	G	G
rs386636125	TC	TC	CT	TC	TC	TC	TC	TC
rs3767641	T	T	C	T	T	T	T	T
rs3767640	C	C	T	C	C	C	C	C
rs3767639	T	T	T	T	T	T	T	T
rs386636127	CA	CA	TG	CA	CA	CA	CA	CA
rs3835613	AC	AC	A-	AC	AC	AC	AC	AC
rs72704062	C	**G**	C	G	G	G	G	G
rs11799952	C	C	T	C	C	C	C	C
rs844	G	G	A	G	G	G	G	G
rs552605155	A	A	A	A	A	A	A	A

	NA12883		NA12884		NA12885		NA12886	
rs780467580	T	T	T	T	T	T	T	T
rs747505037	C	C	C	C	C	C	C	C
rs1459475	C	C	C	C	C	C	C	C
rs3767642	T	C	C	C	C	C	C	C
rs1824385	G	G	G	G	G	G	G	G
rs4656325	G	G	G	G	G	G	G	G
rs10917643	T	C	C	C	C	C	C	C
rs529326020	C-	CA	CA	CA	CA	CA	CA	CA
rs6687928	T	T	T	T	T	T	T	T
rs796412550	A	T	T	T	T	T	T	T
rs10436883	T	C	C	C	C	C	C	C
rs56384835	G	A	A	A	A	A	A	A
rs12567467	G	A	A	A	A	A	A	A
rs12568594	A	T	T	T	T	T	T	T
rs10465555	T	G	G	G	G	G	G	G
rs55989922	C	C	C	C	C	C	C	C
rs6658657	G	A	A	A	A	A	A	A
rs6661332	A	T	T	T	T	T	T	T
rs6694919	A	C	C	C	C	C	C	C
rs10753599	C	T	T	T	T	T	T	T
rs145624007	A	G	G	G	G	G	G	G
rs6664704	T	C	C	C	C	C	C	C
rs4233377	T	T	T	T	T	T	T	T
rs2007773	G	A	G	A	G	A	G	A
rs2007763	C	C	C	C	C	C	C	C
rs55858328	G	G	G	G	G	G	G	G
rs2169051	C	T	T	T	T	T	T	T
rs60081850	C	T	T	T	T	T	T	T
rs4657086	A	G	G	G	G	G	G	G
rs12145090	T	C	C	C	C	C	C	C
rs1383318848	T-	TG	TG	TG	TG	TG	TG	TG
rs7539744	C	A	A	A	A	A	A	A
rs2333845	A	A	A	A	A	A	A	A
rs1984769	G	A	A	A	A	A	A	A
rs922087	C	C	C	C	C	C	C	C
rs1344438491	G	T	T	T	T	T	T	T
rs11811662	A	G	G	G	G	G	G	G
rs4999919	C	C	C	C	C	C	C	C
rs182968886	A	G	G	G	G	G	G	G
rs2125684	C	C	C	C	C	C	C	C
rs2125685	A	A	A	A	A	A	A	A
rs2045571	C	T	T	T	T	T	T	T
rs2793082	C	C	C	C	C	C	C	C
rs2793081	T	T	T	T	T	T	T	T
rs1674755	A	A	A	A	A	A	A	A
rs369224649	A	A	A	A	A	A	A	A
rs1674756	A	A	A	A	A	A	A	A
rs1771554	T	T	T	T	T	T	T	T
rs372700655	A	A	A	A	A	A	A	A
rs1050501	T	C	C	C	C	C	C	C
rs6666965	G	T	T	T	T	T	T	T
rs1771556	TG	TG	TG	TG	TG	TG	TG	TG
rs1674757	A	T	T	T	T	T	T	T
161644663	C	C	C	C	C	C	C	C
161644713	G	G	G	G	G	G	G	G
rs7532925	G	A	A	A	A	A	A	A
rs1674758	C	C	C	C	C	C	C	C
rs6427615	C	A	A	A	A	A	A	A
rs1674759	T	G	G	G	G	G	G	G
rs1674760	C	T	T	T	T	T	T	T
rs1674761	A	C	C	C	C	C	C	C
rs13376485	G	A	A	A	A	A	A	A
rs386636125	CT	TC	TC	TC	TC	TC	TC	TC
rs3767641	C	T	T	T	T	T	T	T
rs3767640	T	C	C	C	C	C	C	C
rs3767639	T	T	T	T	T	T	T	T
rs386636127	TG	CA	CA	CA	CA	CA	CA	CA
rs3835613	A-	AC	AC	AC	AC	AC	AC	AC
rs72704062	C	G	C	G	C	G	C	G
rs11799952	T	C	C	C	C	C	C	C
rs844	A	G	G	G	G	G	G	G
rs552605155	A	A	A	A	A	A	A	A

	NA12887		NA12888		NA12893	
rs780467580	T	A	T	A	T	A
rs747505037	C	G	C	G	C	G
rs1459475	C	C	C	C	C	C
rs3767642	C	C	T	C	T	C
rs1824385	G	G	G	G	G	G
rs4656325	G	G	G	G	G	G
rs10917643	C	C	T	C	T	C
rs529326020	CA	CA	C-	CA	C-	CA
rs6687928	T	T	T	T	T	T
rs796412550	T	T	A	T	A	T
rs10436883	C	C	T	C	T	C
rs56384835	A	A	G	A	G	A
rs12567467	A	A	G	A	G	A
rs12568594	T	T	A	T	A	T
rs10465555	G	G	T	G	T	G
rs55989922	C	T	C	T	C	T
rs6658657	A	A	G	A	G	A
rs6661332	T	T	A	T	A	T
rs6694919	C	C	A	C	A	C
rs10753599	T	T	C	T	C	T
rs145624007	G	G	A	G	A	G
rs6664704	C	C	T	C	T	C
rs4233377	T	T	T	T	T	T
rs2007773	A	A	G	A	G	A
rs2007763	C	C	C	C	C	C
rs55858328	G	T	G	T	G	T
rs2169051	T	T	C	T	C	T
rs60081850	T	T	C	T	C	T
rs4657086	G	G	A	G	A	G
rs12145090	C	C	T	C	T	C
rs1383318848	TG	TG	T-	TG	T-	TG
rs7539744	A	A	C	A	C	A
rs2333845	A	A	A	A	A	A
rs1984769	A	A	G	A	G	A
rs922087	C	C	C	C	C	C
rs1344438491	T	T	G	T	G	T
rs11811662	G	G	A	G	A	G
rs4999919	C	C	C	C	C	C
rs182968886	G	G	A	G	A	G
rs2125684	C	C	C	C	C	C
rs2125685	A	A	A	A	A	A
rs2045571	T	T	C	T	C	T
rs2793082	C	C	C	C	C	C
rs2793081	T	T	T	T	T	T
rs1674755	A	A	A	A	A	A
rs369224649	A	A	A	A	A	A
rs1674756	A	A	A	A	A	A
rs1771554	T	T	T	T	T	T
rs372700655	A	A	A	A	A	A
rs1050501	C	C	T	C	T	C
rs6666965	T	T	G	T	G	T
rs1771556	TG	TG	TG	TG	TG	TG
rs1674757	T	T	A	T	A	T
161644663	C	C	C	C	C	C
161644713	G	G	G	G	G	G
rs7532925	A	A	G	A	G	A
rs1674758	C	C	C	C	C	C
rs6427615	A	A	C	A	C	A
rs1674759	G	G	T	G	T	G
rs1674760	T	T	C	T	C	T
rs1674761	C	C	A	C	A	C
rs13376485	A	A	G	A	G	A
rs386636125	TC	TC	CT	TC	CT	TC
rs3767641	T	T	C	T	C	T
rs3767640	C	C	T	C	T	C
rs3767639	T	T	T	T	T	T
rs386636127	CA	CA	TG	CA	TG	CA
rs3835613	AC	AC	A-	AC	A-	AC
rs72704062	G	G	C	G	C	G
rs11799952	C	C	T	C	T	C
rs844	G	G	A	G	A	G
rs552605155	A	A	A	A	A	A

Appendix 17. The list of samples used in the population genetics analysis				
Haplotype Codes	Sample ID	Pop Code	S Pop. Code	Population Description
CHS1, CHS2	HG00419	CHS	EAS	Southern Han Chinese
CDX1, CDX2	HG00759	CDX	EAS	Chinese Dai in Xishuananna, China
KHV1, KHV2	HG01595	KHV	EAS	Kinh in Ho Chi Minh City, Vietnam
CHB1, CHB2	NA18525	CHB	EAS	Han Chinese in Bejing, China
CHB1_P	NA18555	CHB	EAS	Han Chinese in Bejing, China
JPT1, JPT2	NA18939	JPT	EAS	Japanese in Tokyo, Japan
JPT1_P, JPT2_P	NA18956	JPT	EAS	Japanese in Tokyo, Japan
GBR1, GBR2	HG00096	GBR	EUR	British in England and Scotland
FIN1, FIN2	HG00268	FIN	EUR	Finnish in Finland
IBS1, IBS2	HG01500	IBS	EUR	Iberian Population in Spain
CEU1_FF, CEU2_FF	NA12889	CEU	EUR	Utah Residents (CEPH) with N. and W. Ancestry
CEU1_FM, CEU2_FM	NA12890	CEU	EUR	
CEU1_MF, CEU2_MF	NA12891	CEU	EUR	
CEU1_MM, CEU2_MM	NA12892	CEU	EUR	
TSI1, TSI2	NA20502	TSI	EUR	Toscani in Italia
YRI1_F, YRI2_F	NA18507	YRI	AFR	Yoruba in Ibadan, Nigeria
YRI1_M, YRI2_M	NA18517	YRI	AFR	Yoruba in Ibadan, Nigeria
YRI1_C2, YRI2_C2	NA19129	YRI	AFR	Yoruba in Ibadan, Nigeria
YRI1_C1, YRI2_C1	NA19240	YRI	AFR	Yoruba in Ibadan, Nigeria
ESN1, ESN2	HG02922	ESN	AFR	Esan in Nigeria
ACB1, ACB2	HG01879	ACB	AFR	African Caribbeans in Barbados
GWD1, GWD2	HG02568	GWD	AFR	Gambian in Western Divisions in the Gambia
MSL1, MSL2	HG03052	MSL	AFR	Mende in Sierra Leone
LWK1, LWK2	NA19017	LWK	AFR	Luhya in Webuye, Kenya
ASW1, ASW2	NA19625	ASW	AFR	Americans of African Ancestry in SW USA
PUR1, PUR2	HG01051	PUR	AMR	Puerto Ricans from Puerto Rico
CLM1, CLM2	HG01112	CLM	AMR	Colombians from Medellin, Colombia
PEL1, PEL2	HG01565	PEL	AMR	Peruvians from Lima, Peru
MXL1, MXL2	NA19648	MXL	AMR	Mexican Ancestry from LA, USA
PJL1, PJL2	HG01583	PJL	SAS	Punjabi from Lahore, Pakistan
BEB1, BEB2	HG03006	BEB	SAS	Bengali from Bangladesh
STU1, STU2	HG03642	STU	SAS	Sri Lankan Tamil from the UK
ITU1, ITU2	HG03742	ITU	SAS	Indian Telugu from the UK
GIH1, GIH2	NA20845	GIH	SAS	Gujarati Indian from Houston, Texas