Human alpha defensin CNV haplotype diversity

Xiao Xu, BSc. MSc.
Thesis submitted to The University of Nottingham
for the degree of Doctor of Philosophy

October 2017

Abstract

Humans have highly variable number of alpha defensin genes, with between 3-16 diploid copies. Alpha-defensin genes have important roles in human innate immunity and diseases. Recently, GWAS studies reported this locus associated with IgA nephropathy and periodontitis. However, the underlying mechanism of association is not clear. In this Ph.D. thesis, human alpha defensin CNV flanking haplotype diversity in global populations was studied and the association between diseases and haplotype classes was discussed. Then a novo method to detect variants from inside the *DEFA1A3* CNV was developed and a list of potential disease-related mutations for further functional studies was generated. The association between CNV internal variants and flanking haplotype classes was studied. Non-allelic homologous recombination was found to be the major mechanism of CNV formation of alpha defensin CNV. Analysis results were verified by PCR and Sanger sequencing-based methods. Additional to that, the haplotype diversity analysis highlighted an unusual haplotype 5T/7C which is only found in European populations but highly diverged from other human haplotypes. Further evidence was provided to suggest that this is an introgressed haplotype from Neanderthals. Furthermore, we used Oxford Nanopore to reconstruct haplotype structure in *DEFA1A3* CNV and discussed its advantages and limitations by our analysis results. In brief, this Ph.D. research

greatly improved our understanding of *DEFA1A3* global diversity, evolutionary history, diseases and haplotype association.

Acknowledgements

First and foremost I offer my sincerest gratitude to my supervisor Prof John Armour who provide me with this PhD research opportunity and give me countless guidance, advice and help in both aspect of academic and career development.

Thanks to all present and previous JALA lab members: Nzar Shwan, Eric Moise, Paula Necsoiu, Jess Tyson, Holly Black, Liaque Latif, Ivan Stetsenko, Sugandha Dhar, Omniah Mansouri, Danielle Carpenter, Laura Mitchell, Catherine, Harrison, Andrew Cubbon. Your help and advices contributed a great proportion to this PhD thesis.

Thanks to all my family members, especially to my parents Dr Jie Xu and Ailan Cao for their encouragement and support, both emotionally and financially, without them I will achieve nothing.

Also Thanks to Bill wickstead from university of Nottingham, UK. XueQing Yu's group at Sun Yat-Sen University China, JianJun Liu's group at IGS Singapore. Schäfer Arne's group at Charité – Universitätsmedizin Berlin, Jonathan Barratt's group at university of Leicester for their collaboration advice and discussion during this research work.

At last a special thanks to Miss Yang Wang for her love, support, and accompany during my way to pursuing this PhD degree.

Contents

1 Chapter1: Introduction.....11

1.1 Copy number variation 11

1.2 Human alpha defensin CNV 25

1.3 Diseases associated with *DEFA1A3* CNV.....31

1.4 Neanderthal Introgression.....42

1.5 DNA Sequencing 46

1.6 Aims of this research.....51

2 Chapter2: Materials and methods 53

2.1 Research Materials.....53

2.2 Bioinformatics methods 55

2.3 Experimental methods.....77

3 Chapter3: Human *DEFA1A3* CNV flanking haplotype diversity ...93

3.1 Introduction 93

3.2 Global *DEFA1A3* flanking haplotype analysis.....96

3.3 Human *DEFA1A3* Flanking haplotype diversity 101

3.4 New haplotype classes vs Black's haplotype classes 105

3.5 Tag SNPs for *DEFA1A3* CNV flanking haplotypes 107

3.6 Recombinant haplotypes 110

3.7 Phylogenetic analysis 112

3.8 Experimental verification of recombinant haplotypes 114

3.9 Conclusion 117

4 Chapter4: *DEFA1A3* haplotypes associated with diseases 119

4.1 Introduction 119

4.2 Haplotypes associated with IgA Nephropathy 120

4.3 Haplotypes associated with periodontitis 127

4.4 Looking for causative variants inside *DEFA1A3* CNV 131

4.5 Conclusion 146

5 Chapter5: haplotype 5T/7C as a Neanderthal introgressed haplotype 147

5.1 Introgression of haplotype 5T/7C from Neanderthals 147

5.2 Haplotype 5T/7C is highly similar to Altai Neanderthal haplotype 148

5.3 Copy number measurement of haplotype 5T/7C 150

5.4 Neanderthal derived variants and introgression length..... 151

5.5 Geographic distribution of 5T/7C haplotype..... 153

5.6 Evidence from S* statistics 157

5.7 Evidence from divergence time estimation 159

5.8 Selection test for the 5T/7C haplotype..... 166

5.9 Neanderthal derived variants inside CNV..... 173

5.10 Conclusion 175

6 Chapter6: Construction *DEFA1A3* CNV haplotype structure by Oxford Nanopore MinION sequencing 177

6.1 Introduction to Oxford Nanopore MinION sequencing 177

6.2 Experiment design to construct *DEFA1A3* CNV haplotype structure..... 180

6.3 Problems and limitations for *DEFA1A3* haplotype reconstruction 182

6.4 Conclusions..... 191

7 Chapter 7: Discussion 192

7.1 Functional variants inside DEFA1A3 CNV 192

7.2 Recovering CNV internal variants from next generation sequencing data 193

7.3 Evolutionary history of *DEFA1A3* CNV 195

7.4 CNV internal features of Neanderthal introgressed haplotype 196

List of Abbreviations

AgP	Aggressive Periodontitis
ANI	Ancestral North Indian
BWA	Burrow-Wheeler Aligner
CGH	Comparative Genomic Hybridization
CN	Copy Number
CNV	Copy Number Variation
CP	Chronic Periodontitis
DGV	Database Of Genomic Variants
dNTP	Deoxynucleotide Triphosphate
DSBs	DNA Double-Strand Breaks
EHF-PCR	Emulsion Haplotype Fusion-PCR
EHH	Extended Haplotype Homozygosity
EN	Endonuclease
ESRD	End-Stage Renal Disease
FGS	First Generation Sequencing
FISH	Fluorescent In Situ Hybridization
FoSTeS	Fork Stalling And Template Switching
GATK	Genome Analysis Toolkit
GWAS	Genome Wide Association Study
HPD	Highest Posterior Density
HR	Homologous Recombination
IgAN	Iga Nephropathy
iHH	Integrated Extended Haplotype Homozygosity
iHS	Integrated Haplotype Score
ILS	Incomplete Lineage Sorting
LD	Linkage Disequilibrium
MLPA	Multiplex Ligation-Dependent Probe Amplification

MMEJ	Microhomology-Mediated End Joining
MSV	Multi Site Variants
MY	Million Years
NAHR	Non-Allelic Homologous Recombination
NEB	New England Bio Lab
NGS	Next Generation Sequencing
NHEJ	Non-Homologous End Joining
PCR	Polymerase Chain Reaction
PRT	Paralogue Ratio Test
qPCR	Quantitative Polymerase Chain Reaction
RT	Reverse Transcriptase
SDSA	Synthesis-Dependent Strand-Annealing
SGS	Second Generation Sequencing
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variants
SV	Structural Variants
TGS	Third Generation Sequencing
TMRCA	Time To Most Recent Common Ancestor
UTR	Untranslated Region
ZMW	Zero-Mode Waveguide

1 Chapter1: Introduction

1.1 Copy number variation

1.1.1 Human genetic variation

Since the completion of the Human Genome Project in 2001 (Lander et al., 2001), human variation has become a major research focus in the field of human genetics, leading to human variation research projects like the International HapMap Project and the Human 1000 Genomes Project. The Phase 1 International HapMap Project genotyped more than one million human variants in four populations from Africa, Europe and Asia (The International HapMap Consortium, 2003). This was followed by the human 1000 Genomes Project which used next generation sequencing (NGS) technologies to sequence whole genomes and whole exomes of 2504 individuals from 26 world populations. It allowed discovery of most human variants with frequencies as low as 1% (The 1000 Genomes Project Consortium, 2010). These data made a great contribution to our understanding of human variation and benefited research in areas such as genotype-phenotype association studies, human genetic disorder and disease studies, and human evolution studies.

1.1.2 Types of human genetic variation

According to size, human genetic variation can be classified into single nucleotide polymorphism (SNP) and structural variants (SV). SNP involves substitution of a single nucleotide base to an alternative base; they are the most common type of variation. In humans, there are 135 million validated SNPs reported according to the dbSNP 150 database (Feb 03, 2017). SNPs can have a profound functional effects depending on their location.

Human structural variants (SV) can be further classified according to length. The largest type of SV is chromosome rearrangement, such as inversion and translocation. This is then followed by large copy number variation (CNV) which is defined as the deletion, insertion or duplication of sections of the genome of DNA ≥ 1kb in length. They are repeated and the number of repeats in the genome varies between individuals in the human population (Mccarroll and Altshuler, 2007)(5). Indels are small SVs that involve insertion or deletion of DNA sections less than 1kb in length(Feuk et al., 2006) , it also including single nucleotide insertions and deletions. Apart from those, highly variable small tandem repeats such as microsatellites and minisatellites also make up a great part of human variation. Minisatellites are 10–60 base pairs in repeat unit length and

microsatellites are usually 2-10bp. Both of them are highly repeated sequences, typically repeated 5-50 times.

1.1.3 Copy number variation

In the human genome, one of the most important types of variation is CNV. It is estimated that two thirds of the human genome is composed of repeated sequences of some type (de Koning et al., 2011). Overall, 4.8-9.5% of the human genome can be categorized as involved in copy number variation (Zarrei et al., 2015). According to the Database of Genomic Variants (DGV), there are over 7 million CNVs of size greater than 50bp in the human genome. Most CNVs reported are between 1kb and 10kb.

CNVs have important roles in human phenotypic variation and diseases. It has been reported that high copy number of the salivary amylase gene (AMY1) is associated with high-starch diets in humans and dogs (Perry et al., 2007, Axelsson et al., 2013). Copy number polymorphism of the *FCGR3* gene has been demonstrated associated with lupus glomerulonephritis(Aitman et al., 2006). Deletions and reciprocal duplications of the chromosome 16p13.1 region have also been reported to be associated with schizophrenia(Ingason et al., 2011), autism and mental retardation(Ullmann et al., 2007).

Recently various CNV detection and genotyping methods have been developed. They can be classified into three main categories: (a) Hybridization based methods such as Fibre-FISH**,** array CGH; (b) PCR based methods such as qPCR, PRT, and MLPA. (c) Various NGS sequencing depth based methods.

1.1.3.1 Fibre fluorescent in situ hybridization (Fibre-FISH)

Fluorescent in situ hybridization (FISH) is a technique used for detecting and localizing DNA sequences in a genome. It involves binding of fluorescent probes to complementary DNA sequences. Regions of probe binding can be directly observed under fluorescence microscopy. Fibre-FISH is defined as the application of FISH techniques on a completely relaxed DNA double helix. In most cases, the target DNA is prepared by deproteinizing chromosomal chromatin by extraction with detergents or high salt concentration (Ersfeld, 2004). It can be used to measure copy number in a CNV region. Fibre-FISH fluorescent probes are designed to bind with every repeat unit. The number of repeat units can be directly counted under a microscope. It has a resolution of approximately 1000bp.

1.1.3.2 Array CGH

Comparative genomic hybridization (CGH) was initially developed for comparing the copy number of a test genome with a reference diploid genome by using fluorescence in situ hybridization (FISH) on metaphase chromosomes. Although it has had successful applications in studies of cancer genome copy number variation, the major drawback was the low resolution of metaphase chromosomes, typically about 5–10 Mb (Carter, 2007). With the completion of the Human genome project, the metaphase genome used by CGH was replaced by an array of test and reference oligonucleotides with known copy number. The resolution of CGH was decided by the size and density of probes used in the array. In 2004 Iafrate AJ, et al used array CGH to compare the genomes of normal individuals, with one BAC clone every 1 Mb across the human genome; they detected 255 CNVs in 55 individuals (Iafrate, 2004) and demonstrated the importance of CNV in human variation.

1.1.3.3 Quantitative polymerase chain reaction (qPCR)

Quantitative polymerase chain reaction (qPCR) is also called real-time polymerase chain reaction (Real-Time PCR). It is a kind of polymerase chain reaction which monitors the amplification of a targeted DNA molecule by labeled fluorescent dyes in real time during the PCR reaction. Therefore data is collected throughout the process

rather than collected after PCR reaction. There are generally two types of chemistries used by qPCR, non-specific detection fluorescence dyes such as SYBR Green which binds to all kinds of double stranded DNA molecule, and sequence-specific probes such as TaqMan. A TaqMan probe is a specifically designed DNA oligonucleotide containing a fluorophore covalently attached to the 5’-end and a quencher at the 3’-end (Debski et al., 2017). The quencher molecule quenches the fluorescence emitted by the fluorophore and inhibits any fluorescence signals. During the PCR annealing step both primers and the TaqMan probe anneal with template DNA molecules. During the extension step, when Taq polymerase extends the primer and reaches the TaqMan probe, the 5' to 3' exonuclease activity of the Taq polymerase degrades the probe, releases the fluorophore and quencher, breaking the close proximity of fluorophore and quencher, thus emitting fluorescence signal. Initial DNA template copy number can be determined according to the cycle number before an increase in fluorescence above a threshold is observed. In CNV detection it compares the amount of PCR product made from a test amplicon with the yield from a reference locus in the same multiplex reaction (Armour et al., 2007).

1.1.3.4 MLPA (Multiplex Ligation-dependent Probe Amplification)

MLPA is a multiplex PCR based method. It has the ability to detect the copy number of a maximum of 50 different genomic DNA or RNA targets (Schouten et al., 2002). One MLPA probe is designed to have two oligonucleotides, each pair containing primer sequences that can hybridize with template DNA and amplify products with different sizes. An MLPA reaction consists of five steps (Fig 1.1). First, DNA is denatured by heating in high temperature and mixed with MLPA probes and incubated overnight. Then the two hybridized MLPA probes can be ligated together followed by PCR amplification of ligated probes. Amplicons are then separated by capillary electrophoresis and their yield compared with reference peaks with known copy number. The fragment peak yield is proportional to the copy number of initial template DNA. If one of the primers fails to hybridize, the amplifiable probe will not be exponentially amplified in the PCR reaction, and thus no signal is generated by this probe.

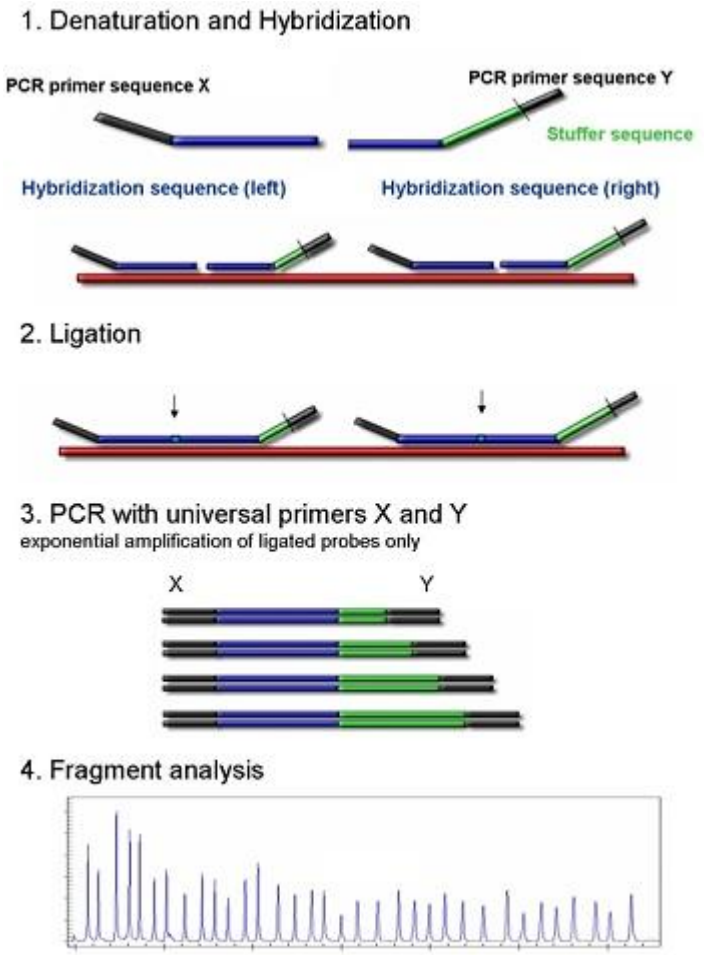


Figure 1.1 Illustration of MLPA reaction processes and capillary electrophoresis analysis result. (From http://mlpa.com/WebForms/)

1.1.3.5 Paralogue ratio test (PRT)

Paralogue ratio test (PRT) determines the representation of a test locus relative to a co-amplified reference locus. The primers are precisely designed to amplify from a copy of the element within the variable repeat unit, plus exactly one other unlinked reference locus (Armour et al., 2007). Unlike the other quantitative PCRs, PRT uses a single pair of primers to amplify both test and reference locus, but the two products differ in size, and can be separated by capillary

electrophoresis. The ratio between test and reference locus can be determined by calculating peak area or height (Hollox, 2017). PRT is a fast, accurate and high throughput method for genotyping CNV; it has been successfully used in the study of the copy number variation of alpha-defensins (Khan et al., 2013), the beta-defensin cluster (Armour et al., 2007), *CCL3L1*/*CCL4L1* (Carpenter et al., 2011), Amylase region (Carpenter et al., 2015), Complement C4 (Fernando et al., 2010) and FCGR3 (Hollox et al., 2009).

1.1.3.6 NGS Read-depth based methods

With the fast development of next generation sequencing technologies, massive amounts of short read sequencing data were generated. Coverage or read depth is the number of short reads that can be matched to a particular position on the reference genome. The read depth theoretically is proportional to the dosage of DNA sequenced, and thus CNV can be detected from the read depth change in a region. Several tools have been developed to detect CNV in humans from genome or exome sequencing data, for example CONTRA (Li et al., 2012), CoNVEX (Amarasinghe et al., 2013), CoNIFER (Krumm et al., 2012), Alex Nord script (Nord et al., 2011) and CANOES (Backenroth et al., 2014). Although different in statistical approaches, they are all based on similar data processing procedures. Firstly, coverage needs to be corrected according to GC

content. Then coverage is normalized and a baseline is generated from multiple sample comparison. Finally, CNV are called according to coverage ratio between test sample and baseline or reference samples. These methods generally have a higher specificity and sensitivity for detecting simple 1-2 copies duplication and deletions which size bigger than one exon. NGS based methods cannot easily handle CNVs with highly polymorphic copy number regions since these methods relied on the accuracy of the reference genome where short reads mapped. Recently a software Genome STRiP was developed for detect large multiallelic copy number variations (Handsaker et al., 2015).

1.1.4 Molecular mechanism of CNV formation

Copy number variation can be generated by four mechanisms: NHEJ, NAHR, FoSTeS and L1-mediated retrotransposition (Feng Zhang, 2009).

1.1.4.1 Non-homologous end joining (NHEJ)

DNA double-strand breaks (DSBs) are normal consequences of cell division and differentiation and must be repaired to maintain genome stability. They can be repaired through three mechanisms: Non-homologous end joining (NHEJ), Microhomology-mediated end joining (MMEJ) and homologous recombination (HR). NHEJ is referred as

"non-homologous" since it repairs the damaged DNA without the help of homologous sequence, and differs from homology directed repair in contrast. It can repair the DSB accurately or lead to 1-4bp deletion or in some cases cause insertion of additional DNA (fig 1.2) (Hastings et al., 2009). MMEJ is an alternative double strand broken repair mechanism which is characterized by using 5-25bp long microhomologous sequences in the alignment of the broken end before joining. MMEJ always causes deletion of flanking sequences of the original break (Mitch and Sang 2008).

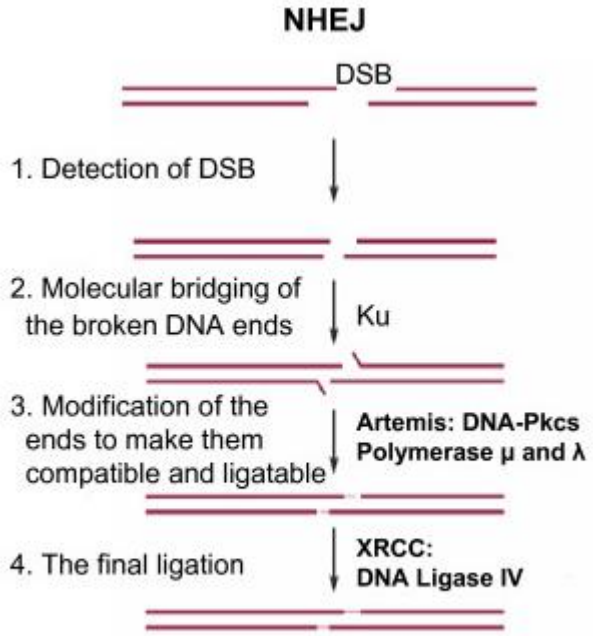


Figure 1.2 Illustration of DNA double-strand breaks repair by NHEJ

1.1.4.2 Non-allelic homologous recombination (NAHR)

Homologous recombination (HR) is another DNA double strand break repair mechanism which relies on homologous sequence. It happens during meiosis and mitosis. During the repair processes, the 5’ end of a DSB is resected and leaves a 3’ overhang which can invade a homologous sequence. Holliday junction formation between homologous sequences can lead to crossing over, gene conversion or synthesis-dependent strand-annealing (SDSA). When HR happens between non-allelic homologous sequences (NAHR), duplication, deletion and inversion events may occur. NAHR can lead to recurrent changes in copy number and form a large copy number variable region. It can happen between homologue chromosomes, between sister chromatids or in an intra-chromatid manner (Fig 1.3).

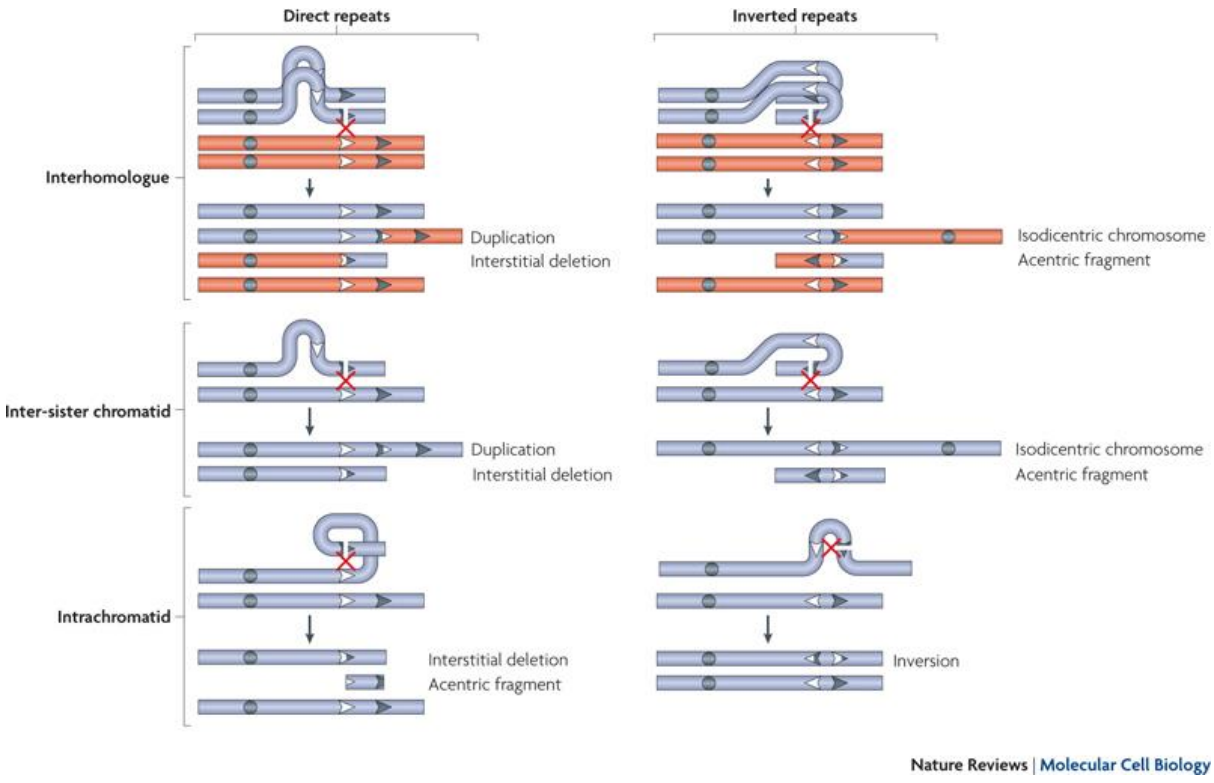


Figure 1.3 Mechanism of NAHR. The process can be inter-homologue chromosomes, inter-sister chromatid or intra-chromatid, leading to duplication, deletion, or inversion.

1.1.4.3 Fork Stalling and Template Switching (FoSTeS)

The FoSTeS (Fork Stalling and Template Switching) model was proposed as a mechanism underlying complex human genomic rearrangements. According to Lee et al (Lee et al., 2007) during DNA replication, the DNA replication fork can stall (Fig 1.4 A); then the lagging strand disengages, and invades an adjacent active replication fork. It anneals to a region of microhomology 1 and extends (Fig 1.4 B). Then the lagging strand disengages again and anneals to another region with microhomology 2 (Fig 1.4 C). Eventually, the lagging

strand returns to the original replication fork and continues to finish the rest of replication, resulting in duplication (Fig 1.4 D). This model can explain regions with complex structural variants.

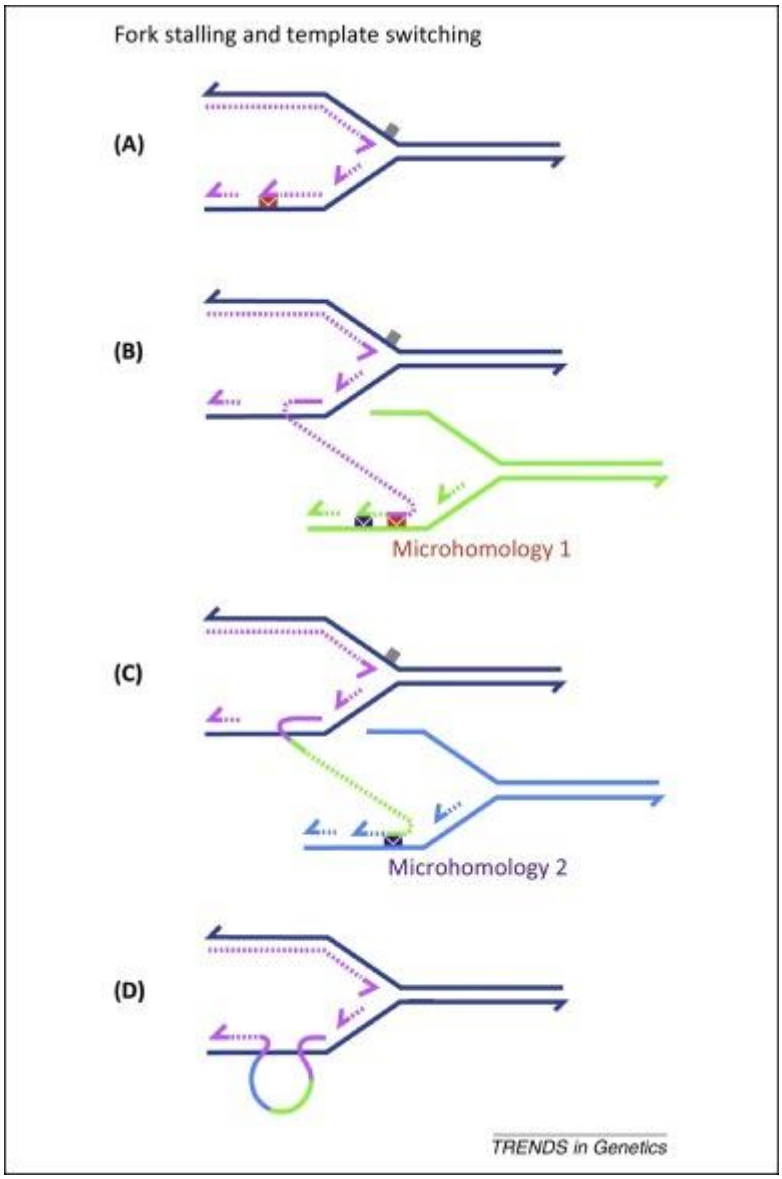


Figure 1.4 Illustration of complex duplication created through FoSTeS mechanism. Diagram from (Diego Ottaviani, 2014)

1.1.4.4 L1-mediated retrotransposition

Long interspersed element-1 (L1) constitutes 16.89% of the human genome (Haig et al., 1998). The full length L1 is about 6kb in size and includes two open reading frames; ORF1 encodes an RNA-binding protein and ORF2 encodes a protein with both endonuclease (EN) and reverse transcriptase (RT) activity (Babushok and Kazazian, 2007). In total 516,000 copies of L1 have been discovered in the human genome but most of them are partial copies. It is reported 30% of structural variants can be attributed to retrotransposition, among them 90% caused by L1 and 8% caused by SVA elements (Korbel et al., 2007). Another report also suggested that 15% of human structural variants were contributed by retrotransposition (Kidd et al., 2008).

1.2 Human alpha defensin CNV

1.2.1 Defensins

Defensins are 2-6 kDa, cationic, cysteine-rich, disulphide-bonded, peptides that participate in innate immunity. They are active against many Gram-negative and Gram-positive bacteria, fungi, and enveloped viruses (White et al., 1995). In humans they are found in three subfamilies: α, β and θ-defensins.

Beta defensins are the oldest defensins found in most vertebrates (Maxwell et al., 2003). They are synthesized in epithelial cells (Bedi et al., 2015). Alpha defensins are found in marsupials and most mammals except Laurasiatheria species. It is believed that alpha defensins are derived from beta defensins (Belov et al., 2007). Theta defensins are reported to be derived from alpha defensins. They are the youngest defensins, and can be found in old world monkeys and certain apes. In human, chimpanzee, and gorilla a nonsense mutation in the signal peptide turns the theta defensin gene into a pseudogene (Tang et al., 1999). The mouse alpha defensin region is located on mouse chromosome 8, and copy number variation has been detected. A study identified 53 active defensin genes and 22 defensin pseudogenes in the mouse; this indicates evolution, duplication and allelic variation of defensin genes is a feature in this region (Amid et al., 2009). Similar to the mouse, in primates, there are 41 functional and 22 non-functional defensin genes reported in a study involving 5 primate species (Das et al., 2010). Evolution of *α*-defensins in primates is governed by two opposing evolutionary forces: one force stabilizes specific amino acid residues and the other diversifies the sequences generating variation (Das et al., 2010).

1.2.2 Human alpha defensins

Defensins are peptides that participate in innate immunity. There are six types of alpha defensin peptides that have been characterized till now in humans. DEFA1 to DEFA4 are primarily kept in azurophilic granules of neutrophils and can make up to 30% to 50% of the total azurophilic granule proteins. Human neutrophil peptides are synthesized in neutrophil precursor cells called promyelocytes within the bone marrow (Lehrer and Lu, 2012). DEFA5 and DEFA6 are produced by Paneth cells of the small intestine (Ghosh et al., 2002), and are further secreted into the lumen of the small intestine where they are proteolytically cleaved into active antimicrobial peptides. The defensin gene cluster is located on human chromosome 8p23.1 (Fig 1.5). The cluster starts from *DEFA6* at chr8: 6782216, followed by *DEFA4*, *DEFA1A3* CNV and ends at *DEFA5* chr8: 6914259 according to the human reference genome hg19.

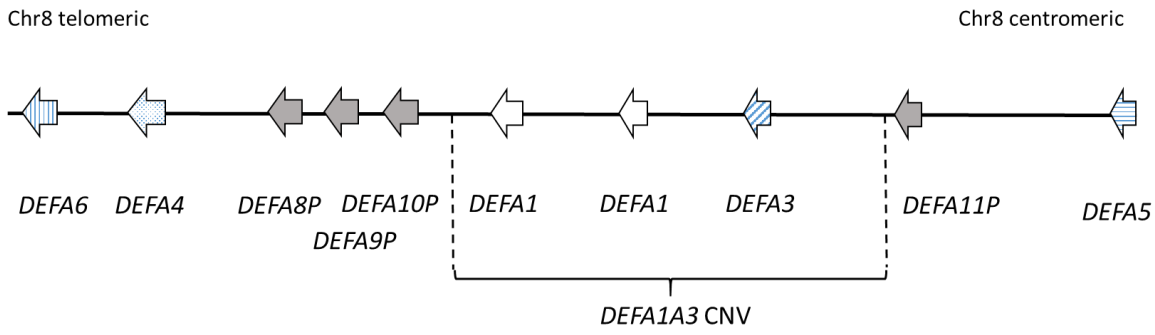


Figure 1.5 Illustration of the human alpha defensin gene cluster on chromosome 8p23.1 according to human reference genome hg19.

1.2.3 *DEFA1A3* CNV

The *DEFA1* and *DEFA3* genes are in a multiallelic CNV region, and individuals can have between 3-16 copies of *DEFA1* or *DEFA3* . The human reference genome hg19 has three copies (Fig 1.6), two full repeats each 19kb in length containing *DEFA1* gene and one 10kb partial repeat which contains the *DEFA3* gene. Intra-allelic rearrangement has been reported as the predominant mechanism responsible for the formation of this CNV (Black et al., 2014). This is a type of NAHR that happened within sister-chromatid or intra chromatid. DEFA1 and DEFA3 differ only by the first amino acid; this is due to a single nucleotide change in exon 3. DEFA2 lacks this residue, and there is no gene coding for DEFA2 that has been discovered (Aldred et al., 2005) and it is thought to derive from proteolytic cleavage of DEFA1 and/or DEFA3.

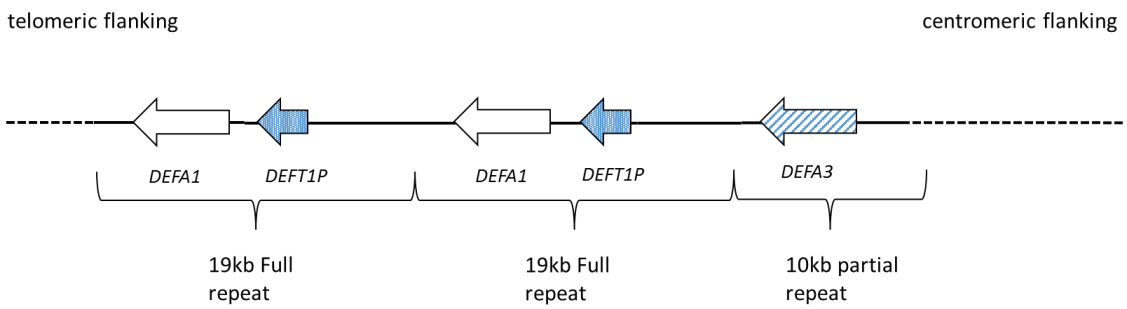


Figure 1.6 Structure of human *DEFA1A3* CNV according to the hg19 reference genome, containing two 19kb full repeats and one 10kb partial repeat. Each full repeat unit contains a *DEFA1* gene and a *DEFT1P* pseudo gene.

The multi-copy and copy-variable features of alpha defensin persist in humans, chimpanzees, and gorillas indicating that this feature has existed for at least 25 million years (Khan and Armour, 2010). On the other hand, *DEFA3* was found to be human specific and formed more recently, and some individuals lack the *DEFA3* gene completely. Approximately 10% of the UK population lack *DEFA3* and the function of DEFA3 in humans is still not known; one hypothesis suggested that it could be the precursor of biologically more active DEFA2, but it also possible that the gene is just an incipient pseudogene (Aldred et al., 2005).

Both *DEFA3* and *DEFA1* are expressed. The mRNA expression level per copy of *DEFA3* is approximately twice that of *DEFA1*. All *DEFA1* copies are apparently expressed at the same level, whereas in neutrophils no correlation was found between total copy number (both *DEFA1* and *DEFA3* copies) and combined expression level. This

may be due to these genes being modulated by an independent *trans*-acting factor (Aldred et al., 2005).

A recent study identified five classes of *DEFA1A3* haplotypes in European population by using 4.1kb centromeric flanking region SNPs, termed Reference, Class1, Class2, Exchange1, and Exchange2. Each haplotype has its own unique profile of features. The reference haplotype is so named because it is the same as the human reference genome (GRCh37/hg19). Exchange 1 haplotypes contain a 140bp centromeric partial repeat sequence replaced by an equivalent region of the full repeats through a gene conversion event. Similarly, Exchange2 haplotype contains a 1990bp replacement of centromeric partial repeat by the equivalent part of the full repeat. Both reference and exchange1 haplotypes usually have high *DEFA1A3* copy number, while class1 and class2 have lower copy number (Black et al., 2014). A small scale analysis of 1000 Genomes Project data showed that different populations in the world have different copy number distributions, Asians have significantly higher Class1 haplotype copy number, and low class2 copy number. Europeans have high reference haplotype copy number but low for Class2 and exchange2 haplotype copy number (Black et al., 2014). One SNP (rs4300027) near the *DEFA1A3* locus is associated with *DEFA1A3* copy number in European populations, its C allele associated with low total copy number,

usually 2-3 copies per haplotype, and the T allele associated with high *DEFA1A3* copies, usually 4-5 copies (Khan et al., 2013).

1.3 Diseases associated with *DEFA1A3* CNV

DEFA1 and *DEFA3* are important genes that participate in host immune response. This locus is associated with various diseases and phenotypes in humans. The high copy of *DEFA1A3* (8 copies) was found to be significantly associated with the risk of severe sepsis in the Chinese population (Chen et al., 2010). It was also reported that high copy number of *DEFA1A3* was significantly associated with Crohn's Disease in Danish Patients (Jespersgaard et al., 2011). Recently GWAS studies also showed associations of SNPs at the *DEFA1A3* locus with IgA Nephropathy and Periodontitis.

1.3.1 IgA Nephropathy

IgA nephropathy (IgAN) was first described by Berger and Hinglais in 1968 (Berger and Hinglais, 1968). It is a common form of primary glomerulonephritis and the most important cause of end-stage renal disease (ESRD) worldwide (Sanders and Wyatt, 2008); analysis of mesangial IgA deposits in post-mortem kidneys suggest that the prevalence of glomerular IgA deposits in Finland may be as high as to 6.9% (Varis J, 1993) and in Japan is 16% (Suzuki et al., 2003). The

symptoms of IgAN are highly heterogeneous; the patient could show haematuria, proteinuria, or upper respiratory tract, gastrointestinal or urinary infection, but the most characteristic feature is deposits of IgA in the renal mesangium. The diagnosis of IgAN relies entirely on a histopathologic renal biopsy evaluation of IgA deposits, either alone or with IgG, IgM, or both in the glomerular mesangium (Wyatt and Julian, 2013, Francois et al., 2012). IgAN is a progressive disease with up to 50% of affected adults having persistent hematuria or proteinuria (D'Amico, 2000). Approximately 30%–50% of patients will develop end-stage renal disease within 20 years of the initial biopsy (Schena, 1990). A kidney transplant is the treatment of choice in individuals that develop ESRD, but the high recurrence rate after kidney transplantation suggested that the primary pathogenic defect of IgAN may be of extra-renal origin (Berger et al., 1975).

Current studies suggest that altered IgA1 O-glycosylation may be an extra-renal cause of IgAN. IgA is an antibody that plays an important role in mucosal immunity; IgA has two subclasses, IgA1 and IgA2. IgA1 is an unusual serum protein in possessing a series of O-linked glycans in the hinge region of the molecule, a feature which distinguishes it from IgA2 and other serum immunoglobulins (Kerr, 1990). Galactose deficient variants of IgA1 (Gd-IgA1) are more common in the sera of IgAN patients compared with those of healthy and disease controls (Moldoveanu et al., 2007). IgA deposition in the

renal mesangium is also a feature of Henoch–Schönlein purpura nephritis, but altered IgA1 O-glycosylation is only found in those patients associated with glomerulonephritis (Allen et al., 1998). Some theories have been developed to explain the auto-immune nature of IgAN and the molecular mechanism behind it. The hinge region of IgA1 consists of two octapeptide repeats (Frangione and Wolfenstein-Todel, 1972) with nine potential *O*-glycosylation sites of which three to six are glycosylated (Takahashi et al., 2012). In patients with IgAN, the hinge regions of the O-glycans lack galactose, leaving the terminal GalNAc residue recognizable by other IgG or IgA1 autoantibodies, and these IgA1 or IgG molecules interact with Gd-IgA1 and eventually form circulating IgA immune complexes which deposit in renal mesangium (Tomana et al., 1999). This deposition will lead to activation of the mesangial cells with release of cytokines and chemokines that can cause glomerular injury (Barrat and Feehally, 2011). However, the source and origin of the Gd-IgA1 are still not clear. The mucosal immunity was thought to be involved in IgAN since the first clinical report (Coppo et al., 2000). IgAN patients frequently show acute tonsillitis inflammation leading to a hypothesis that the Gd-IgA1 is produced in tonsils. More interestingly in East Asia, tonsillectomy has recently been used to modulate mucosal-associated lymphoid tissue activity (Chen et al., 2007, Hotta et al., 2001, Maeda et al., 2012). This is a simple method to lower the

mucosal immunity level, but the result is still inconclusive (Coppo, 2015). Some research shows that the Gd-IgA1 may originate from mucosally primed plasma cells, and that the cross talk between the mucosa and bone marrow may be critical in the pathogenesis of IgAN (Barratt et al., 2007). Most recently, some researchers were interested in intestinal immunity and the permeability of the intestinal barrier. Some interesting data suggest a new hypothesis of a strong intestine–kidney connection in IgAN: defective immune tolerance might favour an abnormal response to microbiota with alterations of the intestinal barrier, activating mucosal-associated lymphoid tissue and intestinal inflammation, leading to increased production of aberrantly glycosylated polymeric IgA1 and eventually resulting in renal mesangial deposits (Coppo, 2015). Study of IgA nephropathy-prone mice shows that injected IgA starts to deposit in glomerular capillaries between 1 min to 2 hrs and the deposit disappeared after 24 hours. This finding indicates that a continuous clearance mechanism of IgA deposition may be at work in the glomerulus, and thus IgAN may explain by an imbalance between IgA deposit and clearance (Kenji et al., 2014).

1.3.2 Genetics of IgA nephropathy

Although the aetiology behind IgAN is still not known, a number of lines of evidence suggest the disease is associated with strong

genetic factors. Immunohistologic findings suggest IgA nephropathy has a striking geographic distribution: the disease is more frequent in the East and Southeast Asian area than in Australia, Europe, and North America, and relevantly rare among African countries. Study of United States ethnicities of different origin still shows this striking distribution; the percentage of IgAN-ESRD was 5-fold greater for "Caucasian" (European Whites) and 15-fold greater for Asian Americans compared to African-Americans (Kiryluk et al., 2012). This suggests the disease is unlikely to be caused by the differences in environmental factors. The male/female ratio shows that males are three times more likely to be affected by the disease, particularly in the second and third decade of life (Schena, 1990). Although most IgAN cases are sporadic, still up to 14% of cases may be familial (Kiryluk et al., 2012). Some familial aggregation of the disease has been studied, and it suggested that IgAN follows an autosomal dominant transmission with incomplete penetrance (Kiryluk and Novak, 2014). Since 2005 genome wide association study (GWAS) has become a very popular tool for locating disease causing variation. Recently several large scale GWAS for IgA nephropathy have been carried out and identified several susceptibility loci. In 2010 the first GWAS study of IgAN involved affected UK individuals compared with controls from a British Birth Cohort; this study showed a strong signal of association on chromosome 6p in the region of the MHC,

suggesting that the HLA region contains the strongest common susceptibility alleles that predispose to IgAN in the European population (Feehally et al., 2010). In 2011 a study involving 1,194 cases and 902 controls of Chinese Han population identified three independent loci in the major histocompatibility complex, as well as a common deletion of *CFHR1* and *CFHR3* at chromosome 1q32 and a locus at chromosome 22q12 (Gharavi et al., 2011). In 2012 another GWAS study of IgAN in Han Chinese population with 1,434 affected individuals and 4,270 controls identified associations at 17p13 (*TNFSF13*) and 8p23 (*DEFA*) as susceptibility genes (Yu et al., 2012). These findings show that IgAN is associated with variants near genes involved in the anti-glycan immune response, mucosal immunity, antigen processing and presentation, and host–intestinal pathogen interactions. In 2014, so far the largest scale of GWAS of IgAN involving 20,612 individuals of European and East Asian ancestry IgAN patients was carried out, discovering two known loci (*HLA-DQB1* and *DEFA*) and four new signals in *ITGAM-ITGAX*, *VAV3* and *CARD9* and two new independent signals at *HLA-DQB1* and *DEFA* locus (Kiryluk et al., 2014). A genetic risk score was calculated based on the weighted sum of the number of risk alleles each IgAN patient carried. This IgAN risk score was highly significantly associated with the age of diagnosis. Most of the loci discovered so far are involved in the maintenance of the intestinal barrier and mucosal immunity.

These GWAS findings support the intestine-renal hypothesis of IgAN pathogenesis.

The 2012 IgAN GWAS result detected a protective signal in rs2738048 G ($p=3.18\times10^{-14}$ OR=0.79) near the alpha defensin locus in the Han Chinese population (Yu et al., 2012). Later on, in 2014, GWAS results confirmed this protective signal and further verified that it was an Asian population specific signal. This study also detected a new signal in rs10086568 near *DEFA5* (OR= 1.16, p= 1.0×10^{-9}) (Kiryluk et al., 2014). The association of rs2738048 with IgA nephropathy is not the only GWAS SNP found near *DEFA1A3* CNV. There are seven SNPs found within *DEFA* genes that were significantly associated with IgAN by a association study with 1,000 IgAN cases and 1,000 controls (Xu et al., 2014). Recently research has started to focus on *DEFA1A3* CNV internal variants; Ai Z, et.al 2016 studied association between several CNV internal variants and IgA nephropathy, and pointed out that low α-defensin gene copy number increases the risk for IgA nephropathy. A noncoding deletion variant (211bp) ($p = 3.50 \times 10^{-16}$; OR=0.75) is the strongest association discovered so far (Ai et al., 2016). This evidence suggests association on the flanking regions are likely to act as an indirect reporter of CNV internal variants and they were present on some common haplotypes.

1.3.3 Periodontitis

Another recent reported associated disease with *DEFA1A3* CNV is periodontitis. Periodontal diseases are complex chronic diseases characterized by inflammation of tissues surrounding the teeth and progressive loss of the alveolar bone. They are the most frequent cause of tooth loss in adults (Lipsky et al., 2017). Periodontal diseases are caused by interaction between oral microbiota that presents on the surface of teeth and an over-aggressive host immune response. The diagnosis of periodontitis relies on inspection of soft gum tissues around teeth, and X-ray films examining the level of alveolar bone loss around the teeth (Savage et al., 2009). The disease can be classified into 7 classes: Gingivitis, Chronic periodontitis, Aggressive periodontitis, Periodontitis as a manifestation of systemic disease, Necrotizing ulcerative gingivitis/periodontitis, Abscesses of the periodontium, and Combined periodontic-endodontic lesions (Armitage, 1999).

Unlike dental caries which is a common modern human disease, and which was hypothesized to reach high prevalence with the high sugar diet change after the first agricultural revolution, periodontal diseases can occur naturally or experimentally induced in various animals, including primates, miniature pigs, dogs, rodent (Helieh S and David A, 2011). In humans, periodontitis has high prevalence affecting

individuals regardless of population of origin. It is the second most prevalent dental disease after dental decay. It is estimated that there are 750 million individuals (10.8% of the total population) affected by chronic periodontitis in 2010 (Vos et al., 2012). Poor oral hygiene is thought to be the main cause of periodontitis, which leads to accumulation of mycotic and bacterial plaque around the teeth. Once the plaque is formed, the composition of biofilm changed from streptococci to actinomyces which will cause gingivitis. If a predominantly gram-negative environment is established, bacteria such as *Porphyromonas gingivalis, Prevotella intermedia*, and *Fusobacterium nucleatum* will settle down and lead to prolonged inflammation (Tsuchida et al., 2017) Other factors such as nutrition, smoking and, diabetes also influence the risk of periodontitis (Borrell and Papapanou, 2005).

1.3.4 Genetics of periodontitis

Recent studies suggested genetic factors play an important role in the pathogenesis of periodontitis. Twin analyses have demonstrated that there is 50% heritability for mean attachment loss in subjects with chronic periodontitis (Michalowicz et al., 2000). Since 2005 GWAS have been applied in understanding the genetic factors of periodontitis. There are several GWAS studies of the chronic form of periodontitis (CP) published so far (Divaris et al., 2013, Feng et al.,

2014, Teumer et al., 2013, Shimizu et al., 2015), but the outcome is inconclusive, as most studies failed to detect variants passing a genome-wide significance threshold. This is partially due to the most prevalent chronic form of periodontitis being influenced by various environmental factors such as long-time smoking and diabetes and also because these association studies are small number and limited by statistical power. Aggressive periodontitis (AgP) is a rare form of periodontitis with severe periodontal attachment loss and severe destruction of the alveolar bone in young adults (age < 35 years). It is a good study material for genetic investigations since the young age of disease onset which means patients are less affected by environmental factors. Several AgP susceptibility GWAS loci have been identified so far, including *GLT6D1* (Schaefer et al., 2010), *IRF5* and *PRDM1* (Schaefer et al., 2014). In 2017 a GWAS study of AgP patients identified that *SIGLEC5* and *DEFA1A3* are associated loci in German and Dutch case-control samples, P=1.09E-08 (rs4284742,-G; OR=1.34, 95% CI=1.21-1.48) and P=5.48E-10 (rs2738058,-T; OR=1.28, 95% CI=1.18-1.38) respectively. These associations were validated in German samples of chronic periodontitis and replicated in Turkish aggressive periodontitis cohorts (Munz et al., 2017).

1.3.5 Potential link between IgA nephropathy and periodontitis

Both IgA nephropathy and periodontitis GWAS studies reported *DEFA1A3* as a susceptibility locus. The associations with the two diseases can be linked together through the same variant rs2738058. As early as in 2013 GWAS of chronic periodontitis in a general German population reported an association with this SNP (Teumer et al., 2013), although in this study none of the SNP markers reached genome wide significance levels, but it reported rs2738058 (T; P=6 $x10^{-6}$ (Age 20-81 years) OR=1.29) among the list of best-associated SNPs. The 2017 GWAS of AgP patients also reported rs2738058 (T; P = 5.48 $x10^{-10}$ OR = 1.28, 95% CI = 1.18–1.38) (Munz et al., 2017) as an associated variant. Moreover, a Chinese patient cohort study discussed the association between IgA nephropathy and periodontitis at a clinical level. They found the prevalence of severe chronic and aggressive periodontitis was higher in patients with IgA nephropathy. Chronic periodontitis is correlated with the onset and development of IgAN. Patients with IgAN have a worse condition with the aggravation of periodontitis (Cao et al., 2016). These findings show a potential link between IgA nephropathy and periodontitis through the *DEFA1A3* locus.

1.4 Neanderthal Introgression

The evolutionary history of the human *DEFA1A3* CNV is another topic to study in this thesis. It can be enlightened through archaic genomic sequencing data.

1.4.1 Introduction to Neanderthals

Neanderthal is classified as a species or subspecies of archaic humans under the genus *Homo*. The first Neanderthal fossil record was discovered in 1829 at Engis caves. Neanderthals diverged from modern humans at least 430,000 years ago based on the fossil record (Arsuaga et al., 2014), or 550,000-765,000 years ago based on a study of genomic data (Prufer et al., 2014). Neanderthals migrated to Eurasia and geographically separated from ancestors of *Homo sapiens* in Africa. Both species started their own path of evolution until modern humans the out of Africa migration about 70,000 years ago (Soares et al., 2012). Neanderthal fossil records have been discovered in a large area spanning from Western Europe to western and northern Asia, but they had a long-term effective population size smaller than that of modern humans according to Neanderthal mitochondria DNA studies (Green et al., 2008). The last Neanderthals remaining were discovered in Iberia, dated to 28,000 years ago. After that, no more Neanderthal remains were found. The reason for

extinction is unclear. Climate changes, modern human competition, a decrease in population size and accumulation of genetic deleterious effects may be the cause of extinction (Sanchez-Quinto and Lalueza-Fox, 2015).

1.4.2 Admixture with modern humans

Neanderthals went extinct at about 40,000 years ago (Higham et al., 2014). Modern humans left Africa, starting the migration to Euro-Asia about 70,000 years ago, and arrived in Europe between 45,000 and 43,000 years ago (Higham et al., 2014). So the two species possible had a long period of co-existence in Eurasia. In 2014, a high coverage (~50X) whole genome sequencing was performed for a Neanderthal sample discovered in the Altai mountains (Prufer et al., 2014). This high-quality whole genome sequencing data allowed the investigation of Neanderthal genetics at high resolution for the first time. In the past few years, several studies have concluded that modern humans had interbred with Neanderthals after the out of Africa migration and identified Neanderthal introgressed genomic regions in the human genome (Sankararaman et al., 2014, Vernot and Akey, 2014, Sankararaman et al., 2016, Vernot et al., 2016). It is estimated that today modern non-Africans have about 1.5-2.1% of genetic material passed from Neanderthals (Green et al., 2010). This

admixture is estimated to date to 47,000–65,000 years ago (Sankararaman et al., 2012).

Neanderthal introgression can be detected by several features. Firstly, Neanderthals are distributed in Europe and Asian, so far no evidence of Neanderthal interbreed with Africans found and no Neanderthal remaining was discovered in Africa. So Neanderthal introgressed haplotypes should at least absent from sub-Sahara African populations. Secondly, Patterson's D is a statistic that measures the excess sharing of derived alleles between each of two populations in a pair (ingroup populations) and an outgroup population (Racimo et al., 2015). If neither of the modern human populations has had any gene flow from Neanderthals, then each of the two human populations should share approximately the same number of derived alleles with the Neanderthal. Population share more alleles with Neanderthals indicate introgression may happen. Thirdly, by sequence divergence. If we know a local mutation rate we can estimate the divergence time between putative introgressed human DNA segments with corresponding Neanderthal DNA segment. An introgressed haplotype should have low sequence divergence to the Neanderthals, but high sequence divergence to other present-day human individuals (Racimo et al., 2015). This method can also distinguish introgression from incomplete lineage sorting. Finally, track length and linkage disequilibrium (LD). Time of Neanderthal

admixture with modern human is estimated to 47,000–65,000 years. The length of Neanderthal coting in human should keep in a certain length. The existence of a long introgressed haplotype should increase long-range linkage disequilibrium (LD) (Racimo et al., 2015), statistics like S* was designed to grab such information and detect introgressed segments.

1.4.3 Adaptive introgression

Introgression of a gene flow from one species into the gene pool of another which help this species to survive and enhance adaptation in new environment is referred to adaptive introgression. Functional analysis of Neanderthal introgressed non-synonymous and regulatory variants shows introgressed alleles contribute proportionally more to expression variation than non-introgressed alleles, suggesting that the major influence of Neandertal introgressed alleles is through their effects on gene regulation (Dannemann et al., 2017). Several studies pointed out that on average Neanderthal introgressions were deleterious, and suggested there was negative selection against Neanderthal introgression in modern humans. Neanderthal introgression was found to be largely depleted from regions near genes, which were thought to be explained by purifying selection (Vernot and Akey, 2014, Sankararaman et al., 2014). Although it seems Neanderthal introgressions caused lower fitness for early

modern humans, there are still several Neanderthal introgressions reported to be beneficial (Racimo et al., 2015). Since Neanderthals inhabited Eurasia for more than 400,000 years, that time was enough for evolutionary adaptation to local environments. Such adaptive introgression may also have helped modern humans to adapt to new environments in Eurasia. These examples of adaptive introgression involved genes important in the immune system, metabolism, pigmentation and integumentary system, among others (Mendez et al., 2012, Khrameeva et al., 2014, Racimo et al., 2015, Sankararaman et al., 2014, Ding et al., 2014). Most strikingly, immune system genes like HLA region, *STAT2*, *OAS* gene clusters have been reported with Neanderthal introgressed haplotypes.

1.5 DNA Sequencing

With the fast development of sequencing technologies there are more and more tools available to study *DEFA1A3* haplotype diversity. Developing methods to reconstruct *DEFA1A3* haplotype structure is also another task in this thesis.

1.5.1 First generation

The world's first DNA sequencing method was developed by Frederick Sanger and his colleagues in 1975. It was based on the selective

incorporation of chain-terminating dideoxynucleotides by DNA polymerase during *in vitro* DNA replication (Sanger et al., 1977, Sanger and Coulson, 1975), and is called the "chain-termination method" or "Sanger sequencing method". In 1977 Maxam and Gilbert also developed their method of DNA sequencing. This method is based on nucleobase-specific partial chemical modification of DNA and subsequent cleavage of the DNA backbone at sites adjacent to the modified nucleotides (Maxam and Gilbert, 1977); it is also called as the "chemical sequencing method." These methods nowadays are referred to as first generation sequencing (FGS). The chain-termination method is less complex and more reliable than "chemical sequencing method" and eventually became the most widely used FGS method. It was used in many famous projects such as the "Human Genome Project". The high accuracy and reliable features of Sanger sequencing make it the most basic and fundamental DNA sequencing method until today.

1.5.2 Second generation

Although Sanger sequencing has the advantage of high accuracy, long read length and reliability, its low throughput and high-cost features limit its power towards genomic level applications. The new wave of revolution came in 2005 when the 454 pyrosequencing system was first introduced by 454 Life Sciences. This was followed

by the Illumina (Solexa) system in 2007 and ABI's SOLiD system in the same year. These new methods are collectively called "Next-Generation Sequencing (NGS)" or "second generation sequencing (SGS). All NGS platforms perform sequencing of millions of small DNA fragments in parallel. This greatly increased sequencing throughput and reduced sequencing time and cost leads to it being also referred to as "high-throughput sequencing". For example, the time required for sequencing a complete human genome was reduced from years by FGS to days by SGS. The major difficulties of SGS come from post-sequencing data processes. The length of small fragments generated by SGS range from 35 to 400bp, and they need to be pieced together by some bioinformatics approaches, either mapping to a known reference genome or by *de novo* assembly. In SGS, Each nucleotide could be sequenced multiple times, and this redundant coverage brings two benefits. Firstly, this provides a high-quality assembly. Secondly, redundant reads will increase sequencing accuracy of each nucleotide. If a nucleotide has a single read error rate at 1%, then 8X coverage will reduce the error rate to 10^{-6} (Micheal et al., 2010). The assembly of SGS short reads is a major challenge, because each individual's genome contains single nucleotide variations, structural variants, and highly repeated regions. If a repetitive sequence is longer than the length of a read, then coverage will never compensate. Paired reads were then

introduced to overcome this drawback, which consists of two reads generated from a single fragment of DNA and separated by a known distance. Paired-end sequencing approaches greatly improve mapping quality, but repetitive regions longer than a paired-end read remain problematic. *DEFA1A3* CNV is a good example; each repeat unit is 19kb in length and the human reference genome (hg19) has 3 such copies, so that SGS short reads inside the CNV region are mapped to multiple locations. Nevertheless, SGS technologies have dominated the sequencing market for the past decade. This advance in sequencing technologies makes human variation projects like the 1000 Genomes Project possible.

1.5.3 Third generation

Despite the successful applications of SGS, its short sequencing length and complex mapping process create a barrier to studying highly repetitive genomic regions such as CNV. So-called "third generation sequencing" (TGS) technologies appeared recently, first by Pacific Biosciences (PacBio) and followed by Oxford Nanopore Technologies in 2014. TGS involves sequencing single DNA molecules in real time and bring three important improvements compared to SGS. Firstly, the length of reads increased from tens of bases to tens of thousands of bases. Secondly, the time required by sequencing a genome reduced from days to hours. Thirdly, a PCR enrichment step

used in SGS is not necessary, so sequencing bias generated by PCR was eliminated (Hengyun et al., 2016).

PacBio SMRT sequencing

In PacBio SMRT sequencing, the four types of nucleotides are labeled by different fluorescent dyes. Each template DNA molecule is sequenced by synthesis by using DNA polymerase in a zero-mode waveguide (ZMW). A ZMW is a small structure that allows observation of only a single nucleotide of DNA being incorporated by DNA polymerase (Levene et al., 2003). The DNA polymerase is attached at the bottom of a ZMW. When DNA polymerase incorporates a nucleotide, the fluorescent dye is cleaved off and fluorescent signal can be detected by a detector. "PacBio RS II" uses 150,000 ZMW holes working in parallel allow a high throughput sequencing, and the “Sequel system” further increased ZMW holes to one million.

Oxford Nanopore sequencing

Oxford Nanopore Technologies released the MinION device in 2014. It is the smallest DNA sequencing machine released so far with a size comparable to a flash memory stick. The device includes a removable flowcell which contains a Nanopore membrane with 512 channels, allowing 512 DNA molecules to be sequenced at the same time. Each channel has 4 wells, giving in total 2048 wells. It connects to a

computer through a USB portal. During sequencing, a voltage is applied to drive DNA through the Nanopore and ion current flow can be measured. The current in the Nanopore is measured by a sensor several thousand times per second. A current change is recorded and characterized when a DNA molecule passes the Nanopore, and this signal is then analyzed to generate base calling (Hengyun et al., 2016).

1.6 Aims of this research

The *DEFA1A3* locus was reported to be associated with IgA nephropathy and periodontitis (Matthias Munz, 2017, Yu et al., 2012, Kiryluk et al., 2014), but the underlying causative factors were not identified. Variants that were not studied in association studies, or the copy number itself, might be the real source of association. They may be linked to reported variants through a common haplotype. Haplotype diversity other than in European populations is also not well characterized (Black et al., 2014). Studying the global population haplotype diversity may reveal the evolutionary history of this locus. So it is important to define *DEFA1A3* haplotype diversity in human populations and discover haplotypes associated with IgA nephropathy and periodontitis.

It is reported that variants inside the *DEFA1A3* CNV have a stronger association with IgA nephropathy than flanking variants (Ai et al., 2016), but *DEFA1A3* CNV internal variants are largely unknown. It is also necessary to develop a method for detecting variations inside *DEFA1A3* CNV and looking for potential disease related functional variants.

The Oxford Nanopore has released their third generation sequencing device MinION. In principle, the revolutionary long sequence read capability should allow us to reconstruct *DEFA1A3* CNV haplotype structures. It is interesting to explore what change the new technology could bring to the CNV studies.

2 Chapter2: Materials and methods

2.1 Research Materials

2.1.1 The 1000 Genomes Project data

The 1000 Genomes Project data were used to understand *DEFA1A3* locus diversity and evolutionary history. This dataset contains whole genome sequencing of 2,504 individuals from 26 populations. We downloaded sequencing data from 1000 Genomes Project phase1 9 populations, 1063 individuals from the 1000 Genomes FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/data/). These are aligned bam files against the human hg19 reference genome. Then we extracted the 400kb (from chr8:6,700,000-7,100,000) *DEFA1A3* CNV region from the genome assembly by the `samtools view` command. These data were used for realignment of *DEFA1A3* CNV region.

Phased VCF files were used for flanking haplotype diversity analysis, downloaded from (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chr8.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz). They are VCF files phased by shapeit2 software. These VCF files

were then converted to ped format by VCF_to_PED_converter.pl obtained from the 1000 Genomes Project website.

2.1.2 Archaic human sequences

Altai Neanderthal (Prufer et al., 2014), Denisovan (Meyer et al., 2012) and Ust-ishim (Fu et al., 2014) whole genome sequencing data were downloaded from the Max Planck Institute for Evolutionary Anthropology (http://www.eva.mpg.de/neandertal/index.html). All genomes were mapped against the human hg19 reference genome. We cut a 300kb section which contains the *DEFA1A3* region from whole genome sequencing bam files (chr8: 6700000-7000000). Although these archaic human samples are un-phased, all 3 samples appear to carry homozygous haplotypes in this region and this allows us to direct phase by hand. Phased VCF files were converted to ped format then to fasta format by a self-made R script (vcf2ped.r).

2.1.3 HapMap project samples

The HapMap phase 1 genomic DNA samples were used in our experiments. These included 180 samples from Utah, USA with northern and western European ancestry (CEU), 90 samples from Ibadan, Nigeria (YRI), 45 samples from Tokyo, Japan (JPT) and 45 samples from Beijing, China (CHB). The CEU samples contain 56 trios, 5 duos, and 2 singletons. YRI samples consist of 30 trios. JPT

and CHB samples are unrelated individuals. All samples were purchased from Coriell (http://ccr.coriell.org) and stored in -20°C freezer.

2.2 Bioinformatics methods

2.2.1 Realignment of 1000 genome project samples

A human individual could have variable *DEFA1A3* copy number but the human hg19 reference genome only has 3 copies. The 1000 Genome Project NGS short reads were randomly mapped to three copies, making the *DEFA1A3* CNV a mapping quality "0" region and it is impossible to utilize these low mapping quality reads. In order to detect variants inside the *DEFA1A3* CNV, we realigned NGS short reads to an artificial one-copy reference sequence, and then called variants according to their read depth copy number.

The data processing flow chart is shown in fig 2.1. 1024 1000 Genomes Project samples were used in this analysis. Low mapping quality reads mapped to *DEFA1A3* CNV between chr8: 6824739-6882259 (hg19) were cut from the genome assembly by the samtools -view command. Then we converted each bam file back to two fastq files by the Picard tool and removed unpaired reads by a script "fastqCombinePairedEnd.py". A 22kb long artificial one-copy *DEFA1*

reference was created in Fasta format. This reference sequence contains one 19kb long *DEFA1* full repeat plus a 3kb long *DEFA3* copy centromeric unique sequence obtained from the UCSC genome browser (hg19). Using this sequence as a reference, we then performed realignment by bowtie2 and converted output Sam format to Bam format. Artificial read groups were added by the Picard tool. Finally, we sorted and indexed bam files with samtools.

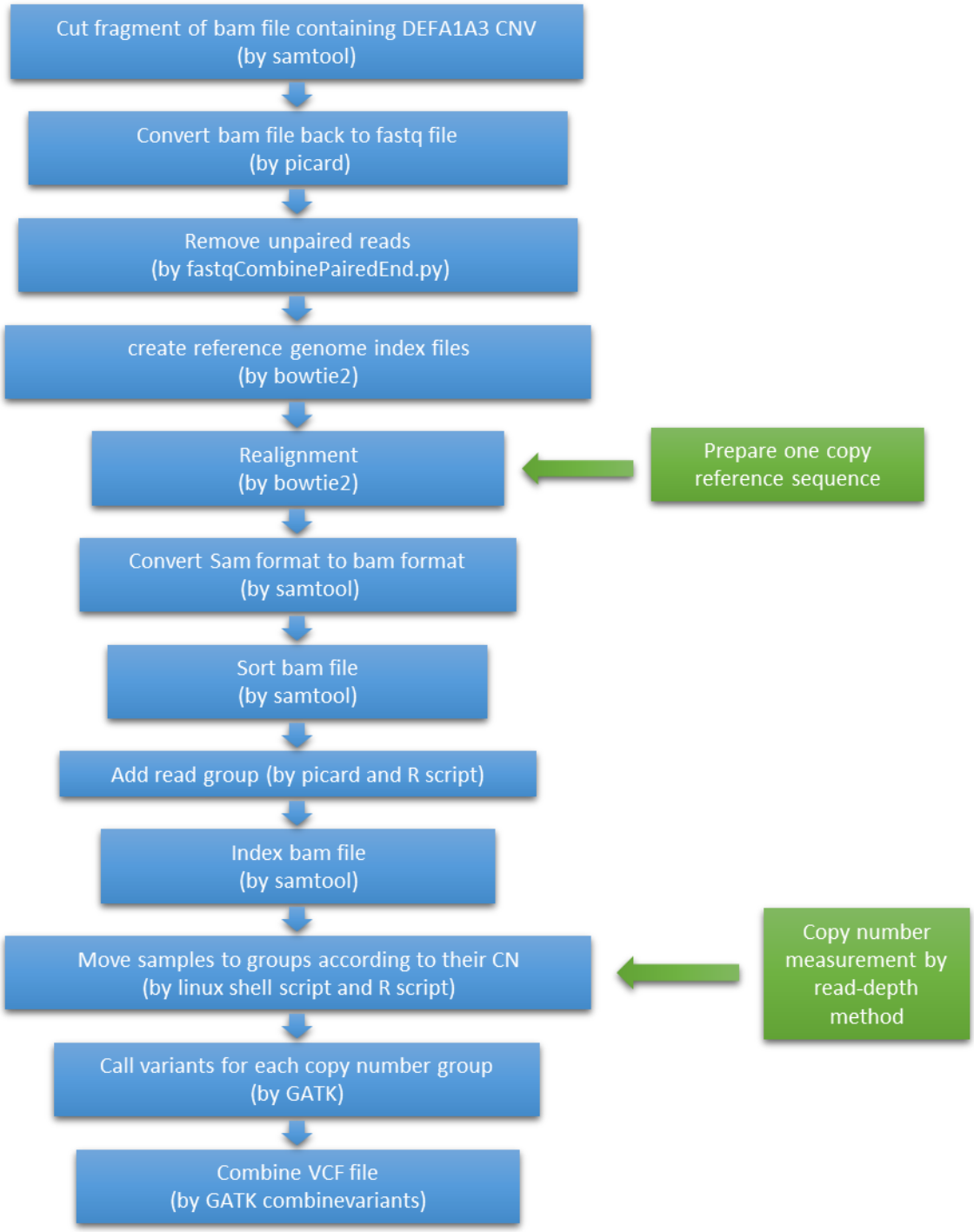


Figure 2.1 flow chart of data processing steps used in the realignment of 1000 Genome Project data and variant calling.

2.2.2 CNV internal variation detection

After realignment of *DEFA1A3* CNV internal short reads, sequences from all copies inside the CNV were collapsed down into one copy reference sequences. This multiple copy alignment is then treated like polyploid sequencing data. For achieving a high sensitivity and specificity for variation calling, we decided to group samples together according to their measured read-depth copy number and call variation separately for each copy number group (fig 2.1). *DEFA1A3* copy number for each sample is measured by a read depth method described in section 2.2.9. Samples were separated to 14 (3-16 copies) folders according to read depth copy number by a home made an R script called "Call_snp_by_CN.r". Variant calling was performed by Genome Analysis Toolkit (GATK) UnifiedGenotyper (according to GATK's standard setting) for each copy number group separately, in each GATK run for different copy number, ploidy were set to their measured total copy number. Both SNPs and indels were called and outputted in VCF 4.0 format. Then VCF files for each copy number group were combined into a single VCF file by GATK "combinevariants" command.

2.2.3 *DEFA1A3* flanking haplotype diversity analysis

To understand *DEFA1A3* flanking haplotype diversity and evolutionary history, we carried out a sequence diversity analysis. 300kb

(chr8:6700000-7000000) 1000 Genomes Project phase1 whole genome sequencing variants calling data were used in this analysis. The region included in this analysis is in an LD block as reported by a previous study, SNPs in 60kb region of the *DEFA1A3* locus displaying high levels of linkage disequilibrium (D' = 1, data from phased HapMap CEU1 individuals) (Black et al., 2014). Also evidence from the recombination hotspot analysis (section 2.2.4). Telomeric and centromeric sequences were analyzed separately since SNP correlation between the two flanking regions is not perfect. Phased VCF files were downloaded from 1000 Genome Project including CHB, CHS, JPT, CEU, GBR, FIN, LWK, ASW, YRI 9 population (in total 1738 phased sequences) and converted to ped file by Perl script VCF_to_PED_converter.pl. Then ped files were converted to fasta format by two home-made R scripts (split.r and ped2fasta.r). SNPs inside the CNV region (from chr8:6829137 to 6879803) and rare SNPs with minor allele frequency <1% were removed from the analysis. Data in the telomeric region (chr8:6822524 to chr8:6829137) and centromeric regions (chr8:6879803 to 6899209) were divided into two separate files.

Haplotype classification was carried out by homemade R script, "Haplotype.R". This script detected unique sequences in the population and counted their frequency. Highly similar haplotypes were grouped together if the differences at SNP positions were

smaller than 1%. This was achieved by collapsing down branch lengths of 0.01 or less in a neighbour-joining tree constructed in Mega 6 software. Haplotypes were named in a descending order from highest frequency haplotype. Haplotypes with a frequency smaller than 1% were marked as "rare" haplotype. Haplotypes were named separately for the telomeric and centromeric regions.

We used a heat map to display *DEFA1A3* common centromeric and telomeric haplotype correlation (frequency >1%). This is achieved by R script "heat_map.R". It links common telomeric and common centromeric haplotypes by a two-dimensional matrix, colour representing the frequency of each combined haplotype. The R package "lattice" was used as the plot function.

2.2.4 Recombination analysis

Population scaled recombination rates (rho) plot was drawn by RDP 4 software (Martin et al., 2015) for viewing local recombination hotspots and determining regions for sequence analysis. This analysis included a 300kb (chr8:6700000-7000000) *DEFA1A3* flanking region, in 1738 sequences from 1000 Genomes Project phase 1, drawn from 9 populations.

Evidence for recombinant centromeric and telomeric haplotypes was obtained by a visual check of sequence similarities in recombination

analysis tool (RAT) software (Etherington et al., 2005). Only haplotypes with frequency >1% were included, and this contains 14 major telomeric and 17 major centromeric haplotype sequences in Fasta format. High-frequency haplotypes were considered as parental haplotypes. A haplotype with segmental high similarity to more than one parental haplotype was marked as a recombinant haplotype.

2.2.5 Linkage disequilibrium calculation

The Linkage disequilibrium D' value was calculated by the equation D=(p11*p22)-(p12*p21), D_{max}=min(p1*q2,p2*q1), D'=D/Dmax and r^2 value by r^2=D*D/(p1*p2*q1*q2) and significance was tested by chi-square $\chi^2=r^2*n$, with one degree of freedom and significance threshold $p<0.05$.

2.2.6 Flanking haplotype tag SNP search

DEFA1A3 flanking haplotype tag SNPs were identified as the SNPs best associated with one flanking haplotype, but not (or less frequently) found in other haplotypes. *DEFA1A3* CNV flanking SNP and haplotype association was measured by R-squared value. Here we calculated the R-squared value between each SNP and each haplotype by a home-made R script "R_square.R". For each SNP and each haplotype association, a 2X2 contingency table was generated (Table 2.1).

Table 2.1 2X2 contingency table used in variants haplotype association analysis

	Allele in the test haplotype	Allele in the non-test haplotype
SNP allele 1	A	B
SNP allele 2	C	D

SNPs with more than 2 alternative alleles were removed from this analysis. Rare haplotype classes with a frequency smaller than 1% were also removed. R-squared is calculated as:

$$\frac{((\mathrm{A}/\mathrm{tot}) * (\mathrm{D}/\mathrm{tot}) - (\mathrm{C}/\mathrm{tot}) * (\mathrm{B}/\mathrm{tot}))2}{((\mathrm{A}+\mathrm{C})/\mathrm{tot}) * ((\mathrm{B}+\mathrm{D})/\mathrm{tot}) * ((\mathrm{A}+\mathrm{B})/\mathrm{tot}) * ((\mathrm{C}+\mathrm{D})/\mathrm{tot})}$$

Where "tot" is the total number of samples included in this analysis (= A + B + C + D); A, B, C, and D are values from the 2 by 2 table in Table 2.1. Finally, we obtained a list of associations between each flanking variant and each haplotype class.

2.2.7 CNV internal haplotype tag multi-site variants search

Variants present in a multiple copy variable region and also referred as Multi-Site Variants (MSV). In this analysis, a CNV internal tag MSV of a haplotype class should be present on every repeat unit of this haplotype but not present on any copies of other haplotype classes. To search for such tag variants we use samples carrying homozygous telomeric and centromeric haplotypes. 38 samples from 1000

Genome Project were eligible for analysis. We searched for tag MSVs for each haplotype class by calculating the R-squared value between each MSV and flanking haplotype classes. Two R scripts were made for this purpose (Step-1_random_phase.R, Step-2_R_squared.R). First, we randomly phased each sample according to their read-depth copy number and labelled them by their flanking haplotype classes. Then we pooled copies from each sample together and calculated the R-squared value between each MSV and each haplotype classes. We selected MSVs with the highest R-squared value as tag MSVs.

2.2.8 CNV internal variants and flanking haplotype association

To detect *DEFA1A3* CNV internal variants associated (but not necessary present on every copies, as in section 2.2.7) with specific flanking haplotypes, we carried out a Cochran-Armitage trend test for all CNV internal variants with all flanking haplotype classes. This test is appropriate only when one variable has two levels and the other variable is ordinal, as a modification of the Pearson chi-squared test. The Cochran-Armitage trend test can be applied in our case since variant calling inside the CNV region is un-phased and each sample carries two haplotypes. 1738 sequences were used in this analysis, obtained from the 1000 Genomes Project.

The calculation was achieved by a self-made R script called “Cochran-Armitage.r”. This script takes two input files: a VCF file of CNV internal variants for all samples, and a CSV table containing the two flanking haplotype identities for each sample. First, we converted variant calling into presence (marked as “1”) or absence (marked as “0”) format. If a variant is detected in a sample we code it as “1”, no matter how many copies carry this variant. Then we use a loop to test variants associated with each haplotype class. Each sample’s haplotype information was represented in three categories “++”,”+-”,”--” corresponding to sample carrying homozygous test haplotype, heterozygous test haplotypes, and non-carrier respectively. Then the program will generate a 2 by 3 contingency table (Table 2.2). Finally, Cochran-Armitage test was carried out by “CochranArmitageTest” command from an R package called "DescTools". Bonferroni correction was then applied to output p values.

Table 2.2 Scheme of 2 by 3 contingency table used in Cochran-Armitage trend test

	-- (non-test haplotypes carrier)	+- (heterozygous test haplotypes)	++ (homozygous test haplotypes)
1(present)	A	B	C
0(not present)	D	E	F

2.2.9 Measuring copy number by next generation sequencing read depth

In the 1000 Genomes Project NGS data, the copy number of *DEFA1A3* CNV can be measured according to rise and drop in sequencing read depth. The higher copy numbers usually have a higher read depth compared to low copy number regions. This method was first developed by Prof John Armour (unpublished). It selected 3 regions inside *DEFA1A3* CNV named core, core2, core3 and compared them with a non-copy number variable region outside the CNV with a known copy number of 2 (Table 2.3). All regions have similar GC content so that the influence of PCR bias is minimized. The human hg19 reference genome has three copies of *DEFA1A3*, so reads within these three regions should be added together. The total number of reads in the defined region was counted by a Perl script called "JAscript.pl". A previous analysis concluded that core3 and flankR ratio divided by 1.06 gives the most realistic *DEFA1A3* CNV copy number estimation when compare to PRT method of copy number measurement (Unpublished data). Then all samples from 1000 Genomes Project and three archaic samples *DEFA1A3* copy number were measured by this method.

Table 2.3 Summary of regions used in copy number measurement for 1000 Genomes Project samples.

	GC content	mean length	region 1 position	region 2 position	region 3 position
Core	44.30%	8258 bp	6829298-6837591	6848458-6856701	6867561-6875800
core2	43.70%	5485 bp	6832100-6837591	6851217-6856701	6870320-6875800
core3	45.70%	8140 bp	6832100-6840247	6851217-6859357	6870320-6878452
flankL.bed	45.60%	130000 bp	6700000-6830000		
flankR.bed	44.20%	100000 bp	6900000-7000000		

2.2.10 Flanking haplotype phylogenetic analysis

We constructed median joining networks to study the phylogenetic relationship between *DEFA1A3* flanking haplotypes. Genomic regions included in this analysis are same to haplotype classes analysis (~6kb telomeric and ~20kb centromeric). We removed haplotypes with strong evidence of recombination (section 2.2.4). Neanderthal (Nea), Denisovan (Den) and Ust-ishim (UST) haplotypes were added to analysis for understanding their phylogenetic relationships. Flanking haplotype sequences were prepared in fasta format. Then we converted fasta format to rdf format by DNAsp software (Librado and Rozas, 2009). Then we loaded the rdf file into network 5.0.0.1 software (Bandelt et al., 1999) for construction of a median-joining network. Chimpanzee haplotype (Pantro4 reference) was used as a root. Default setting were used: All site weights= 10, Transitions:

transversions weight=1:1, epsilon = 0. (Epsilon 0,10,20,30 were tried, and no major change in topology found.)

2.2.11 Flanking haplotypes and disease associations

GWAS studies reported *DEFA1A3* CNV flanking SNPs associated with both IgA nephropathy and periodontitis. To understand CNV flanking haplotypes and disease relationships, we linked disease association with haplotype association through SNPs. Haplotype-SNP association was calculated by r^2. SNP-disease association was reported by previous GWAS studies. We linked them together by Pearson's correlation coefficient between SNP haplotype R-squared values with GWAS P-value.

The calculation is implanted in a home-made R script "R_square.R". This calculates Pearson's correlation coefficient for each haplotype classes with each IgA nephropathy or periodontitis GWAS SNP P-value. First, SNPs not included in GWAS studies were removed. For IgA nephropathy, GWAS P-value was obtained from IgA Nephropathy TDT analysis data of UK samples with IgA nephropathy (Feehally et al., 2010). Periodontitis GWAS data was from Matthias Munz, et al 2017 (Matthias Munz, 2017). Pearson's correlation coefficient was calculated by R function "cor.test".

2.2.12 Neanderthal introgression length analysis

We defined Neanderthal-derived variants as variants which are absent from African Yoruba (YRI) population and can be matched with the Altai Neanderthal genome assembly. Neanderthal-derived variant detection was done by an R script called "Nea_variants.R". It takes a test variants list, a YRI population variants list and Altai Neanderthal variants list as input. It output a variants list containing "0" and "1", where "1" means identification as a Neanderthal-derived variant. Then this list was used by another r script "length_plot.r" to output a stack image of introgressed segments in the *DEFA1A3* region for European samples analyzed (1000 Genome Project CEU, GBR, TSI). Continuous Neanderthal introgressed variants were merged together to form a longer introgression segment. With this plot, we can estimate the average introgression length and determine the boundary of Neanderthal introgression. Sankararaman in 2014 showed that the median length of Neanderthal contigs is 129kb when merging all Neanderthal haplotypes in 1000 Genomes Project (Sankararaman et al., 2014). So the expected length of introgression at *DEFA1A3* region should also close to this estimation.

2.2.13 Geographic distribution of Neanderthal introgressed haplotype

To study geographic distribution of Neanderthal introgressed haplotype (5T/7C), an R script named "world_distribution.r" was made. It plotted frequency of a haplotype for given populations on a world map once the longitude and latitude coordinates were determined. The frequency data of introgressed haplotypes come from the 1000 Genomes Project, Human Genome Diversity Project (HGDP) (http://www.hagsc.org/hgdp) and The Allele Frequency Database (ALFRED) (alfred.med.yale.edu). For populations from 1000 Genome Project dataset, we used 5T/7C haplotype tag SNP rs62487514 "A" allele frequency to represent introgressed haplotype frequency. For populations from HGDP and ALFRED data set we use SNP rs4500099 due to the absence of rs62487514. This SNP is in good LD with rs62487514 (r^2=0.73). It is also a Neanderthal-derived variant identified as in section 2.2.12.

2.2.14 S* statistics

S* statistics were used for identifying the signature of archaic introgression even without known source of introgression. S* statistics are both a measure of linkage disequilibrium in a phased chromosome and a method to discover linked SNPs. When

introgression first happened, a big block of sequence was transferred from one species to another, recombination can break down this big block into smaller pieces. Under a simple assumption when no migration happened, the length of introgressed tract depends on the time of introgression and local recombination rate. So introgression should generate longer tract than shared ancestry (Racimo et al., 2015). In simple words, S* marks SNPs that are derived and tests the LD between these SNPs, and longer LD will obtain a higher S* mark. In this study, we used YRI population as representative of an ancestral population since the YRI population was less influenced by introgression. S* statistics were calculated by a self-made R script (S.r). This script uses a sliding window algorithm with a window size 6kb and step size 3kb to scan over a 300kb region of *DEFA1A3* CNV. S* values for each sample were calculated by the sum of S(i,j) in each window, S(i,j) is calculated according to the scheme below.

$$S(i,j) = \begin{cases} -\infty \text{ if } j \text{ is a singleton} \\ -\infty \text{ if } |i-j| < 10 \text{ or } |i-j| > 50{,}000 \\ 5{,}000 + |i-j| \text{ if } d(i,j) = 0 \text{ (no mismatches)} \\ -10{,}000 \text{ if } 1 \le d(i,j) \le 5 \text{ (up to 5 mismatches)} \end{cases}$$

$d(i,j)$ = Number of haplotypes that do not match on SNPs i and j
$|i-j|$ = Distance in base pairs between SNPs i and j

Where i and j are the position of two SNPs, if the distance of i and j(|i-j|)is more than 50kb or smaller than 10 bp then we give S(i,j) minus infinity value. If there is no mismatch between two SNPs in a data set for the test (d(i,j)=0), we give S(i,j) 5000 score plus the distance between two SNPs. If there are more than 5 mismatches we

give S(i,j) -10000 value. S* in a window is the sum of all possible combinations of i and j which makes the S(i,j) value maximum. Then S* value is plotted in each window by R plot function. A cluster of windows with high S* value indicates the pattern of highly associated non-ancestral SNP alleles expected from introgression.

2.2.15 Divergence time calculation

2.2.15.1 By Double Poisson distribution plot

To obtain evidence of 5T/7C haplotype as a Neanderthal introgression, we estimated divergence time between 7C and the most closely related human haplotype (3C), and between 7C and the Altai Neanderthal haplotype. If we assume the number of SNP differences between two sequences over a given period of time follows a poisson distribution, then we can calculate the expected divergence time between the two sequences. We obtained 19,406bp (chr8:6879803-6899209) *DEFA1A3* CNV centromeric flanking sequences for all samples in the test. We first call variant for each site in this region by GATK unifiedgenotyper (standard setting, accept genotypes which pass filters), then use GATK FastaAlternateReferenceMaker command to generate full sequence fasta files from vcf files. The *DEFA1A3* local mutation rate was estimated by chimpanzee pantro4 reference sequence and human hg19 reference sequences. We counted the number of nucleotide

differences between two pairs of test sequences by mega6 pairwise distance matrix calculation function.

The divergence time calculation was done by an R script "poisson divergence time.r". It first estimates a local mutation rate by setting chimpanzee and human reference sequence divergence to 6 million years (Scally et al., 2012). There are still some conflicts on human-chimpanzee divergence time. Estimated date range between 3MY to 11MY, this range of dates comes from different estimates of evolutionary distance as well as differences in calibration(Jensen - Seaman and Hooper - Boyd, 2013). Recently, some paper studied generation time in wild chimpanzees and gorillas and found a longer generation time than previous expected, thus suggest a longer divergence between human and chimpanzee at least 7–8 MY (Langergraber et al., 2012). Due to real split date is unknown and the large ancestral effective population size, the divergence time estimated by pedigree calibrated global rate at 6 MY is a safer option and it is also in general agreement with palaeontological evidence. Next we applied this local rate and estimated the distribution of 7C-3C and 7C-Neanderthal divergence time. The expected number of mutations over a period of time is calculated as follows:

$$number\ of\ mutations = \frac{2 * divergence_time - Nea_time}{\text{generation time}} * rate * length$$

Where Nea_time is 50300 years, since the Altai Neanderthal is an ancient sample. This is the carbon dating age of Altai Neandertal sample (Prufer et al., 2014). "Generation time" was set to 25 years for hominin (Scally et al., 2012). The "rate" is estimated *DEFA1A3* local mutation rate by human hg19 reference and chimpanzee pantro4 sequences. "Length" is the total number of nucleotides analyzed (19406bp). Then expected number of mutations and the actual observed number of mutations was then input in a Poisson distribution density function to find out its probability. We did the above calculation for both 7C-Altai Neanderthal and 7C-3C sequence pairs and plotted them together in a 2D plot by the ggplot2 package. Finally, we can calculate 95% and 99% confidence intervals and peak point.

2.2.15.2 By BEAST software

Additional evidence of divergence time estimation between 7C haplotype and Altai Neanderthal, and 7C-3C, comes from BEAST software. BEAST is a software for Bayesian analysis of molecular sequences by using MCMC. It can be used as a method of reconstructing phylogenies but is also a framework for testing evolutionary hypotheses without conditioning on a single tree topology (Drummond et al., 2012). Here we used BEAST version 1.8.1 to estimate divergence time. We prepared 19406bp *DEFA1A3*

CNV centromeric flanking sequence for 7C, Altai Neanderthal, human 3C haplotype and pantro4 chimpanzee reference sequences. All sequences were aligned by MUSCLE and saved in fasta format.

The parameter setting was as follows: Three sets of TMRCA were set: human 3C to chimpanzee, 7C to human 3C, 7C to Altai Neanderthal. Substitution model: HKY model. Site heterogeneity model: gamma. Clock Model: Lognormal relaxed clock (uncorrelated). Speciation: Yule process. TMRCA between human chimpanzees: normal distribution, mean 6 million, standard deviation 0.5 million. MCMC chain length set to 10,000,000. Log file was viewed by Tracer v1.6.

2.2.16 Selection analysis by integrated haplotype score

Integrated Haplotype Score (iHS) is a way of measuring the level of positive selection for a core variant. It is derived from extended haplotype homozygosity (EHH), where EHH measures the level of LD for a given SNP. It is defined as the probability that two randomly chosen chromosomes carrying the core haplotype of interest are identical by descent for the entire interval from the core region to the point x (Sabeti et al., 2002). The haplotype with high EHH value and high frequency usually indicates it rose in frequency recently under positive selection. iHS is defined as follows:

$$\text{unstandardized } iHS = \ln\left(\frac{iHH_A}{iHH_D}\right).$$

$$iHS = \frac{\ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{\text{SD}_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}$$

Integrated EHH (iHH) is the area under EHH curve. iHH_A is the area of ancestral allele and iHH_D is the area of derived allele. iHS computes the area differences between the ancestral and derived alleles. It is usually standardized (mean 0, variance 1) empirically to the distribution of a set of observed iHS in a larger region. If the standardized iHS value is negative, it means that the derived allele maintains a larger LD than the ancestral allele and vice versa.

We calculated iHS values for 2455 SNPs in a 300kb region flanking the *DEFA1A3* CNV (chr8:6700000-7000000), including 38 Neanderthal-derived SNPs identified in section 2.2.12. Calculation is achieved by a self-made R script "EHH.r". R package "rehh" is used in this script, which contains a set of functions for the computation of EHH, iHS and Rsb or XP-EHH (across pairs of populations) statistics and various plot functions (Gautier and Vitalis, 2012). iHS values for young and rare SNPs will tend to be more negative than other SNPs, since time is too short for recombination events to occur, so the iHS value is only comparable after standardization with SNPs of similar

frequencies. We standardized our raw iHS value by using a set of genome-wide iHS values calculated from the 1000 Genome Project phase 3 CEU samples. This file contains pre-calculated mean iHS value for every 0.02 frequency category obtained from Benjamin F Voight's lab (unpublished, http://coruscant.itmat.upenn.edu/).

To evaluate whether the iHS selection analysis method was biased by introgression, We plotted empirical iHS distributions of whole genome Neanderthal introgressed SNPs and non-Neanderthal introgressed SNPs. Whole genome SNP iHS data was acquired from Voight et al 2006, including 8,040,865 SNP iHS scores calculated from the 1000 Genomes Project CEU population. Sankararaman et al 2014 inferred 97,365 Neanderthal introgressed SNPs from the whole genomes of the 1000 Genomes Project CEU population samples. We filtered for SNPs with an average posterior probability of Neanderthal on derived alleles >0.9 and with Neandertal DAF=1, YRI DAF=0. Then we combined these two data sets by finding the intersection of SNPs between them. There are 55,044 Neanderthal-derived SNPs with an iHS score calculated. The plots were generated by self-made r script "iHS_plot.r".

2.2.17 *DEFA1A3* CNV internal nucleotide diversity analysis

To compare the within-haplotype nucleotide diversity between different haplotype classes, we calculated the average nucleotide

diversity per base pair (π). This calculation is only possible for a sample carrying homozygous haplotypes as phasing inside CNV is not possible. 38 samples with homozygous telomeric and centromeric haplotypes from the 1000 Genomes Project were included in this analysis. The calculation is done by a self-made r script Pi.r. This script takes VCF variant calling from inside CNV as input. A R function nuc.div() from R package "pegas" calculated π value for each position. A histogram plot displayed the average π value for each haplotype class analyzed.

2.3 Experimental methods

2.3.1 General methods and reagents

2.3.1.1 Primer design

All primers described in this thesis were designed by primer3 web primer design tool (Untergasser et al., 2012). Human reference genome hg19 was used as the source of target DNA sequences. Primers were chosen to have similar annealing temperature, length (~20bp), GC content and thermal conditions, without primer hetero-dimer, self-dimer or hairpin structure. Primers were further checked against dbSNP databases to ensure no variations overlap primers. UCSC genome browser in-silico PCR tool and Blast tool were used to

make sure of unique primer binding sites in the human genome. All primers used were produced by Integrated DNA Technologies (IDT) Company.

2.3.1.2 10X L.D. mix buffer

10X L.D. mix buffer was used when Taq polymerase was chosen as enzyme in PCR reaction. 10X L.D. mix is a locally-made reagent containing 500mM Tris HCl (pH8.8), 125mM Ammonium Sulphate, 14mM $MgCl_2$, 75mM 2-mercaptoethanol, 2mM each dNTP and 1.25mg/ml BSA.

2.3.1.3 Capillary electrophoresis

Capillary electrophoresis was performed on an ABI genetic analyzer 3130xl machine. For each 16 sample input, 2µl ROX 500 marker and 170µl of Hi-Di formamide were mixed (Applied Biosystems). Hi-Di is highly deionized formamide used to resuspend samples before electrokinetic injection on capillary electrophoresis systems. For each sample 10µl ROX/Hi-Di is added, and denatured by heating at 96°C for 3 minutes in a PCR machine and then put on ice for 2 minutes. After that samples were ready to load into the ABI machine with usually 1KV, 30s injection time. Resulting data was analyzed by GeneMapper v4.1, and peak height or area was used to measure the ratio between different DNA fragments.

2.3.1.4 Sanger sequencing

We cleaned PCR products by AmpureXP (Agencourt) according to the standard protocol. Approximately 20ng of DNA was used in a Sanger sequencing reaction. After the sequencing reaction, the PCR product was further cleaned by CleanSEQ (Agencourt) and DNA powder was sent to DBS genomics (https://www.dur.ac.uk/biosciences/services/dna/) or DeepSeq (http://www.nottingham.ac.uk/deepseq/) for capillary electrophoresis. Returned sequencing trace data was visualized by BioEdit or SeqMan (DNA Star) software.

2.3.2 Emulsion haplotype fusion-PCR

Emulsion haplotype fusion-PCR (EHF-PCR) was used to identify the position of *DEFA1A3* CNV internal variants in the repeat array. It is a combination of emulsion PCR, fusion PCR and haplotype PCR developed by Tyson et al. in 2012 (Tyson and Armour, 2012). It links one region from inside CNV with one region from flanking region that are far away from each other which cannot be coamplified by ordinary PCR (fig 2.2). The two regions can be fused together by a tailed primer. Oil droplet PCR makes sure the fusion only happens between PCR products initially arising from single DNA molecules. A second round amplification using allele specific primers distinguished

the two chromosomes. Finally, by Sanger sequencing, we could identify the position of variants in the CNV repeat array.

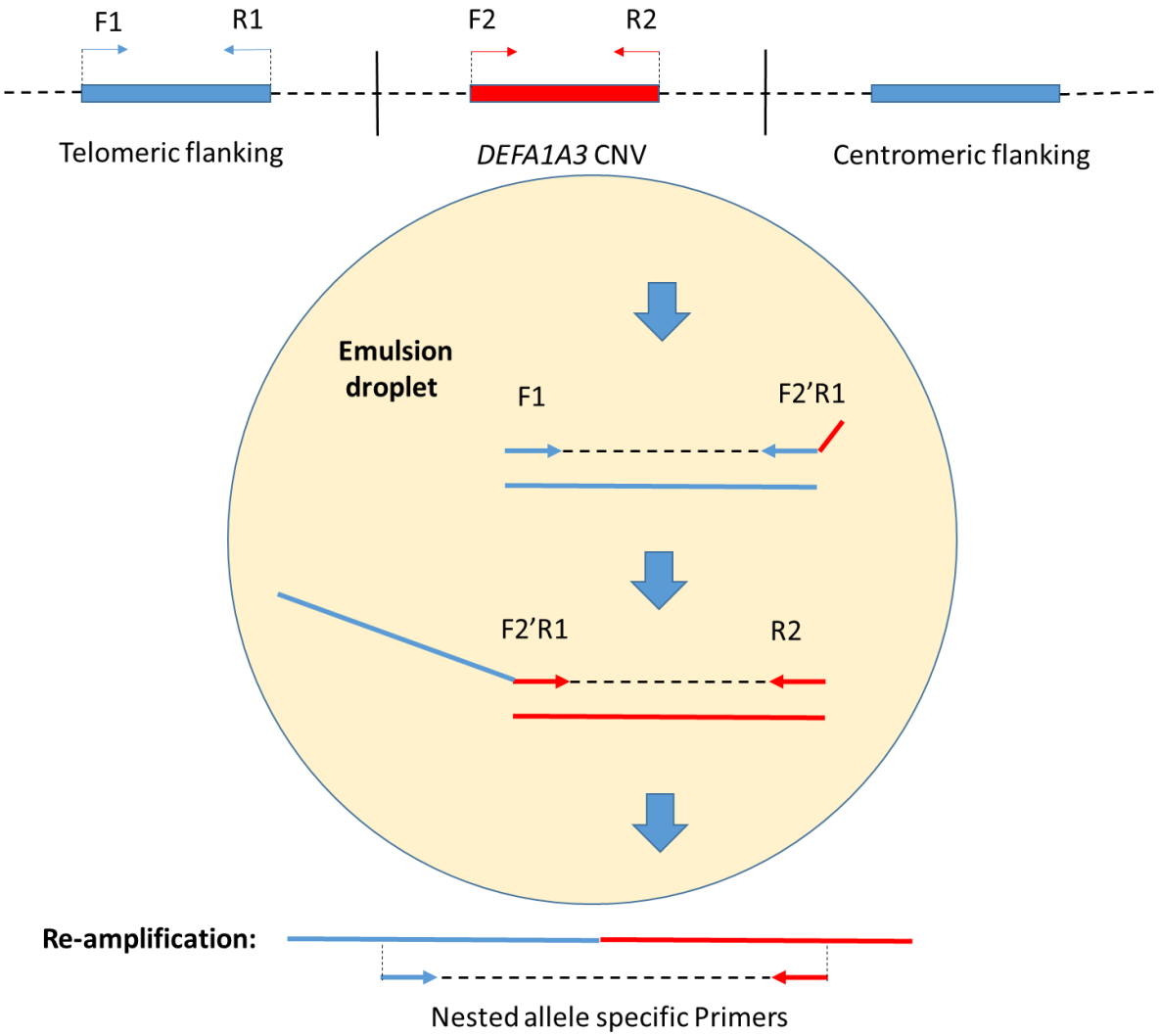


Figure 2.2 Principle of emulsion fusion haplotype PCR: Two pairs of primers (F1, R1, and F2, R2) amplify a *DEFA1A3* flanking region and a CNV internal region, F2'R1 was designed to fuse two regions together. Emulsion droplet PCR guarantees that the initial PCR reactions occurred on the same DNA molecule. Second round amplification phases the two chromosomes by using nested allele-specific primers.

For each sample a standard 25µl EHF-PCR master mix contains, 10µM F1 primer and 10µM R2 primer, 1µM F2'R1 primer, 1X Phusion, 5X GC

buffer, 10mM dNTPs and 10ng/µl input DNA. Primers used are shown in table 2.5.

50µl silicone oil was added to the tube after PCR master mix was prepared. Per 100ml this oil mix contains 40ml silicone polyether/ Cyclopentasiloxane (Dow Corning), 30ml Cyclopentasiloxane/ trimethyl siloxy silicate (Dow Corning) and 30ml AR20 silicone oil (Aldrich) as described by Turner et.al (Turner and Hurles, 2009). A 3mm tungsten bead was added to the tube and shaken at speed 5 in a vortex mixer for 1min 30s for making emulsion droplets.

The PCR conditions for the first round of amplification were as table 2.4.

Table 2.4 PCR conditions for the first round amplification of emulsion fusion haplotype PCR

Step	Temperature	time
Initial denaturation	98°C	30 seconds
40 cycles	98°C	10 seconds
	64°C	30 seconds
	72°C	2 minutes
Final extension	72°C	5 minutes

After first round amplification, PCR products need to be recovered from oil phase for second round amplification. For this purpose, we performed hexane extraction. 25µl 1XGC buffer was added to PCR products and transferred to a fresh tube to dilute the fused PCR

product. Then we added 200µl hexane to each reaction tube and vortex mixed well. The mixture was then centrifuged under 13000g for 3 minutes. The supernatant was removed by pipette. We repeated hexane extraction three times. Finally, we keep the samples at room temperature to allow evaporation of residual hexane.

Table 2.5 primer sequences used for EHF-PCR systems to study position of MSV 10479.

Primer	MSV 10479 (Tel)
F1	GCAGCAAAGACACAGGCAAT
R1	AACATCCCAACCCAGCCAAT
F2	AGTCAGCAACAGGGACCTTC
R2	GTGTCAGCTTAAAGTTCACGCA
F2'R1	GAAGGTCCCTGTTGCTGACTAACATCCCAACCCAGCCAAT
nested F	TTTGGAGGGACGCTGTGA
nested R	TGTCAGCTTAAAGTTCACGCA

Second round amplification was used for enrichment and purification of fused PCR products. Nested allele-specific primers were also used for phasing the two chromosomes. For each sample re-amplification PCR master mix contains, 10µM Forward primer, 10µM Reverse primer, 0.05U/µl Taq polymerase, 10X L.D. mix, and 10ng/µl input DNA. PCR conditions shown in table 2.6.

Table 2.6 PCR conditions for the second round amplification of emulsion fusion haplotype PCR

Step	Temperature	time
Initial denaturation	94°C	30 seconds

30 cycles	94°C	30 seconds
	65°C	30 seconds
	65°C	3 minutes
Final extension	65°C	10 minutes

After the second round of PCR amplification, the PCR product was purified and set up for Sanger sequencing as described in section 2.3.1. Returned sequence data were analyzed by BioEdit and SeqMan software. The positions of variants in the CNV repeat array were determined by the strength of peak signal. The variants with the highest peaks are located in most centromeric or telomeric repeat in the array. A lower peak suggests it was located in the second most centromeric or telomeric repeat. Third copy variants are hardly seen on the sequencing trace.

2.3.3 Experimental verification of recombinant haplotypes by family segregation

Haplotype tag SNPs were selected for main haplotype classes according to the results in chapter 3.5. We Sanger sequenced these tag SNPs for each individual in HapMap Project family samples. Then we used their family relation information to work out the exact phase of two haplotypes and verify the presence of recombinant haplotypes from NGS data analysis. We genotyped three families, described in chapter 3.8. A summary of tag SNPs and their position is shown in

table 2.7 and table 2.8. Primer design for amplifying these SNPs is shown in fig 2.3 and table 2.9.

Table 2.7 best associated tag SNPs for major telomeric haplotype classes

Telomeric haplotype	SNP rs ID	R Square	Haplotype tag allele	Hg19 Reference allele
4T	Rs2702912	1	C	C
1T	Rs2977789	0.92	T	T
2T	Rs2702910	0.93	A	G
5T	Rs79734934	1	T	G

Table 2.8 best associated tag SNPs for major centromeric haplotype classes

Centromeric haplotype	SNP rs ID	R Square	Haplotype tag allele	Hg19 Reference allele
1C	Rs7814783	0.8	T	C
2C	Rs6605578	0.81	T	T
3C	Rs4313182	0.93	T	T
7C	Rs62487513	1	G	C
9C	Rs7841223	0.93	A	G

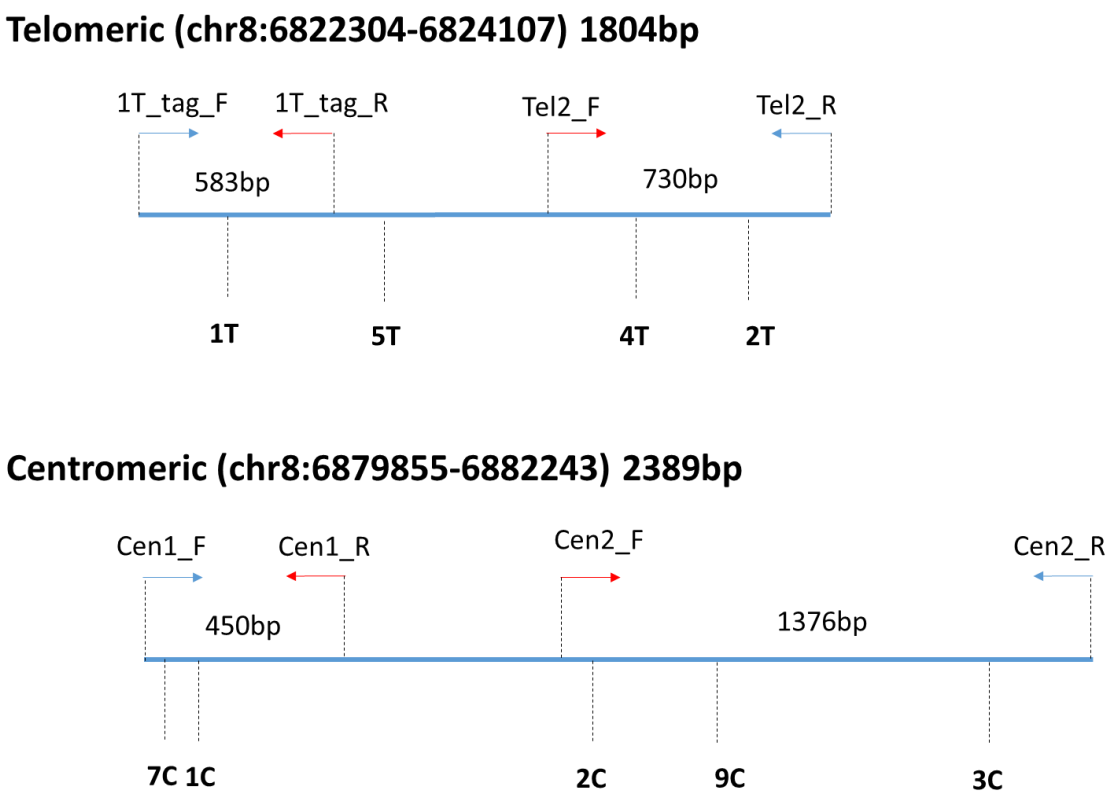


Figure 2.3 Experiment design for verifying recombination between *DEFA1A3* CNV telomeric and centromeric haplotypes.

Table 2.9 Primers used for genotyping *DEFA1A3* flanking haplotypes.

primer name	primer sequences	product size	tag SNPs included
1T_tag_F	GTCAGACAATGCTGCCTTCA	583bp	1T
1T_tag_R	CAGTCCTGTCCTGCTTCAAG		
Tel2_F	GCGAAAATTACCCTGGTTGTC	730bp	4T,2T
Tel2_R	AGAAGGTGAGCTGCAAGTTT		
Cen1_F	GATCAGGCGCTAGTGAACTG	450bp	1C,7C
Cen1_R	GTCACCCTCACTACCTCTCAG		
Cen2_F	CTTGGTGAGAAGTTGAATGTGTG	1376bp	2C,9C,3C
Cen2_R	TTTCACCAGCTCACCCTAGG		

All primers were designed to have similar thermal conditions, so they can be used in multiple combinations. For each sample 10μl PCR master mix contains, 10μM Forward primer, 10μM Reverse primer, 0.05U/μl Taq polymerase, 10X L.D. mix, and 10ng/μl input DNA. PCR conditions in table 2.10 (same for all primer pairs):

Table 2.10 PCR condition used for recombinant haplotype verification experiment

Step	Temperature	time
Initial denaturation	95°C	5 minutes
33 cycles	95°C	30 seconds
	63°C	30 seconds
	65°C	4 minutes
Final extension	65°C	10 minutes

PCR products were cleaned by AmpureXP, followed by standard Sanger sequencing procedures. 1T tag SNP was sequenced by primer 1T_tag_F, 4T tag sequenced by Tel2_F, 7C and 1C tag were sequenced by Cen1_f, 2C and 9C sequenced by Cen2_F, 3C tag SNPs sequenced by Cen2_R. After gathering diploid genotypes for all

samples, we phased individual telomeric and centromeric haplotypes by hand.

2.3.4 Experimental verification of MSV 10479

Our analysis suggested MSV 10479 is a 1T/1C haplotype tag variant inside the *DEFA1A3* CNV. It appears to be present on almost every copy of 1T/1C haplotypes but hardly present on other haplotypes. To verify this conclusion we designed an experiment that amplifies a 1004bp fragment which contains MSV 10479 from all copies from inside *DEFA1A3* CNV. We sequenced 20 samples from HapMap CEU, JPT, CHB, YRI populations. 1003bp fragments were amplified by forward primer (GAGGATCCAGCACAGAGAGG) and reverse primer (TCTCAAAGAGCTCCCGTGTG). 10ul PCR master mix was prepared for each sample, contains 10μM Forward primer, 10μM Reverse primer, 0.05U/μl Taq polymerase, 10X L.D. mix, and 10ng/μl input DNA. PCR conditions were as in table 2.11.

Table 2.11 PCR conditions for MSV 10479 verification experiment

Step	Temperature	time
Initial denaturation	95°C	5 minutes
34 cycles	95°C	30 seconds
	63°C	30 seconds
	65°C	3 minutes 30 seconds
Final extension	65°C	10 minutes

The Sanger sequencing protocol was described in section 2.3.1. The reverse primer was used as sequencing primer since MSV 10479 is located at 243bp from the reverse direction. The genotype of MSV 10479 was analyzed by BioEdit version 7.2.5.

2.3.5 Experimental verification of MSV 2284

A PCR-RFLP was designed to verify MSV 2284. HapMap CHB/JPT samples were used as experimental material. The presence of MSV 2284 G will change a CCGC site to GCGC site which is a restriction enzyme cutting site for *Hha*I or *Hin*PI, and thus can be used for verification of MSV 2284. A pair of primers (Forward: TGAGACCCAACTCTGGAACC Reverse: CACTCCAGGCAAGAGCTGAT) was designed to amplify 188bp long sequences containing MSV 2284. The forward primer was labeled with FAM. In the case of presence of MSV2284 GCGC site is created and can be digested by *Hha*I and produces a 135bp and 53bp fragment. The larger 135bp fragment is labeled with FAM and can be detected by capillary electrophoresis.

10μl PCR master mix was prepared for each sample, consists 10μM Forward primer, 10μM Reverse primer, 0.05U/μl Taq polymerase, 10X L.D. mix, and 10ng/μl input DNA. PCR conditions shown in table 2.12.

Table 2.12 PCR condition for MSV 2284 verification experiment

Step	Temperature	time
Initial denaturation	95°C	5 minutes
36 cycles	95°C	30 seconds
	60°C	30 seconds
	70°C	30 seconds
Final extension	70°C	60 minutes

Capillary electrophoresis by ABI 3130xl genomic analyzer was used to visualize the FAM labeled fragments. DNA fragment peak area was used for ratio calculation between cut and uncut fragments.

A PCR followed by Sanger sequencing was applied to MSV 2284 for further verification. A pair of primers (Forward: GTGAAGCCACCTCAGTGACA Reverse: ACCTCCAGAGAACAGCAGGA) was designed to amplify 688bp long DNA sequences containing MSV 2284. 10µl PCR masters mix contain 10µl Forward primer, 10µl Reverse primer, 0.05U/µl Taq polymerase, 10X L.D. mix, and 10ng/µl input DNA. PCR condition shown in table 2.13.

Table 2.13 Sanger sequencing PCR condition MSV 2284 verification experiment

Step	Temperature	time
Initial denaturation	95°C	5 minutes
35 cycles	95°C	30 seconds
	57°C	30 seconds
	70°C	30 seconds
Final extension	70°C	60 minutes

A pair of nested primers (forward: TGAGACCCAACTCTGGAACC Reverse: CATTTGTTCCCTGAGCCTGT) was used in the Sanger sequencing reaction. The sequencing capillary electrophoresis was carried out by DBS Genomics, Durham University. Data was viewed and analyzed by BioEdit software.

2.3.6 Experiment to detect PCR-mediated recombination in long PCR

To test whether PCR-mediated recombination was occurring in our *DEFA1A3* CNV long PCR haplotype reconstruction approach, we designed a restriction enzyme digest experiment. We used this experiment to examine PCR1 of sample NA06990. The first step was to use PCR1 forward primer (ACCCATGCACAGGAGAGGAT) and reverse primer (CCAATGCTTAAACACCAGCAG) to amplify a 10208bp product from genomic DNA input. For each sample 20ul PCR master mix was prepared, containing 10μM Forward primer, 10μM Reverse primer, 5U Long Amp polymerase, 5X Long Amp buffer, 10mM dNTPs and 10ng/μl input DNA. PCR conditions in table 2.14.

Table 2.14 PCR condition for detect PCR-mediated recombination experiment

Step	Temperature	time
Initial denaturation	95°C	30 seconds
28 cycles	94°C	30 seconds
	62°C	15 seconds
	65°C	11 minutes
Final extension	65°C	20 minutes

For second round PCR amplification we used the PCR product as input and diluted 20 times, with the PCR condition is same to the first round amplification, but reducing the PCR cycles to 15. Then the PCR product was digested with NEB *Sma*I restriction enzyme along with NE buffer 4 at 25°C overnight.

Finally, we loaded the digested PCR product with 5X NEB blue loading dye on 1.5% agarose gel and ran electrophoresis for up to 40 hrs under 100V voltage. NEB 1kb DNA ladder was used as a marker to determine fragment size. Gel image was checked under UV light.

2.3.7 Experiment to reduce PCR-mediated recombination

To reduce PCR-mediated recombination, PCR with double primer and double dNTPs concentrations were evaluated. For the double primer concentration experiment, the PCR master mix and conditions are the same as above (Section 2.3.6) except that both forward and reverse primer input was doubled from 0.6µl to 1.2µl (i.e. from a final concentration of 0.3µM to 0.6µM). For the double dNTPs experiment we doubled dNTPs input from 0.6ul to 1.2µl (final concentration from 0.3µM to 0.6µM) and also added 1.2µl of 20mM $MgSO_4$ since it is a cofactor for dNTPs. The *Sma*I enzyme digestion and gel electrophoresis steps are also the same as above.

3 Chapter3: Human *DEFA1A3* CNV flanking haplotype diversity

3.1 Introduction

Previous association studies showed there are multiple factors near the *DEFA1A3* CNV region that are associated with several diseases, and the haplotype diversity within this locus may be key to understanding the associations (Li et al., 2015, Kiryluk et al., 2014, Yu et al., 2012, Matthias Munz, 2017). Therefore it is important to understand haplotype diversity of *DEFA1A3* CNV in the global human populations.

Black et al carried out a diversity study in a European population based on a 4.1kb centromeric flanking region and identified five classes of *DEFA1A3* haplotypes, which were termed Reference, Class1, Class2, Exchange1 and Exchange2 (Black et al., 2014). Each haplotype has unique SNP background and characteristic CNV internal structure. Reference haplotype is so named as it has the same SNP background as the human reference genome hg19; it is usually associated with a high number of copies of *DEFA1A3* repeats. In these haplotypes, there is usually a *DEFA3* copy, which mostly appears on the most centromeric position. This is supported by

emulsion fusion haplotype PCR experiments. Exchange-2 and Exchange-1 haplotypes are named because they carry replacement polymorphisms: For exchange-2, a 1.99kb sequence of the *DEFA1A3* centromeric partial repeat has been exchanged with sequence from the equivalent region of the *DEFA1A3* full repeats. Similarly, a ~2.2kb telomeric replacement polymorphism is also often found on these haplotypes. Exchange-2 haplotypes usually do not carry *DEFA3* copies. Exchange-1 haplotype has a 140bp centromeric replacement polymorphism and similarly to Exchange-2, no *DEFA3* copy was found in the repeat array. Class-2 haplotype has a characteristic feature of association with low copy number and the most telomeric copy is usually occupied by a *DEFA3* copy. All haplotype classes can be identified by a set of SNPs in the centromeric flanking region (table 3.1).

Table 3.1 Five major haplotypes in European populations and their SNP tags in the CNV centromeric flanking region.

Haplotype class	SNP Genotypes			
	rs4300027	rs7825750	rs7826487	rs62487514
Reference Sequence	T	T	A	C
Class 1	C	T	G	C
Class 2	C	C	A	C
Exchange 1	T	C	A	A
Exchange 2	C	C	A	C

3.1.1 Black's 5 haplotype classification limitation and problems

Although Black's classification explained diversity in European populations very well, it is still lacking evidence to describe global haplotype diversity. Populations other than Europeans could have very different diversity and using European tag SNPs to classify other populations may cause problems. For example, IgA Nephropathy is a rare disease in Africans and has the highest risk in Asian populations. Association in rs2738048 was found to be Asian-specific, and this suggested there is a potential difference between European and Asian diversity. Moreover, on the basis of initial analysis of the 1000 Genomes Project data we observed that Class-1 haplotypes in Europe nearly always associated with low copy number, but almost all Asian Class-1 haplotypes have very high copy number (Figure 3.1). Black's classification was based on centromeric flanking sequences only, and did not consider telomeric flanking sequences. In fact, we observed some different haplotype classes having identical telomeric flanking sequences or same haplotype classes having very different telomeric flanking sequences (Figure 3.2). These observations suggest the haplotype diversity in the *DEFA1A3* region is more complex at a global level and Black's 5 haplotype classification has limitations to cope with this complexity. Therefore a more comprehensive and

detailed classification system is needed to form a base for our understanding of complex multiple association signals at this locus.

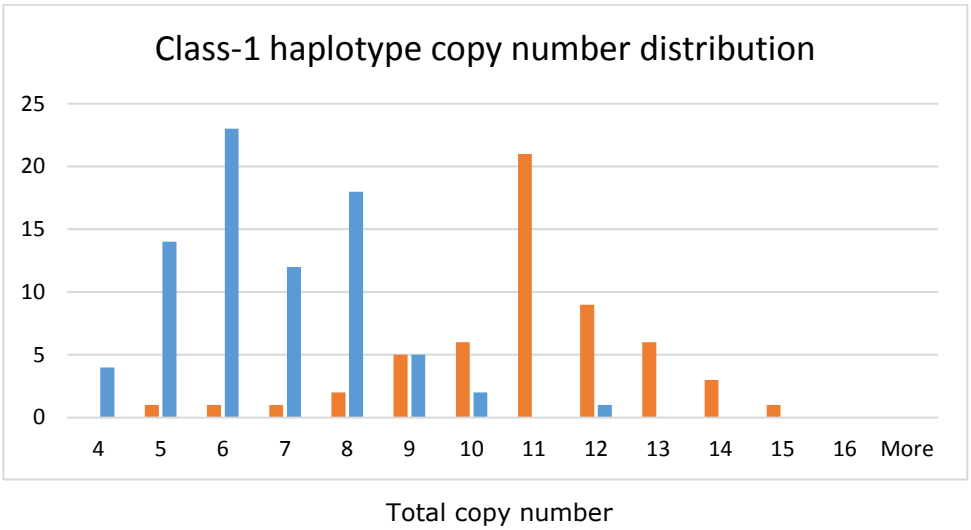


Figure 3.1 Class-1 haplotype carrier total copy number distribution in the 1000 Genomes Project European (blue) and Asian (red) populations. Asian Class-1 haplotypes appear to have much higher total copy number than European Class-1 haplotypes.



Figure 3.2 Comparing 6kb telomeric variable sites shows Japanese Class-1 haplotype (NA18968) is highly similar to Reference haplotype (HG00132) but highly diverged from another (European) Class-1 haplotype (NA12812).

3.2 Global *DEFA1A3* flanking haplotype analysis

The best CNV haplotype construction would use direct sequencing across the full CNV repeat unit block to reconstruct their spatial

arrangement. However, this is not possible based on the existing methods as each full repeat unit is 19kb long and on average each person carries 6 such copies. We decided to study CNV flanking sequence diversity as CNV and nearby flanking region are in a good LD block, SNPs either side of the *DEFA1A3* locus displaying high levels of linkage disequilibrium (D' = 1, data from phased HapMap CEU1 individuals) (Black et al., 2014), so the flanking haplotypes should be able to represent CNV haplotypes.

We used the 1000 Genomes Project phase 3 nine populations for haplotype analysis; the population codes used in this study can be found in table 3.2. The overall data analysis flow chart is depicted in fig 3.3 and method details can be found in chapter 2.2.3. We first downloaded 300kb (from chr8:6,700,000-7,000,000) of phased SNP data from the 1000 Genomes Project website. Then we did a population recombination rate analysis by INTERVAL implanted in RDP4 software (Martin et al., 2015). Mean rho per base plot is shown in fig 3.4. We observed a recombination peak in the *DEFA1A3* CNV position, and this suggested there is a number of recombination events that happened inside the CNV. Therefore we decided to analyse telomeric and centromeric haplotypes separately. We chose a 6kb telomeric region and a 20kb centromeric region between two recombination peaks to minimise the complexity of diversity caused by recombinant haplotypes. We then clustered closely related

haplotypes into one group by pairwise differences smaller than 1% in 1000 genomes SNP positions. This resulted in discovering 126 telomeric haplotypes and 171 centromeric haplotypes. If we consider that haplotypes with frequency >1% are common haplotypes, there are 14 common telomeric and 17 common centromeric haplotypes. This subset accounted for 81.8% (1421 sequences) and 74.7% (1299 sequences) of total sequences analysed in telomeric and centromeric regions.

Table 3.2 The 1000 Genomes Project populations used in haplotype analysis

Population code	Super population	population
CEU	European	Utah Residents (CEPH) with Northern and Western Ancestry
GBR	European	British in England and Scotland
FIN	European	Finnish in Finland
CHB	Asian	Han Chinese in Bejing, China
CHS	Asian	Southern Han Chinese
JPT	Asian	Japanese in Tokyo, Japan
ASW	African	Americans of African Ancestry in SW USA
LWK	African	Luhya in Webuye, Kenya
YRI	African	Yoruba in Ibadan, Nigeria

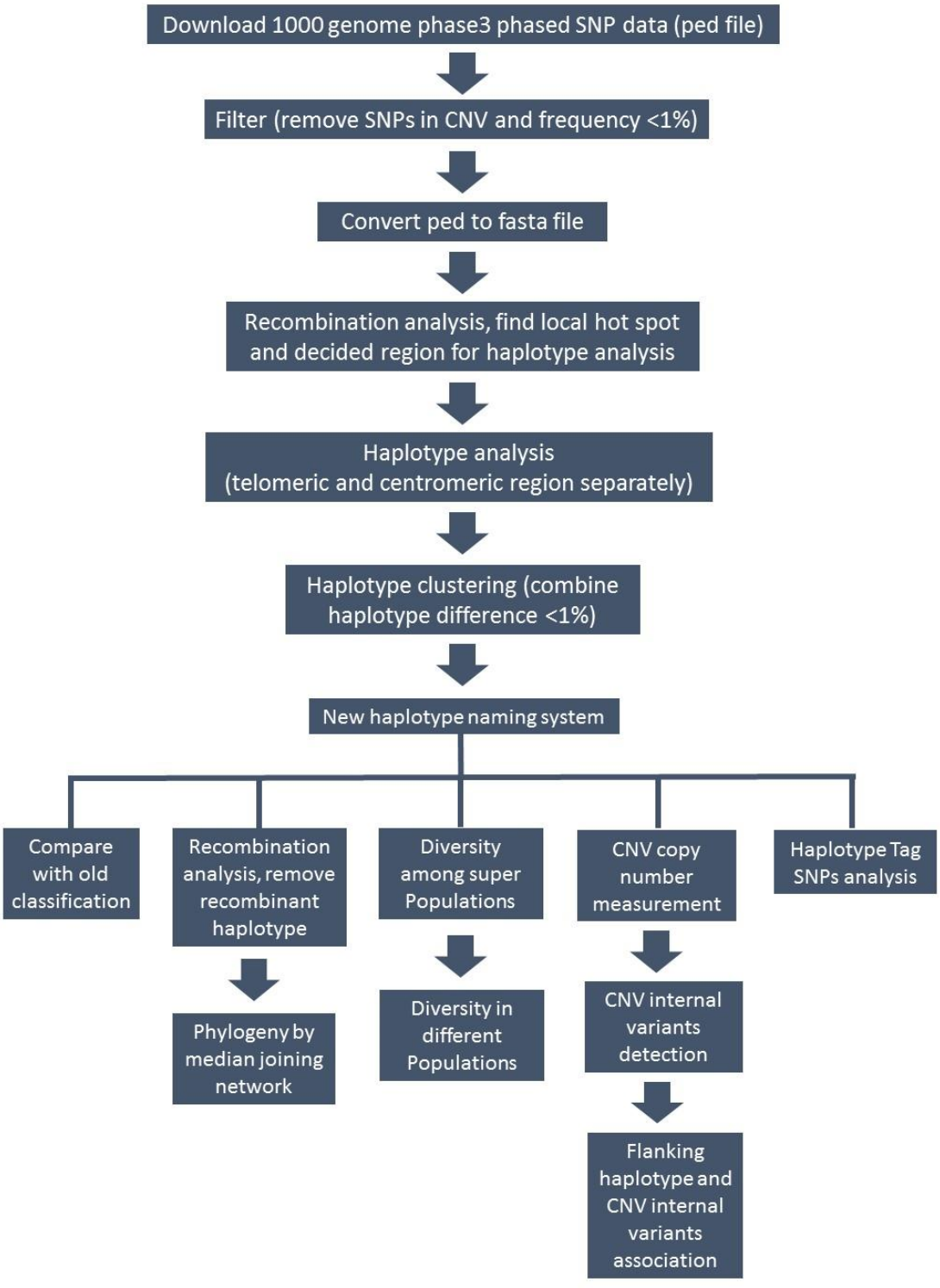


Figure 3.3 Flow chart of *DEFA1A3* CNV flanking haplotype diversity analysis.

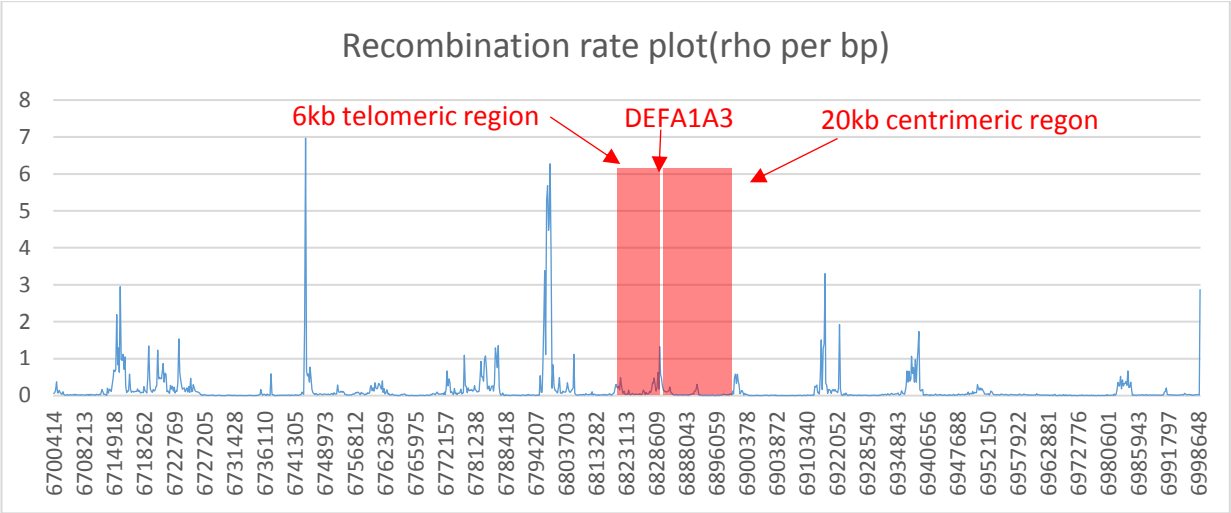


Figure 3.4 Population scale recombination rate estimation (rho per bp) for a 300 kb region near the *DEFA1A3* CNV. The 6kb telomeric region and 20kb centromeric region chosen for haplotype analysis are shown in red blocks.

3.2.1 Naming of new haplotype classes

We analysed telomeric and centromeric flanking haplotypes separately, so each *DEFA1A3* haplotype contained two parts. Telomeric haplotypes were named as XT, where X is the haplotype number ranked according to their frequency in world populations. For example, 1T is the most frequent telomeric haplotype and 5T is the fifth most frequent telomeric haplotype in the world. Similarly, centromeric haplotypes were named as XC, X also ranked according to frequency in world populations. Finally, we combined telomeric and centromeric haplotypes together to name a full haplotype, for example 1T/1C.

3.3 Human *DEFA1A3* Flanking haplotype diversity

3.3.1 Global haplotype diversity

We used a matrix to display relationships between common telomeric and centromeric haplotypes (fig 3.5). Telomeric haplotypes are shown in the x-axis and centromeric haplotypes are shown in the y-axis. The number in each grid represents the frequency of this haplotype. Among them, 1T/1C is the most frequent haplotype combination in humans, followed by 2T/2C, 3T/4C, 2T/5C, 5T/7C and 1T/6C. A lot of haplotypes have shared telomeric/centromeric parts with other haplotypes, and these are probably caused by recombination; this observation is also consistent with recombination hotspot analysis mentioned in section 3.3.2. If the recombination breakpoint is present inside CNV, then this is very likely to change CNV internal variant components.

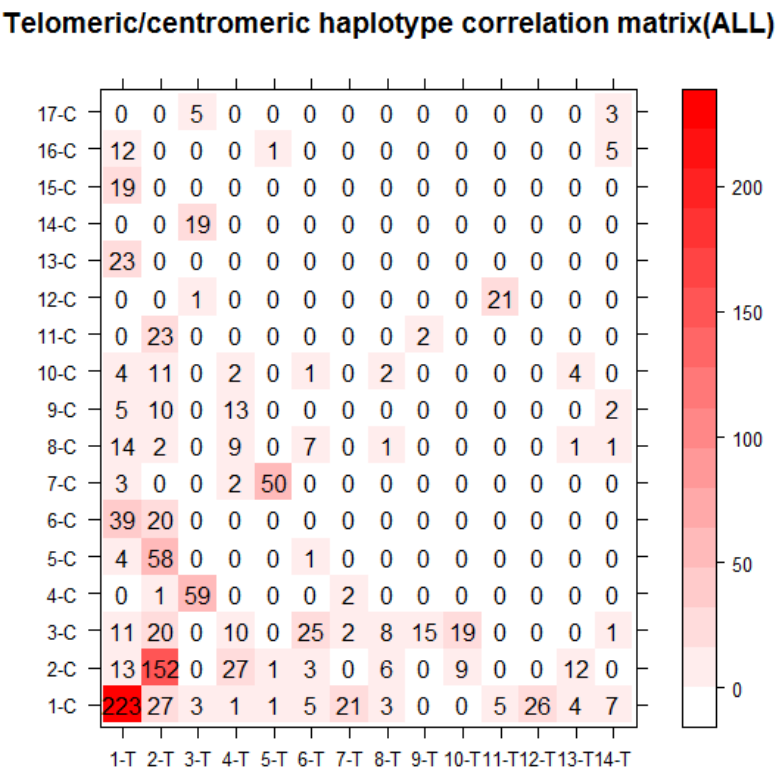


Figure 3.5 Diversity and relationships of *DEFA1A3* telomeric and centromeric flanking haplotypes in all populations.

3.3.2 Diversity between Asian, African and European populations

To understand the difference in haplotype diversity among different super populations, we analysed diversity in Africans, Asians and Europeans separately (fig 3.6). From the plot, we can see that Europeans and Asians share a lot of common haplotypes, but they also possess their own unique haplotypes. In both super populations 1T/1C and 2T/2C are the two most frequent haplotypes. But the third common haplotype in Europe is 5T/7C and is a unique haplotype

found only in Europe. In Asia the third most frequent haplotype is 2T/5C; it is also a unique haplotype found nowhere else. Haplotype diversity is very different in the African population; the top two haplotypes 1T/1C and 2T/2C are relatively rare among Africans. Instead, unique haplotype 3T/4C is the most frequent. Haplotype 11T/12C, and recombinant haplotypes 3T/14C, 12T/1C and 2T/11C are also only found in Africa. These are new haplotypes that were not described by Black's 5 haplotype classification. In all, this analysis shows the haplotype diversity in different super populations is very different in terms of haplotype classes and frequency.

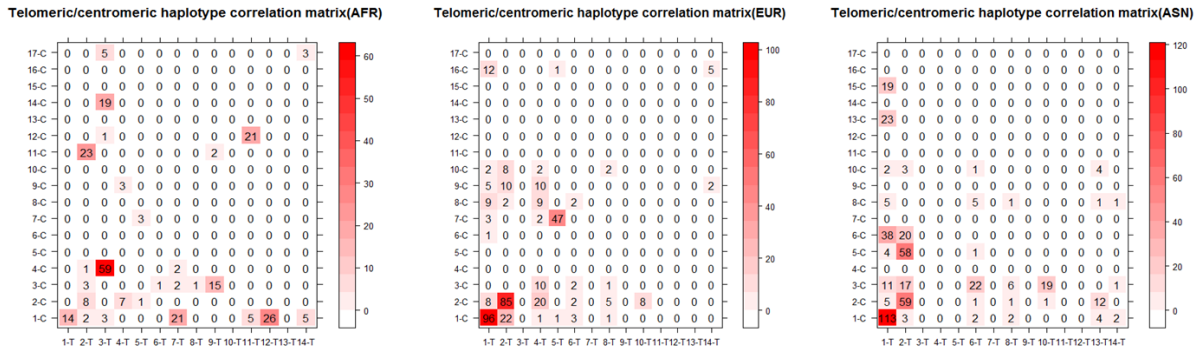


Figure 3.6 Diversity and relationships of *DEFA1A3* telomeric and centromeric flanking haplotypes in African, European and Asian populations, showing that Asians and Europeans share most common haplotypes, but unique haplotypes are also present. In Africa most haplotypes are not seen in Europe and Asia, shared haplotype also in low frequency.

3.3.3 Diversity between Chinese and Japanese

A previous PhD thesis studied all SNP frequency differences between Chinese and Japanese on chromosome 8. It reported SNPs associated with copy number in Chinese Japanese populations also highly differentiated in the two populations (among top 2% of frequency differences) (Khan, 2012). This indicates there may be big diversity differences between the two populations. Here we compared flanking haplotype diversity between 1000 Genomes Project JPT and CHB+CHS samples (Fig 3.7). Unlike what we see in Europe and China, in the Japanese population the most frequent haplotype is 2T/5C; the frequency reached 17.8%, and 33% of Japanese carry at least one copy of this haplotype, while in the Chinese population the frequency is only about 5%. The most frequent haplotype among Chinese is 1T/1C (21%) which is also the highest among Europeans. Other than frequency differences between the two populations, there are more recombinant haplotypes present in the Chinese populations; for example, 1T/15C, 2T/6C and 13T/2C are more frequent among Chinese.

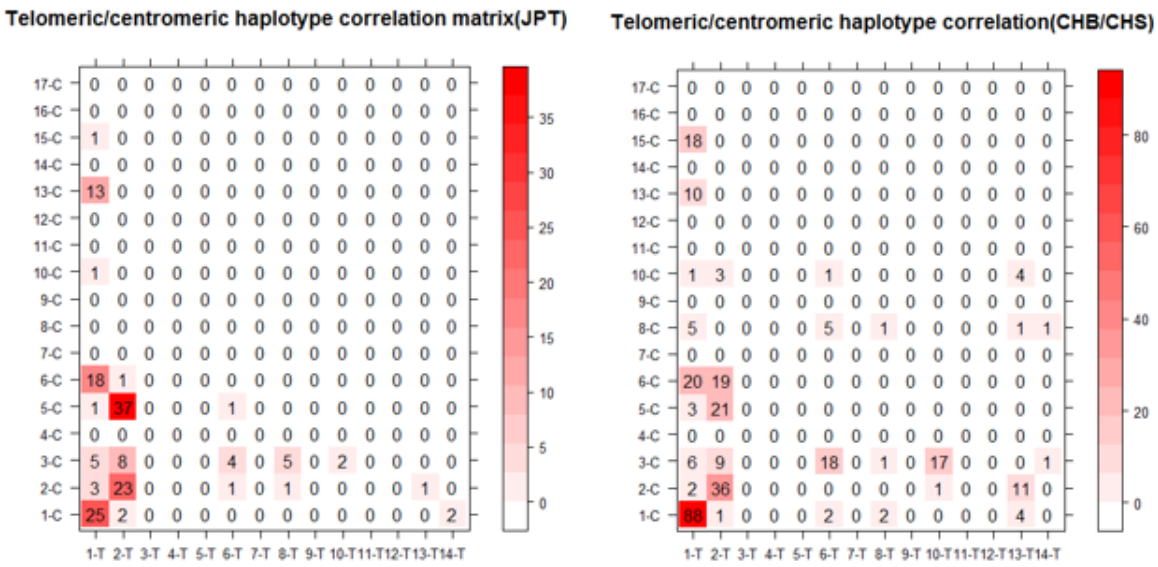


Figure 3.7 Flanking haplotype diversity between 1000 Genomes Project CHB and CHS (right) and JPT (left) populations. The most common haplotype is 1T/1C in Chinese, while in Japanese is 2T/5C.

3.4 New haplotype classes vs Black's haplotype classes

We compared our new haplotype classification with Black's haplotype classification in European (CEU+GBR+FIN), Asian (CHB+CHS+JP) and African (ASW+LWK+YRI) populations. In the centromeric region, the correlation is better than the telomeric region. This is because Black's haplotype is based on 4.1kb centromeric region only and recombination disrupted the association between centromeric and telomeric haplotypes, although the rate of recombination is not high enough to destroy the correlation totally. This suggested that the

DEFA1A3 CNV is a recombination warm spot. The correlation between old and new classification is better in European populations than Asian and African populations. This is caused by misclassifying new haplotypes in the two populations, especially for Class-1 and Exchange-2 haplotypes. Class-1 haplotypes in three super populations have very different components; in Europe it is mainly 4T/9C, in Asia it is 2T/5C and in Africa it is 3T/4C (table 3.3).

Table 3.3 Correlation between Black's haplotype classification and new haplotype (telomeric and centromeric haplotypes) classification in European, Asian, African populations.

European:

	EUR-T								
	1-T	10-T	14-T	2-T	4-T	5-T	6-T	8-T	rare
C1	6	0	2	15	37	0	0	0	3
C2	137	0	6	28	1	2	0	0	8
EX1	3	0	0	0	2	56	0	0	2
EX2	0	0	0	1	10	0	13	15	21
REF	12	8	0	108	28	1	2	7	43

	EUR-C									
	1-C	10-C	16-C	2-C	3-C	6-C	7-C	8-C	9-C	rare
C1	1	0	0	0	0	0	0	11	29	22
C2	124	0	18	0	0	0	0	11	0	29
EX1	0	0	0	0	0	0	53	0	0	10
EX2	6	0	0	0	13	0	0	6	0	35
REF	0	19	0	152	0	2	0	0	0	36

Asian:

ASN-T								
	1-T	10-T	13-T	14-T	2-T	6-T	8-T	rare
C1	3	0	0	0	65	1	0	2
C2	172	0	4	4	4	4	1	4
EX2	11	23	2	1	19	29	10	42
REF	77	1	17	0	98	2	1	26

ASN-C										
	1-C	10-C	13-C	15-C	2-C	3-C	5-C	6-C	8-C	rare
C1	1	0	0	0	0	0	63	0	0	7
C2	127	0	23	0	0	1	0	0	9	33
EX2	3	0	0	0	1	103	0	0	14	16
REF	0	12	0	19	99	0	0	61	0	31

African:

AFR-T													
	1-T	11-T	12-T	14-T	2-T	3-T	4-T	5-T	6-T	7-T	8-T	9-T	rare
C1	0	1	0	0	2	98	4	0	1	20	0	0	59
C2	15	5	26	4	2	14	0	0	0	21	0	14	16
EX1	0	0	0	0	0	0	0	4	0	0	0	0	0
EX2	1	26	2	4	31	4	0	0	2	2	1	19	65
REF	9	0	0	0	13	12	10	1	0	1	0	1	26

AFR-C												
	1-C	11-C	12-C	14-C	17-C	2-C	3-C	4-C	5-C	7-C	9-C	rare
C1	0	0	0	22	4	1	0	72	2	0	7	77
C2	83	0	0	0	8	0	0	0	0	0	0	26
EX1	0	0	0	0	0	0	0	0	0	3	0	1
EX2	5	30	25	0	0	0	35	0	0	0	0	62
REF	0	0	0	0	0	20	0	0	0	0	0	53

3.5 Tag SNPs for *DEFA1A3* CNV flanking haplotypes

Haplotypes without recent recombinations generate and accumulate unique variants, and these variants can be utilised as tags for certain

haplotypes. Tag variants are especially useful for experiment design. A simple genotyping assay of tag SNPs could identify entire haplotype classes. We searched for tag SNPs for each haplotype by a home-made R script (R_square.R). This script calculates the r^2 value between each SNP and each common haplotype. We did this calculation for all 1540 SNPs and 1738 haploid sequences. The result is plotted in fig 3.8 and fig 3.9. Detailed r squared values of the best tag SNPs for each haplotype class can be found in table 3.4 and table 3.5. The r^2 value is also useful to identify recombinant haplotypes as young recombinants have variants that come from parental haplotypes and do not generate their own unique tag SNPs.

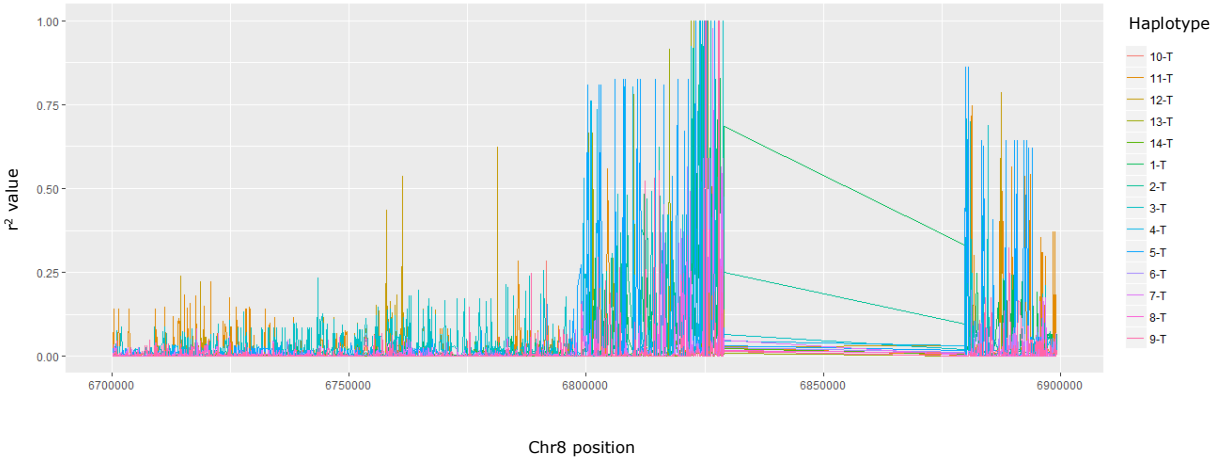


Figure 3.8 SNPs association with each telomeric haplotype.

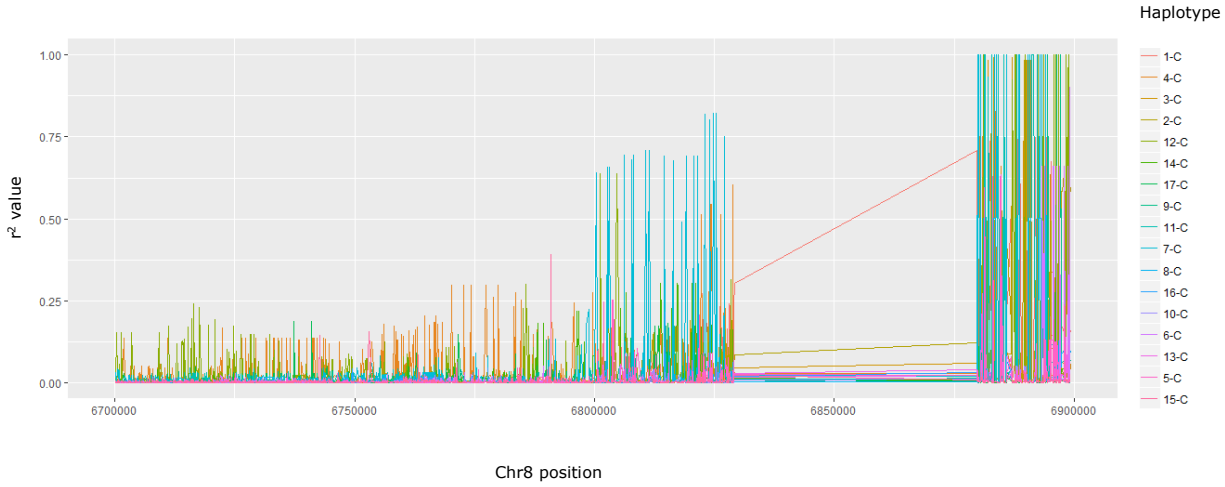


Figure 3.9 SNPs association with each centromeric haplotype.

Table 3.4 Best tag SNPs for each telomeric flanking haplotypes and their r squared value.

Tag	position	r squared
1T	6828241	0.827345
2T	6823891	0.92916
3T	6823933	1
4T	6807839	0.801745
5T	6823113	1
6T	6824631	0.591207
7T	6826717	0.976564
8T	6824631	0.377937
9T	6825168	1
10T	6791649	0.283294
11T	6881441	0.746593
12T	6822624	1
13T	6822139	1
14T	6806724	0.108735

Table 3.5 Best tag SNPs for each telomeric flanking haplotypes and their r squared value.

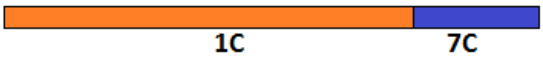
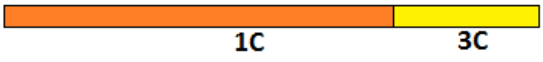
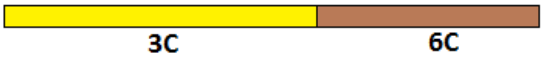
tag	position	r squared
1C	6888565	0.811199
2C	6897885	0.638736
3C	6892418	1
4C	6883604	0.827657
5C	6884828	0.634916
6C	6899070	0.902412
7C	6879879	1
8C	6896688	0.127656
9C	6894409	1
10C	6880963	0.05804
11C	6881209	1
12C	6881116	1
13C	6898386	0.137038
14C	6896115	0.541216
15C	6790836	0.391755
16C	6814637	0.214123
17C	6737382	0.185185

3.6 Recombinant haplotypes

Although we chose our CNV flanking haplotype analysis region by carefully avoiding the recombination hotspot, there is still a small number of recombinant haplotypes involved. Recombinant haplotypes tend to contain intervals similar to their parental haplotypes, and a clear recombination breakpoint between them can be seen for a young recombinant. This can be detected by calculating sequence similarity by a sliding window algorithm across whole sequences. We did recombination analysis by RAT software (chapter 2.2.4) and identified haplotypes 5C, 8C, 10C, 13C, 14C, 16C, 17C on the

centromeric side and 7T, 8T, 10T on the telomeric side as having clear evidence of recombination (table 3.6). Another line of evidence is that recombinant haplotypes usually have a low r^2 value close to 0, as they do not have haplotype-unique tag SNPs (table 3.4 and table 3.5). Removing recombinant haplotypes is important for phylogenetic analysis because recombinants usually do not have a fixed position on the phylogenetic tree. We conclude that there are 11 telomeric flanking haplotypes (1T,2T,3T,4T,5T,6T,9T,11T,12T,13T,14T) and 8 centromeric haplotypes (1C,2C,3C,4C,6C,7C,9C,11C,12C,15C) that are pure, non-recombinant haplotypes.

Table 3.6 Structure of recombinant haplotypes

Haplotype	Parental haplotypes	Diagram of recombination
centromeric		
5C	1C, 7C	
8C	2C, 3C	
10C	2C, 6C	
13C	1C, 3C	
14C	1C, 12C	
16C	1C, 9C	
17C	3C, 6C	
Telomeric		
7T	1T, 11T	
8T	1T, 14T	
10T	1T, 2T	

3.7 Phylogenetic analysis

In order to understand the relationships between identified non-recombinant haplotypes, we carried out a phylogenetic analysis. We constructed networks by median-joining methods for telomeric and centromeric haplotypes. Median-joining networks are established

under the assumption of absent recombination, so we removed haplotypes with strong evidence of recombination identified in the previous section. This analysis also includes three archaic human haplotypes: Neanderthal (Nea), Denisovan (Den) and Ust-ishim (UST) with data obtained from Max Planck Institute for Evolutionary Anthropology. We used chimpanzee as an outgroup to root this network, and this chimpanzee sequence is reference genome Pantro 4 downloaded from the UCSC genome browser. The phylogeny of centromeric haplotypes was shown in fig 3.10, and telomeric phylogeny shown in fig 3.11.

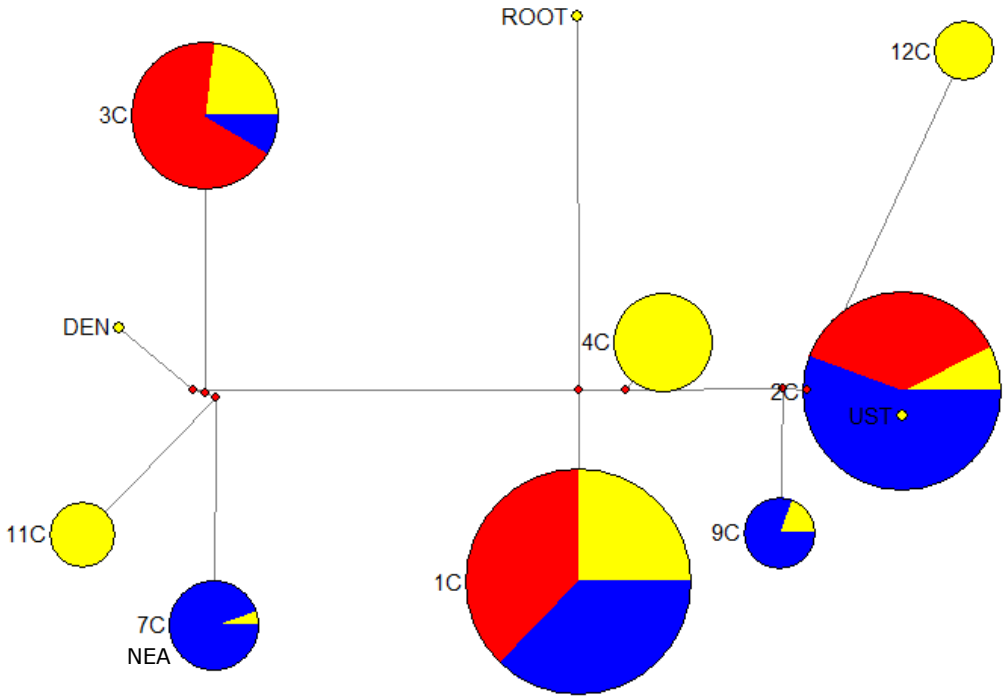


Figure 3.10 Median joining network of centromeric flanking haplotypes. Red colour represents Asian populations, blue colour represents European populations and yellow colour represents African populations.

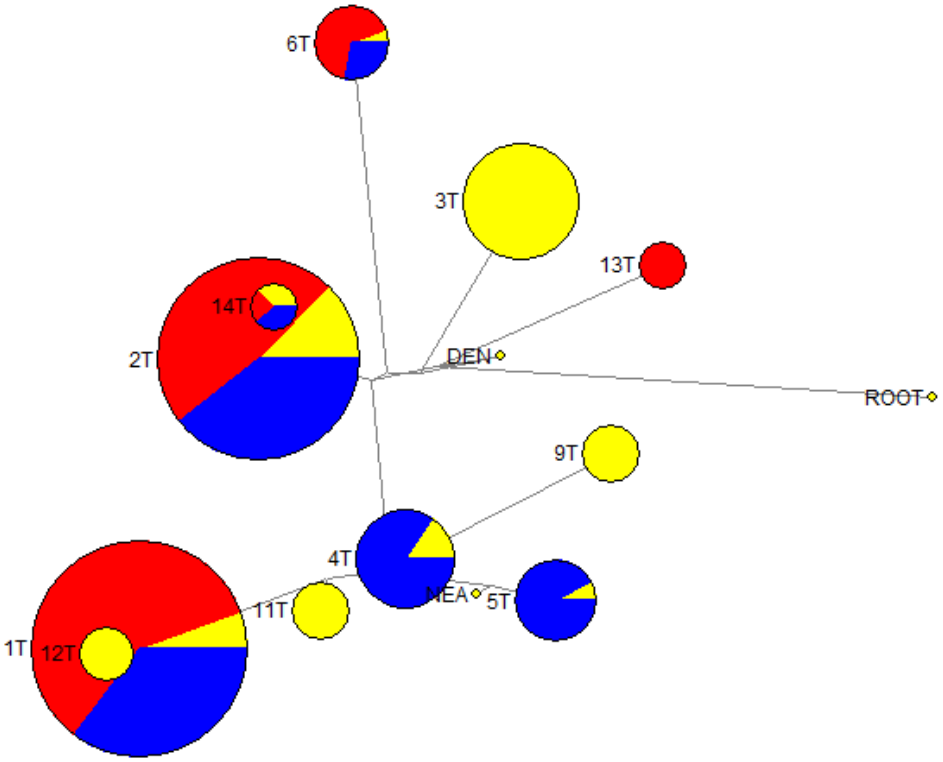


Figure 3.10 Median joining network of telomeric haplotypes. Red colour represents Asian populations, blue colour represents European populations and yellow colour represents African populations.

3.8 Experimental verification of recombinant haplotypes

The sequence data used from the 1000 Genomes Project is mostly based on statistically inferred phasing, and it is possible that some haplotypes were generated by phasing error or sequencing error. We tried not to draw conclusions about those rare and low-frequency haplotypes. High-frequency recombinants are unlikely to be caused

by phasing errors as homozygous individuals are present. There are some haplotypes without homozygous unambiguous phasing individuals, and we do need to make sure that the recombinant haplotypes from our analysis are real, especially for those in which predicted recombination breakpoints lie inside CNV as they might change CNV internal components. To address this question we thought the best approach was to utilise tag SNPs for each haplotype. We used PCR and Sanger sequencing to genotype individuals who were predicted to carry recombinant haplotypes. Then we used family segregation to phase telomeric and centromeric tag SNPs unambiguously. The experimental design is shown in chapter2.3.3; we PCR amplify a 2.6kb region on the telomeric side which contains 5 telomeric haplotype tag SNPs, and another PCR on the centromeric side to amplify a 2.4kb region containing 5 tag SNPs. Then we used two pairs of original PCR amplification primers and 4 nested sequencing primers to genotype these SNPs by Sanger sequencing. We did this analysis for 3 CEPH pedigrees: family 1375, 1408 and 1463.

In family 1375, sample NA12155 was included in the 1000 Genomes Project and identified to carry two haplotypes of 2T/1C and 5T/7C. 2T/1C was a recombinant haplotype. The genotype result of these family members is shown in table 3.7. The segregation result shows

that the two haplotypes of NA12155 are A-T and G-C, suggesting that the recombinant haplotype 2T/1C is real.

Table 3.7 Genotyping result of family 1375

Id		haplotype1	haplotype2	2T(6823891 A)	1C(6879885 T)
NA12155	father	**2T/1C**	5T/7C	A/G	C/T
NA10831	child	unknown	unknown	A/G	T/T
NA12156	Mother	unknown	unknown	G/G	T/T

In family 1408 NA12234 was included in the 1000 Genomes Project and identified to carry two haplotypes of 2T/2C and 1T/7C. Here, 1T/7C was a recombinant haplotype. The genotype result of family members is shown in table 3.8. The segregation result shows that the two haplotypes of NA12234 are T-G and C-C, confirming that the recombinant haplotype 1T/7C is real.

Table 3.8 Genotyping result of family 1408

id		haplotype1	haplotype2	1T (6822494 T)	7C(6879879 G)
NA12264	father	unknown	unknown	T/C	C/C
NA12234	Mother	2T/2C	**1T/7C**	T/C	G/C
NA10863	child	unknown	unknown	C/C	C/C

In family 1463 NA12889 was included in the 1000 Genomes Project and identified to carry two haplotypes of 2T/1C and 1T/7C. Both haplotypes were recombinant. The genotype result of these family members is shown in table 3.9. The 2T position cannot be phased,

due to all three samples being heterozygous at this position. Segregation result shows two haplotypes of NA12889 are A-C-T-C and G-T-C-G or G-C-T-C and A-T-C-G. This suggest that the recombinant haplotype 1T/7C is real but 2T/1C is not sure.

Table 3.9 Genotyping result of family 1463

id		haplotype1	haplotype2	2T(6823891 A)	1T(6822494 T)	1C(6879885 T)	7C(6879879 G)
NA12889	Father	**2T/1C**	**1T/7C**	A/G	T/C	T/C	C/G
NA12890	Mother	unknown	unknown	A/G	C/C	T/C	C/C
NA12877	Child	unknown	unknown	A/G	C/C	T/T	C/C

3.9 Conclusion

We carried out a comprehensive *DEFA1A3* flanking haplotype analysis by using the 1000 Genomes Project data and proposed a new haplotype naming system. Our analysis confirmed there are 5 major haplotype classes in the European population as described by Black et.al, but we disagree that the same haplotype classification can be used to explain Asian and African populations, especially for Class-1 and Exchange-2 haplotypes. New haplotypes were discovered in Africa and Asia. Samples with recombination between telomeric and centromeric flanking haplotypes are also commonly seen, indicating that the CNV is a recombination warm spot and the recombination breakpoints lying inside CNV may alter the CNV composition. CNV internal tag SNPs are also identified for some haplotypes. In all, the

new haplotype classification explained diversity in Asian and African population better, and improved our understanding of diversity and evolution in this locus.

4 Chapter4: *DEFA1A3* haplotypes associated with diseases

4.1 Introduction

Since 2005 genome-wide association studies (GWAS) have become a popular tool for studying genotype-phenotype association. This approach looks for genetic variants that are associated with a phenotype by using millions of SNP markers throughout the genome. GWAS opened a new gate towards understanding complex diseases in humans. Until 2013 there were more than 1600 publications that reported ~2000 associations for ~300 complex diseases and traits (Manolio, 2013). However, only a small proportion of them have been validated at a functional level. These GWAS hits require a large number of follow-up investigations. The human alpha-defensin locus is one of the loci implicated by GWAS studies. Since 2012 there have been several GWAS reported at this locus associated with multiple diseases (Yu et al., 2012, Matthias Munz, 2017, Kiryluk et al., 2014, Li et al., 2015). The causative variants are poorly understood, not only because there are multiple independent association signals near this locus but also because the copy number variation polymorphism added extra complexity. The real causative variant may hide inside

CNV and the GWAS signals may be just indirect reporters. In this chapter, we link the new *DEFA1A3* CNV flanking haplotypes with reported disease association signals to predict potential causative variants inside CNV.

4.2 Haplotypes associated with IgA Nephropathy

4.2.1 Association between IgA nephropathy and the alpha defensin locus

IgA nephropathy is the most common glomerulonephritis in the world. Epidemiology studies show disease prevalence is strikingly different in world populations; it is most common in the Asian populations and less frequent among Europeans and rare in Africans (Kiryluk et al., 2012). Recent genome-wide association studies and familial aggregation studies suggested there is a strong genetic factor in the prevalence of this disease. More than one study has reported the association of the alpha defensin locus with IgA nephropathy. So far there are six association studies implicating the alpha defensin locus in IgA nephropathy. Three of them are genome-wide association studies, and the other three association studies are based on genotyping a few local variants in or near *DEFA1A3* CNV (table 4.1

and 4.2). XueQing Yu et al. (2012) produced the first paper to describe the association and reported rs2738048 G ($p=3.18x10^{-14}$ OR=0.79) allele have a protective effect in a Chinese population. This finding was further confirmed by Kiryluk 2014 in a wider range study of world IgA patients, and he also pointed out that the association in rs2738048 is an Asian specific signal. Following these two GWAS, Ricong Xu et al. (2014) and YY Qi et al. (2015) conducted case-control studies by genotyping local variants in the *DEFA1A3* flanking region and replicated the GWAS result. They also discovered more than one causative factor within this region. Zhen Ai et al. (2016) for the first time studied the association between CNV internal variants and IgA nephropathy and reported that several internal variants have stronger association signal than flanking region variants (table 4.2); Indel 211bp (a 4bp deletion in intron 2 of *DEFA1*/*DEFA3*) have the smallest p value of disease association discovered so far (p value=3.5E-16). This suggested the real target of disease association may lay within the *DEFA1A3* CNV. Therefore it is important to understand disease association signals and *DEFA1A3* haplotypes. This may help us to understand and locate real causative variants inside *DEFA1A3* CNV.

Table 4.1 List p value and odds ratio of IgA Nephropathy associated SNPs reported by Recent 5 association studies of IgA Nephropathy near *DEFA1A3* CNV flanking region and best associated haplotypes.

		XueQing Yu(2012)		**Kiryluk(2012)**		**Ming Li(2015)**		**Ricong Xu(2014)**		**YY Qi(2015)**			
Population		Chinese		Asian, European		Chinese		Chinese		Chinese			
sample size		1434 cases, 4270 controls		7658 cases, 12954 controls		8313 cases, 19680 controls		1000 cases, 1000 controls		1194 cases, 902 controls			
variants genotyped		GWAS		GWAS		GWAS		16SNPs		60 SNPs			
rs number	Position	p value	odds ratio	p value	odds ratio	p value	Odds ratio	p value	odds ratio	p value	odds ratio	best haplotype	Max r2
rs9314614	6697731					4.25E-09	1.129					NA	NA
rs712276	6782271							8.30E-01	0.97			NA	0.04
rs2738118	6784971							6.60E-01	1.04			NA	0.02
rs3758139	6786220							9.80E-01	1			NA	0.14
rs2738116	6786417							1.70E-01	0.85			NA	0.03
rs12545378	6786557							1.00E-02	0.73			NA	0.03
rs736227	6793540							1.10E-01	1.12			NA	0.03
rs2738100	6793581							5.80E-01	1.04			NA	0.04
rs2615787	6811149							1.00E-04	0.74			2T	0.68
rs2741658	6811559									1.70E-03	0.8	2T	0.84
rs2702852	6813501									4.29E-03	0.81	2T	0.28
rs2738081	6815518							3.00E-04	0.72			2T	0.43
rs12680018	6820464									2.06E-02	1.22	6T	0.69
rs2738058	6821617					1.15E-19	1.232	1.00E-04	0.73	4.64E-05	0.76	2T	0.71
rs2738048	6822785	3.18E-14	0.79	1.60E-04	1.1			7.00E-04	0.77	2.13E-03	0.81	2T	0.96
rs2702910	6823891									1.48E-03	0.8	2T	0.96
rs7825750	6879803									1.95E-02	1.16	1C	0.68
rs4300027	6880575									5.37E-04	0.8	2C	0.36
rs4512398	6880594									6.65E-04	0.8	2C	0.37
rs7826487	6880925									4.45E-03	0.69	5C	1.00
rs4288398	6881638							8.00E-04	0.78	1.19E-02	0.85	2C	0.35
rs4313182	6881988							1.20E-01	1.15			3C	0.98
rs9774483	6883291									4.23E-02	1.18	3C	0.98
rs6984215	6884878							3.00E-03	0.81			2C	0.27
rs6982814	6887956									4.74E-02	1.17	3C	0.98
rs12716641	6898998					9.53E-09	1.154	2.00E-05	0.71			2C	0.83
rs10086568	6900336			1.00E-09	1.16							1C	NA
rs4304345	6901619							1.10E-01	0.82				NA
rs12682030	6903607									1.17E-02	0.76		NA
rs6999181	6904509									1.04E-02	0.75		NA
rs9644778	6909654									4.78E-03	1.21		NA
rs10095331	6913850									9.45E-03	0.75		NA

Table 4.2 Zhen Ai et al. (2016) association study of IgA Nephropathy and *DEFA1A3* CNV internal variants. Showing association of CNV internal variants with IgA Nephropathy (Data from Zhen Ai et al. 2016)

CNV internal variants	p value	OR
DEFA1A3 total copy number	3.99E-09	0.88
DEFA1 copy number	6.71E-05	0.92
DEFA3 copy number	6.55E-05	0.82
211bp copy number	3.50E-16	0.75
124bp copy number	2.89E-08	0.87
275bp copy number	2.51E-07	0.89
129bp copy number	1.37E-02	0.88

4.2.2 Haplotype association with IgA Nephropathy

In the previous chapter, we have discussed *DEFA1A3* CNV flanking haplotype diversity in world populations. To understand and dissect GWAS signals discovered in this locus, we studied reported IgA nephropathy association signals and their relationships with flanking haplotype classes.

rs2738048 and rs2738058 are the strongest disease-associated SNPs in the telomeric flanking region, and they are in a good LD with each other. Both of them are good tag SNPs for 2T haplotype in Asian populations, with r^2 value 0.96 and 0.71 respectively. Another strong GWAS association signal on the centromeric side is rs12716641; this is more significant than rs2738048 according to Ming Li et al. (2015), and this SNP most frequently appears on 2C haplotype background

(r^2 = 0.82). From Ricong Xu et al. (2014) and YY Qi et al. (2015)'s analysis, we can see SNPs which are more associated with 2T haplotype on the telomeric side or 2C on the centromeric side, and they are more strongly associated with the IgA Nephropathy protective effect.

In Asia, 2T haplotype mainly exists in the form of 2T/2C, 2T/3C, 2T/5C and 2T/6C. 2C haplotype mainly exists in 2T/2C and 13T/2C forms. According to Ricong Xu et al. (2014) and YY Qi et al. (2015), 2C, 3C, 5C tag SNPs were included in their association study. 2C and 5C tag SNPs were reported as associated with IgA Nephropathy. 3C tag SNP rs4313182 (r^2=0.98) has a non-significant p-value and odds ratio (p=0.12, OR=1.15) which indicates that the 3C haplotype is unlikely to be relevant to the disease risk. 2C haplotype tag SNP rs12716641 (r^2=0.83) has the smallest p-value and a low odds ratio (p= 0.00002 OR=0.71). 5C haplotype tag SNP rs7826487 (r^2=1) has a moderate p-value but has the smallest odds ratio (p=0.00445 OR=0.69). According to Ming Li's (2015) analysis of three IgA Nephropathy GWAS SNP haplotypes (table 4.3), if we condition on rs9314614 which is a completely independent factor not in LD with *DEFA1A3* CNV, the rs2738058-rs12716641 CC combination (2T/2C haplotype) has the strongest association, followed by CT (2T/non-2C), then TC (non-2T/2C). This analysis shows there is an additive protective effect for the two components of the 2T/2C haplotype, as

the pure lineage has a stronger association than recombinant haplotypes. Zhen Ai et al. (2016) showed variants inside the *DEFA1A3* CNV have much stronger association signals than the flanking region, which means the real target of disease association may be within the CNV. If there is a single causative factor inside the CNV then a variant present in 2T/2C haplotypes could explain this association most simply. Pure 2T/2C haplotypes would carry more such variants and show a stronger protective effect than recombinant 2T or 2C haplotypes.

Table 4.3 Haplotype association of three IgA Nephropathy SNPs reported by Ming Li 2015 showing additive effect. (Data from Ming Li 2015)

Haplotype rs9314614-rs2738058-rs12716641	# risk alleles	P	Average frequency (controls)	OR (each vs all)	Phet	OR with CTT as reference
CTT	3	4.71E-22	24.6%	1.257	0.365	1.000
CTC	2	7.88E-01	4.0%	0.987	0.914	0.850
CCT	2	1.95E-01	3.7%	0.934	0.284	0.792
GTT	2	1.34E-03	34.2%	1.068	0.738	0.896
CCC	1	4.20E-05	3.9%	0.809	0.689	0.709
GTC	1	2.31E-02	4.5%	0.898	0.004	0.767
GCT	1	9.74E-08	13.4%	0.847	0.060	0.744
GCC	0	3.51E-13	11.5%	0.777	0.643	0.688

Kiryluk et al. (2012) carried out a GWAS analysis on IgA nephropathy in European and Asian populations. They also reported an association with rs2738058, and further discovered this is an Asian specific association. Our haplotype diversity analysis shows the 2T/2C and

related recombinant haplotype is commonly distributed worldwide. The reason 2T/2C tag SNP does not show association in European samples may be explained either by population-specific variants inside Asian 2T/2C haplotypes, or by Asian populations having some other local risk factors that put them at risk of IgA Nephropathy, and European populations lack of higher background risk factor the 2T/2C haplotype is protective of. We must keep in mind there are many other IgA nephropathy association signals discovered in the genome. Until the discovery of the *DEFA1A3* association, the strongest associations were reported in the HLA region. Kiryluk et al. (2012) also reported that rs10086568 is another SNP associated with IgA nephropathy ($p=1.00 \times 10^{-9}$ OR=1.16). The A allele is a good tag SNP for 1C haplotype and odds ratio bigger than 1 suggests that the 1C haplotype is an IgA nephropathy risk factor. Unfortunately, there is no 1T haplotype tag SNP included in all association studies, and therefore we cannot make a further inference.

Ai Zhen et al. (2016) studied association inside *DEFA1A3* CNV, and found that CNV internal variants have a stronger association than previously reported flanking variants (table 4.2). High *DEFA1A3* total copy number, high DEFA3 number, and high 211bp number are significantly associated with protection from IgA nephropathy. This result was also consistent with our 2T/2C copy additive effect hypothesis. 2T/2C and its recombinants have higher copy number

than other haplotypes, and especially the 2T/5C haplotype (specific to Asian populations) has an average copy number of about 7 which is the highest among all haplotypes. 2T/2C haplotypes also have a higher average *DEFA3* and 211 copy number than other haplotypes.

4.3 Haplotypes associated with periodontitis

4.3.1 Periodontitis

Periodontitis is an inflammatory disease with loss of connective tissue and alveolar bone and if left untreated, it may eventually lead to tooth loss (Pihlstrom et al., 2005). It has unknown aetiology, but is assumed to be caused by interaction between oral microbiota and the host immune system. Although environmental factors such as diet and smoking may play an important role in the pathogenesis of the disease, there is also evidence of genetic factors involved. GWAS also detected multiple disease association signals, but until recently none of them passed genome-wide significance levels (Teumer et al., 2013).

4.3.2 GWAS of Periodontitis

Recently GWAS studies of chronic periodontitis also showed an association of the *DEFA1A3* locus with the disease. In 2013 a GWAS was conducted for chronic periodontitis in 4032 German individuals,

although none of SNPs tested reached genome-wide significance level. This study reported that rs2738058-T ($p=6 \times 10^{-6}$, OR=1.29) is a risk factor for the disease and it is among the top list of association signals (Teumer et al., 2013). In 2016 another GWAS based on 9629 cases plus control samples with chronic and aggressive periodontitis in European populations reported two associations in the region, at rs2738058 ($P=1.04 \times 10^{-9}$, OR=1.28) and rs2978951 ($p=2.22 \times 10^{-6}$ OR=0.7559). These results indicate that alpha-defensins may participate in the pathology of periodontitis. Alpha-defensin peptides 1-4 are primarily stored in neutrophil azurophilic granules. Neutrophils are abundant in gingival tissues and they participate in innate immune defenses. It is reported alpha-defensin 1-3 in saliva and gingival fluid can be 15-fold up-regulated in cases of aggressive periodontitis and 60-fold upregulated in chronic periodontitis (Puklo et al., 2008).

4.3.3 1T/1C haplotype associated with periodontitis risk effect

To understand the relationship between *DEFA1A3* haplotypes and periodontitis, we compared SNP-haplotype association and SNP-disease association. SNP-disease association data for periodontitis were obtained from Arne Schafer's GWAS study. This included 2287 SNPs in a 200kb (chr8:6700000-chr8:6900000) region near *DEFA1A3*

CNV. The SNP-haplotype association was obtained by calculating r^2 values for each SNP and each haplotype class. Then we linked SNP-haplotype association and SNP-disease association by calculating Pearson's correlation coefficient across all typed SNPs (table 4.4). Our result shows the best haplotype associated with periodontitis is 1T haplotype with correlation coefficient 0.74 (fig 4.1). This suggests that SNPs which are more specifically present on 1T haplotypes are also more associated with periodontitis. A similar effect also can be seen for 1C haplotype since 1T/1C are in relatively good LD in the European population. 5T/7C is a Neanderthal introgressed haplotype (discussed in chapter 5), but it has a poor correlation with periodontitis; the value is close to 0 (fig 4.2), indicating the Neanderthal introgressed haplotype does not associate with periodontitis.

Table 4.4 Pearson's correlation coefficient between telomeric/centromeric haplotypes and $-\log_{10}$ p-value of periodontitis association across all SNPs typed, showing 1T is the best telomeric haplotype correlated with periodontitis. 1C is the best centromeric haplotype correlated with periodontitis.

Telomeric haplotype	4T	1T	2T	8T	5T	10T	6T	14T	16T
Pearson's correlation coefficient	0.210	0.744	0.545	-0.009	-0.028	0.529	0.013	0.143	-0.054

Centromeric haplotype	1C	2C	3C	6C	7C	8C	9C	10C
Pearson's correlation coefficient	0.504	0.435	-0.023	-0.052	-0.054	0.302	0.002	0.320

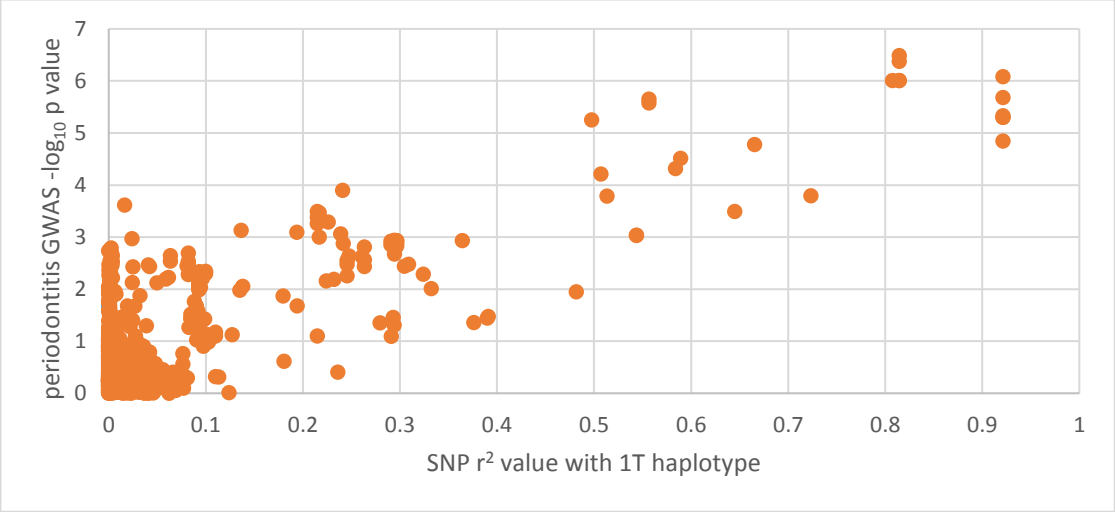


Figure 4.1 Plot showing SNP r^2 with 1T haplotype and $-\log_{10}$ P value for association with periodontitis have a positive correlation (Pearson's correlation coefficient=0.744); X axis: SNP r^2 value with 1T haplotype. Y axis: $-\log_{10}$ p-value of periodontitis GWAS association.

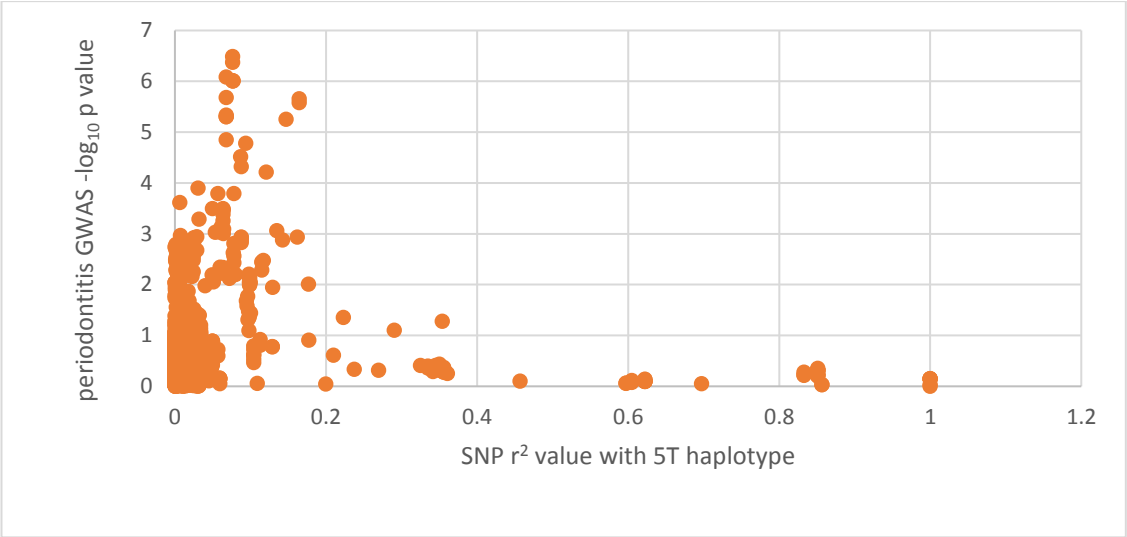


Figure 4.2 Plot showing 5T haplotype and periodontitis have no correlation (Pearson's correlation coefficient=-0.028), X axis: SNP r^2 value with 5T haplotype. Y axis: -log10 p-value of periodontitis GWAS association.

4.4 Looking for causative variants inside *DEFA1A3* CNV

4.4.1 Detecting sequence variants inside *DEFA1A3* CNV

Since IgA nephropathy and periodontitis association factors may be inside the CNV repeat unit, it is useful to have a full list of CNV internal variants. We called CNV internal variants from the 1000 Genomes Project data by using a multiple copy collapse down method (described in chapter 2.2.1). Here we realigned sequences in the *DEFA1A3* CNV onto a single reference sequence copy, then measured copy number for each sample by read depth method and treated each

copy number group as a polyploid, finally calling variants by GATK according to their ploidy number. In total, we discovered 1774 variants from 1047 samples in 9 populations (CHB, CHS, JPT, CEU, GBR, FIN, YRI, ASW, LWK). 1704 variants passed GATK phred quality score of 20 which the error rate <1%. 24 of them are located in coding regions (table 4.5). Among them, there are 8 synonymous mutations, 15 missense mutations, and 1 nonsense mutation. Variants detected here are present in a multiple copy variable region and also referred as Multi Site Variants (MSV). Among the 15 missense mutations, one MSV (at position 1625) is the well-known *DEFA3* MSV. One Asian population specific missense mutation (Asp-His) particularly draws our attention. We named this missense mutation "2284" as it is located at the coordinate 2284 of the one-copy reference sequence we created. GATK picked up in total 77 "2284" carrier samples, most of them are Asians.

Table 4.5 list of variations detected within *DEFA1A3* CNV region from the 1000 Genome Project realigned samples.

position	region	reference allele	alternative allele	allele frequency	type of mutation	amino acid change
exon3 (1414-1643)						
1425	3'-UTR	G	A	9.17E-04	-	-
1436	3'-UTR	G	A	0.247	-	-
1459	3'-UTR	G	C	3.67E-04	-	-
1468	3'-UTR	G	T	0.058	-	-
1470	3'-UTR	A	T	5.50E-04	-	-
1472	3'-UTR	T	C	0.032	-	-
1520	3'-UTR	C	G	1.83E-04	-	-
1583	Exon3 coding	C	T	3.67E-04	missense	R>H
1595	Exon3 coding	G	A	1.83E-04	missense	A>V
1603	Exon3 coding	C	T	3.12E-03	synonymous	A
1604	Exon3 coding	G	A	7.34E-04	missense	A>V
1625	Exon3 coding	G	T	0.206	missense	D>A
1630	Exon3 coding	G	A	1.83E-04	synonymous	N
1643	Exon3 coding	C	T	9.17E-04	missense	G>D
exon2 (2225-2401)						
2231	exon2 coding	C	G	1.10E-03	missense	K>N
2240	exon2 coding	C	T	1.83E-04	synonymous	L
2248	exon2 coding	C	T	5.50E-04	missense	E>Q/E>K
2249	exon2 coding	G	A	3.48E-03	synonymous	D
2270	exon2 coding	C	G	0.089	synonymous	V
2275	exon2 coding	C	G	3.67E-04	missense	E>Q
2284	exon2 coding	C	G	0.014	missense	D>H
2300	exon2 coding	C	T	6.42E-03	synonymous	P
2317	exon2 coding	C	T	1.84E-04	missense	E>K
2326	exon2 coding	T	C	1.84E-04	missense	R>G

2347	exon2 coding	G	T	3.68E-04	missense	Q>K
2351	exon2 coding	C	T	1.10E-03	synonymous	Q
2353	exon2 coding	G	A	1.65E-03	Nonsense	Q>stop
2371	exon2 coding	T	G	0.018	missense	I>L
2387	exon2 coding	G	A	4.59E-03	synonymous	L
2397	exon2 coding	A	G	1.83E-04	missense	M>T
2401	exon2 coding	C	G	9.18E-04	missense	A>P
exon1 (3771-3850)						
3774	exon1 non-coding	T	A	3.67E-04	-	-
3790	exon1 non-coding	A	G	1.83E-04	-	-
3810	exon1 non-coding	G	C	3.67E-04	-	-
3825	exon1 non-coding	C	T	5.50E-04	-	-
3833	exon1 non-coding	G	A	3.67E-04	-	-
promotor (3841-3901)						
3845	promotor	C	G	0.019	-	-

4.4.2 Haplotype-specific CNV internal variants.

After we detected all variants present inside *DEAF1A3* CNV, another question began to emerge: Are these variant present on specific haplotype classes, or they can be freely exchanged between different haplotype classes? What variants are only present on 2T or 2C haplotypes? To answer these questions we carried out a Cochran-Armitage trend test between all CNV internal variants and all flanking haplotype classes (chapter 2.2.8). We set the p-value cut off at 0.05. In total, we discovered 449 significant variants that associated with

at least one telomeric haplotypes. 453 variants significantly associated with at least one centromeric haplotypes. This result suggests there is a barrier of inter-homologue transfer of variants inside the *DEFA1A3* CNV. This means that exchange of genetic material between copies of different haplotypes is not common, and that a variant tends to stay within the same haplotype class. We also generated a list of 63 CNV internal variants that are tightly associated with 2T and 2C haplotype. If IgA nephropathy causative variants are located within *DEFA1A3* CNV, then this list provided good candidates for further analysis.

Table 4.5 list of *DEFA1A3* CNV internal variants that are associated with 2T and 2C haplotype.

MSV position	Allele frequency	P value(2-T haplotype)	P value(2-C haplotype
231	0.09	5.99E-13	1.72E-34
478	0.449	1.42E-09	1.62E-07
594	0.21	2.23E-17	6.50E-09
638	0.23	2.24E-09	0.000598
886	0.179	7.73E-05	0.006637
1036	0.15	2.45E-43	9.65E-19
1148	0.069	3.54E-14	1.85E-61
1197	0.087	0.00018	8.57E-32
1887	0.217	2.13E-09	0.086888
1920	0.086	2.31E-12	3.11E-11
2270	0.089	3.11E-51	7.33E-29
2539	0.431	0.004543	0.000116
2703	0.338	2.29E-05	3.58E-06
3608	0.046	3.80E-20	3.67E-58
3918	0.056	2.16E-21	1.47E-40
4223	0.509	1.39E-06	4.07E-07
4293	0.582	0.000233	9.99E-07
4294	0.58	0.000233	9.99E-07
4345	0.206	2.20E-42	2.55E-32
5524	0.139	1.04E-22	5.41E-15
5738	0.334	0.00011	6.92E-05
6787	0.212	1.26E-29	4.08E-41
6975	0.026	0.006298	0.000699
7020	0.475	8.05E-05	7.18E-05
7084	0.192	1.34E-12	1.71E-06
7351	0.393	7.83E-21	4.45E-18
8681	0.171	1.71E-21	1.21E-15
8774	0.265	1.67E-08	1.86E-07
9575	0.257	1.61E-13	1.22E-05
10284	0.278	4.38E-16	1.31E-08
10319	0.125	6.51E-44	3.50E-22
10596	0.106	8.31E-61	1.98E-32
10656	0.239	5.35E-12	1.80E-05
11384	0.224	3.63E-12	1.49E-06
11445	0.281	2.49E-05	4.91E-05

11870	0.24	1.03E-26	3.45E-23
12361	0.109	2.37E-30	1.74E-31
12891	0.158	8.83E-08	5.02E-18
13634	0.196	2.72E-41	4.66E-30
13711	0.155	2.42E-07	7.00E-19
14066	0.489	0.000101	0.000629
14235	0.222	7.42E-24	4.78E-27
14239	0.225	4.13E-24	2.78E-26
14325	0.113	2.14E-24	1.38E-41
14577	0.165	9.47E-06	6.53E-12
14588	0.109	6.54E-39	4.06E-27
14736	0.444	0.000835	0.002531
15386	0.276	2.87E-07	7.30E-07
15404	0.09	1.05E-58	2.67E-35
15856	0.117	9.85E-17	3.42E-30
16132	0.098	7.57E-53	8.43E-31
16352	0.117	7.76E-38	1.06E-09
16366	0.195	6.33E-08	0.001004
16557	0.079	1.65E-17	8.32E-47
16568	0.177	4.42E-12	1.30E-06
16605	0.104	0.003049	1.63E-11
16639	0.085	3.48E-62	4.25E-28
16693	0.085	4.38E-76	5.28E-21
16936	0.11	3.07E-06	9.14E-14
16942	0.081	7.17E-52	2.10E-29
17163	0.067	2.73E-15	0.003173
17219	0.084	1.04E-54	8.19E-26
18575	0.081	2.51E-17	8.57E-55

4.4.3 Haplotype specific tag MSVs inside the CNV

For non-recombinant haplotypes which have two consistent flanking haplotypes in both sides of the CNV, we can predict they should also have a consistent CNV component if the gene conversion events between different haplotypes are not frequent enough to destroy their consistency. As well as the overall arrangement of variant

repeats, we may also be able to identify haplotype-specific tag SNPs for repeat units inside the CNV. We searched for unique haplotype tags by calculating the r^2 value between internal SNPs and their flanking haplotype classes; this method is similar to the one used for the CNV flanking region, but we can only analyse individuals who carry homozygous haplotypes since phasing inside the CNV region is not possible. In this analysis, we used 38 samples with homozygous non-recombinant flanking haplotypes. We phased their CNV internal variants randomly according to their copy number and labelled each copy's flanking haplotype classes and pooled them together. This resulted in the generation of 180 *DEFA1A3* copies in total. Finally, we can use the r^2 value to study the association between each CNV MSVs and each flanking haplotype. The result is shown in fig 4.3 and the best CNV tag MSVs for each haplotype are displayed in table 4.6. The Denisovan sequence obtained from (Meyer et al., 2012), and found to carry two identical haplotypes (see 2.1.2). We discovered there are good CNV tag MSVs for 5T/7C, 1T/1C, 3T/4C and Denisovan haplotypes (r^2 value>0.8). Gene conversion event can transfer small segments of sequence from one haplotype to another without disrupt flanking sequences, but a lot of MSVs were unique to certain haplotypes were discovered. This indicated that a gene conversion event between different haplotypes is not common, but SNPs that

first occurred in one copy could be easily transferred to other copies of the same haplotype.

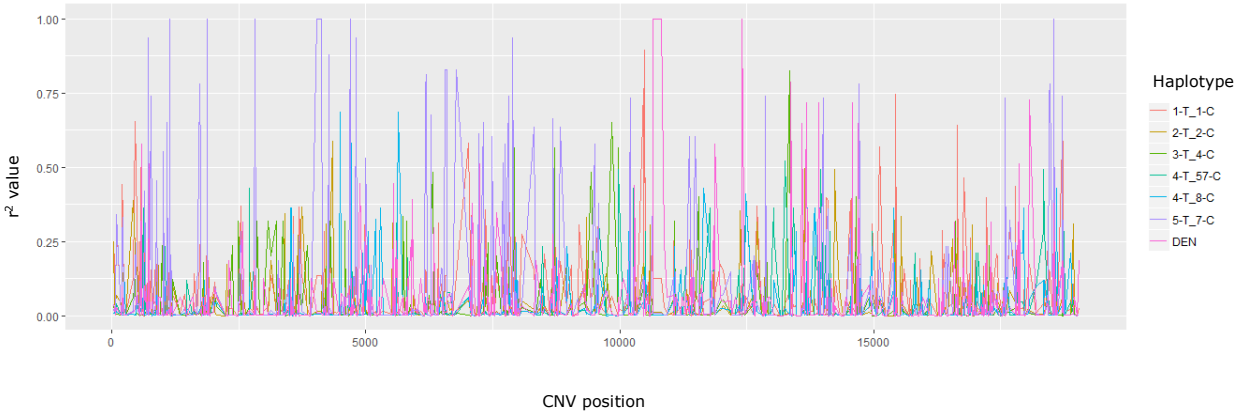


Figure 4.3 *DEFA1A3* CNV internal MSV association with telomeric/centromeric flanking haplotypes

Table 4.6 Best tag MSVs inside the *DEFA1A3* CNV for each flanking haplotype.

haplotype	position	r squared
1-T_1-C	10479	0.893162
2-T_2-C	4345	0.588336
3-T_4-C	13341	0.823529
4-T_8-C	4495	0.684899
5-T_7-C	1145	1
DEN	10661	1

4.4.4 Verification of MSV "2284"

To verify our CNV internal variant analysis results, a PCR-Restriction fragment length polymorphism (PCR-RFLP) experiment was designed (chapter 2.3.5). We applied this experiment to 96 HapMap CHB/JPT samples. In total, we detected 22 carriers and 73 non-carriers of the

2284 variant. All experimental results are consistent with our CNV internal variants analysis result using the 1000 Genomes Project data. We further verified the 2284 variant by Sanger sequencing for 12 carrier samples and 5 control samples from HapMap JPT/CHB samples. Sequencing results show that all samples are consistent with NGS data.

4.4.5 Measuring MSV "2284" copy number

The accuracy of capillary electrophoresis allows quantitative analysis of the PCR-RFLP experiment result. The number of *DEFA1A3* copies carrying MSV 2284 G can be calculated according to 135bp and 188bp peak area ratio. 20 MSV2284 G allele carrier samples from HapMap project JPT and CHB populations were used in this analysis. The results are shown in table 4.7. Our analysis shows most MSV2284 carrier samples have at least one 2T/5C haplotype (70%), and all samples have at least one 2T telomeric haplotype, indicating that MSV 2284 is most strongly associated with the 2T haplotype. All 2T/5C homozygous samples carry 2 copies of MSV 2284 G; most heterozygous or recombinant individuals carry 1 MSV2284. However, there are 3 exceptions carrying 2 copies of MSV2284 G suggesting that the MSV 2284 can be transferred to other copies of the same haplotype.

Table 4.7 list of MSV 2284 carrier and their haplotype classes; diploid copy number and MSV 2284 copy number

samples name	population	haplotype1	haplotype 2	total CN	CN carrying SNP2284G
NA18636	CHB	2-T/5-C	1-T/1-C	8.8	1.1
NA18555	CHB	16-T/2-C	2-T/5-C	9.3	0.9
NA18942	JPT	2-T/5-C	2-T/35-C	12.2	1.1
NA18961	JPT	1-T/15-C	2-T/5-C	10.7	1.0
NA18998	JPT	2-T/130-C	1-T/6-C	11.0	0.7
NA18976	JPT	2-T/5-C	1-T/1-C	10.7	0.8
NA18968	JPT	2-T/5-C	2-T/5-C	14.5	1.7
NA18978	JPT	1-T/135-C	2-T/5-C	10.7	1.0
NA18974	JPT	2-T/3-C	2-T/2-C	10.5	0.8
NA18943	JPT	2-T/5-C	8-T/44-C	10.8	1.0
NA18959	JPT	1-T/1-C	2-T/5-C	10.2	0.9
NA18973	JPT	2-T/2-C	2-T/3-C	8.8	2.1
NA18952	JPT	2-T/35-C	2-T/5-C	13.1	0.9
NA18980	JPT	2-T/5-C	2-T/5-C	15.1	2.1
NA18990	JPT	2-T/2-C	2-T/5-C	12.8	1.0
NA19007	JPT	1-T/1-C	2-T/2-C	7.8	2.0
NA18956	JPT	2-T/5-C	44-T/8-C	11.0	1.0
NA18960	JPT	2-T/5-C	2-T/5-C	13.9	1.9
NA19012	JPT	44-T/8-C	2-T/2-C	7.1	1.8
NA19003	JPT	2-T/3-C	1-T/3-C	11.6	0.9

4.4.6 Distribution of MSV 2284 in the *DEFA1A3* CNV repeat array.

In order to understand the distribution of MSV2284 in the *DEFA1A3* CNV repeat array, we carried out two emulsion fusion PCRs by fusing a CNV internal region which includes MSV2284 with telomeric or centromeric flanking region amplicons (see chapter 2). According to peak height of nucleotide types in the MSV2284 position, we could tell the allele type in first two copies counting from the telomeric or centromeric side. Our experiment result shows the position of MSV 2284 is not fixed and could duplicated/transferred to different

locations in the repeat array (table 4.8), although the most commonly seen position is the second copy counted from the telomeric side, but it can also jump to the first position of telomeric side and first or second copy from the centromeric side. Even for three 2T/5C homozygous samples, the position of MSV2284 is different. This result suggests that MSV 2284 is more easily transferred to different copies of same haplotype classes than to copies of different haplotypes classes.

Table 4.8 Result from Emulsion fusion PCR experiment. Suggesting that the MSV 2284 position is not fixed.

samples ID	population	Haplotype1	Haplotype2	Diploid DEFA1A3 CN	MSV 2284 CN	telomeric EHF-PCR	centromeric EHF-PCR
NA18968	JPT	2T/5C	2T/5C	14.5	1.7	2nd copy	none
NA18980	JPT	2T/5C	2T/5C	15.1	2.1	1st copy	2nd copy
NA18960	JPT	2T/5C	2T/5C	13.9	1.9	2nd copy	none
NA18973	JPT	2T/2C	2T/3C	8.8	2.1	1st copy	none
NA19007	JPT	1T/1C	2T/2C	7.8	2.0	1st copy	2nd copy
NA19012	JPT	2T/2C	44T/8C	7.1	1.8	2nd copy	1st copy
NA18974	JPT	2T/2C	2T/3C	10.5	0.8	none	none
NA19003	JPT	1T/3C	2T/3C	11.6	0.9	2nd copy	none

4.4.7 Experimental verification of CNV internal tag MSV 10479 for 1T/1C haplotype copies

In chapter 4.4.3 we analysed tag SNPs inside the *DEFA1A3* CNV, and discovered that there is a CNV tag MSV 10479 "A" allele that is present in nearly every copy of 1T/1C haplotype but not present in

any other haplotype ($r^2 = 0.89$, table 4.6). To verify this finding, we designed a simple experiment to Sanger sequencing MSV 10479 (chapter 2.3.4). We did this experiment for 20 HapMap samples who were included in the 1000 Genomes Project. In samples who carry homozygous 1T/1C haplotypes, their MSV 10479 position should have "A" allele in all copies if it is perfectly correlated with repeats found on a 1T/1C haplotype. For individuals with heterozygous 1T/1C haplotypes, some copies should have an "A" allele and some should have a "G" allele. For samples who do not carry 1T/1C copies, all copies should have a "G" allele. From our experimental result we can see (table 4.9), for almost all samples sequenced in MSV 10479 position, the genotype is same to our prediction (except NA12718 and NA07000). Sample NA12718 is a 1T/1C homozygous, but "G" allele can be seen on some copies. Sample NA07000 carries recombinant haplotype, but all copies have "A" allele; this might be determined by the location of recombination breakpoint. In all, our sequencing results prove the 1000 genome data analysis is reliable and MSV 10479 is a good CNV internal tag MSV for 1T/1C copies.

Table 4.9 experiment verification of 1T/1C haplotype tag MSV 10479.

ID	Classes	haplotype-1	haplotype-2	Predicted genotype	Sequencing genotype
NA18980	Non 1T/1C	2T/5C	2T/5C	G/G	G/G
NA18948	1T/1C heterozygous	1T/1C	101T/3C	A/G	A/G
NA18552	1T/1C heterozygous	1T/1C	6T/3C	A/G	A/G
NA19206	Non 1T/1C	3T/4C	3T/4C	G/G	G/G
NA18547	Non 1T/1C	2T/6C	2T/2C	G/G	G/G
NA11993	Non 1T/1C	5T/7C	5T/7C	G/G	G/G
NA12155	1T/1C recombinant	2T/1C	5T/7C	A/G	A/G
NA12234	1T/1C recombinant	2T/2C	1T/7C	A/G	A/G
NA12889	1T/1C recombinant	2T/1C	1T/7C	A/G	A/G
NA18632	1T/1C homozygous	1T/1C	1T/1C	A/A	A/A
NA18975	1T/1C homozygous	1T/1C	1T/1C	A/A	A/A
NA07056	1T/1C homozygous	1T/1C	1T/1C	A/A	A/A
NA12751	1T/1C heterozygous	1T/1C	4T/32C	A/G	A/G
NA11843	1T/1C homozygous	1T/1C	1T/1C	A/A	A/A
NA11894	1T/1C homozygous	1T/1C	1T/1C	A/A	A/A
NA12718	1T/1C homozygous	1T/1C	1T/1C	A/A	A/G
NA12282	1T/1C heterozygous	1T/1C	4T/3C	A/G	A/G
NA12341	Non 1T/1C	4T/3C	2T/9C	G/G	G/G
NA07357	Non 1T/1C	2T/2C	2T/10C	G/G	G/G
NA07000	1T/1C recombinant	4T/2C	2T/1C	A/G	A/A

4.4.8 Experimental verification of 1T haplotype recombination events by Emulsion fusion haplotype PCR

From our recombination analysis in the *DEFA1A3* locus, we showed that telomeric and centromeric recombinant haplotypes were present but the rate of recombination is not high enough to randomise association between the two sides. We also showed that MSV 10479 "A" allele is a tag MSV for 1T/1C haplotype, present in almost all 1T/1C haplotype copies but hardly found in any other haplotypes. Therefore we can predict that 1T recombinant haplotype should have MSV 10479 "A" allele in the most telomeric copies in the repeat array

if recombination breakpoints are not involved the first MSV 10479 position. We designed an emulsion fusion haplotype PCR to genotype the first MSV 10479 position for 1T recombinant haplotypes. We selected five 1000 Genome Project CEU samples with 1T recombinant haplotype. From NGS data analysis all samples have A/G genome type at MSV10479 position. The Sanger sequencing result is shown in table 4.10. Except for NA12234 having a "G" allele, all other recombinant haplotypes have "A" allele at their first MSV 10479 position. Sample NA12234 might have additional gene conversion events that replace the first copy with "G" allele. Nevertheless, this result suggests flanking sequence recombination also involves CNV components. It is the major mechanism of transferring variants from one haplotype class to another.

Table 4.10 Emulsion fusion haplotype PCR verification of 1T haplotype recombination events

	1000 genome NGS data analysis		Sanger sequencing result	
samples ID	NGS haplotypes	NGS MSV10479	1T tag SNP	10479 position
NA11931	4T2C, 1T/64C	A/G	A	A
NA12234	2T/2C, 1T/7C	A/G	A	G
NA07037	1T/16C,5T/7C	A/G	A	A
NA11930	1T/1C, 4T/3C	A/G	A	A
NA11920	1T/8C, 4T/3C	A/G	A	A
NA12889	2T/1C, 1T/7C	A/G	A	A

4.5 Conclusion

We analyzed haplotypes associated with reported IgA Nephropathy and periodontitis GWAS signals and concluded that 2T/2C is a protective haplotype in the Chinese population and that recombinant 2T or 2C also have a (lower) protective association, so there seems an additive effect for 2T/2C copies. CNV internal variants that are specifically associated with 2T/2C haplotypes are potential causative variants. 1T/1C haplotype is a risk factor for IgA Nephropathy in both Asian and European populations. This haplotype is also a risk factor for chronic periodontitis in European populations. We further detected *DEFA1A3* CNV internal variants. We studied the association of all CNV internal variants with all flaking haplotypes by Cochran-Armitage trend test and generated a list of variants that are associated with 2T and 2C haplotypes. We also demonstrated that haplotype-specific variants are commonly present, which suggests that an inter-homologue transfer of genetic material is not common.

5 Chapter5: haplotype 5T/7C as a Neanderthal introgressed haplotype

5.1 Introgression of haplotype 5T/7C from Neanderthals

Among all haplotypes described in chapter 3, haplotype 5T/7C display unique features. Firstly, it is mostly present in European populations and almost absent from African or East Asian populations (Figure 5.3). Secondly, this haplotype has a relatively long divergence from its most closely related human haplotypes. Thirdly, although this haplotype has a higher than average copy number (8 compare to average 6 (Black et al., 2014)), it also seems to have an extremely homogeneous CNV internal SNP background when compared to other haplotypes. This was shown from unpublished CNV internal sequencing work done by previous lab members Holly Black and Omniah Mansouri.

A haplotype with restricted geographic distribution usually indicates young age in evolution. On the other hand deep divergence from other human haplotypes is usually caused by long separate

evolutionary history. These contradictory assumptions reflect the unusual evolutionary history of haplotype 5T/7C. Therefore we began to consider another possible explanation of these observations that this haplotype is introgressed from another archaic hominin into European populations. We think the most possible source of introgression is Neanderthals as they were widely distributed throughout Eurasia. Our following research focuses on evidence of haplotype 5T/7C as a Neanderthal introgressed haplotype and evaluation of positive selection of this haplotype.

5.2 Haplotype 5T/7C is highly similar to Altai Neanderthal haplotype

The most basic feature to suggest introgression of Haplotype 5T/7C from Neanderthal is sequence similarity. An introgressed haplotype should share more alleles with Neanderthal than any other human haplotypes.

In this analysis we compared sequence similarity in ~6kb telomeric and ~20kb centromeric CNV flanking region for 3 archaic samples and 8 common modern human haplotypes obtained from the 1000 Genomes Project phase 3 data set. Altai Neanderthal, Denisovan and Ust-ishim high coverage whole genome sequencing data was obtained from Max Planck Institute for Evolutionary Anthropology. Ust-ishim is

a 45,000 year-old modern human found in western Siberia. A visual (by Integrated Genomics Viewer) check of all three sources of archaic NGS data at the *DEFA1A3* locus shows almost all variant positions are homozygous which indicates these individuals are each carrying two identical haplotypes. The unambiguous allele status allows us to do further analysis without being challenged by the phasing problem.

In comparisons of 5T/7C haplotype (NA12273) in both telomeric and centromeric *DEFA1A3* flanking sequence with Altai Neanderthal sequence, the similarity reached 98.1% in 258 variable positions which minor allele frequency greater than 1%. To display their relationships we constructed a Neighbour-joining tree with Mega 6 software, on the basis of common SNP positions (frequency >1%) in the ~20kb centromeric flanking region. The tree was tested by the bootstrap method, with 1,000 replications. We used chimpanzee reference sequence PanTro 4 as an outgroup. The centromeric NJ tree shows 7C and Altai Neanderthal sequence have the closest relationship among all sequences analysed (Fig 5.1). In 20kb centromeric flanking region, they have only 4 different variant sites. Archaic sample Ust-ishim is most close related to present day haplotype 2C and a close counterpart of the Denisovan haplotype was not seen in the modern human sequences analysed so far.

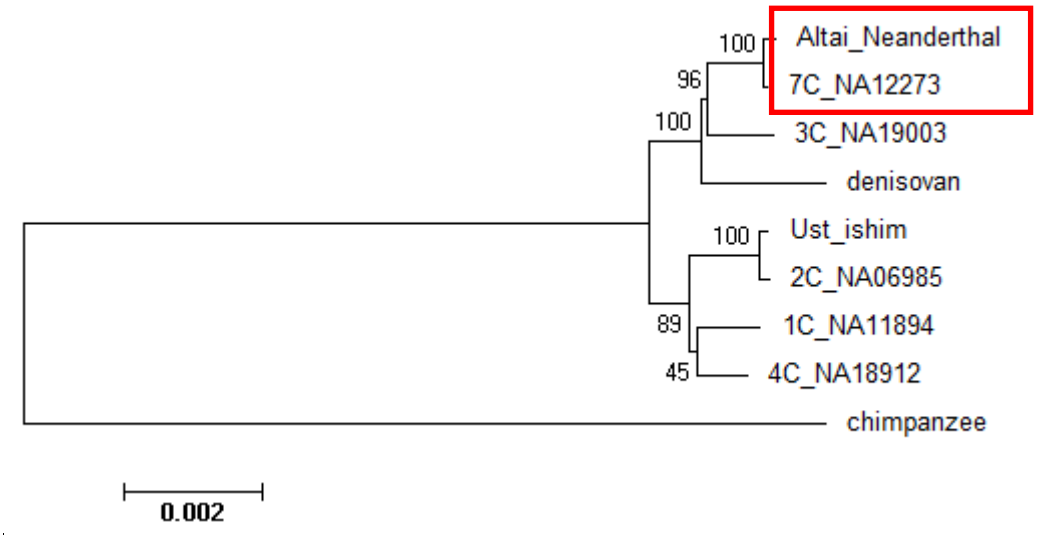


Figure 5.1 Neighbour-joining tree showing 7C haplotype more closely related to Altai Neanderthal haplotype than any other human haplotypes.

5.3 Copy number measurement of haplotype 5T/7C

Although the most accurate experimental haploid copy number typing method for *DEFA1A3* CNV is by a combination of three allelic ratio and two paralog ratio test methods (Khan et al., 2013), it is not applicable here since Altai Neanderthal and other archaic genomic DNA is not available. However, we still can measure diploid copy number from NGS whole genome sequencing data. This involved comparing sequence read depth of a CNV internal region with a CNV flanking region which we are confident has 2 copies. The detail of the method is described in chapter 2.2.9. Our analysis result shows that the Altai Neanderthal diploid copy number estimate is 8.9, which is

very close to modern human 5T/7C average haploid copy number of 7.6. We also used the same method to estimate copy number of other archaic samples. (Table 5.1)

Table 5.1 Copy number measurement for archaic samples

Sample	Core3 read depth	Read depth CN
Altai Neanderthal	29540	8.9
Denisovan	32290	15.0
Ust-ishim	31036	17.1

5.4 Neanderthal derived variants and introgression length

To infer the boundary and length of Neanderthal introgression in the *DEFA1A3* locus, we first defined Neanderthal-derived SNPs as human derived alleles (differing from chimpanzee) that are absent from the African Yoruba population and match with the Altai Neanderthal alleles. The African Yoruba population carries least Neanderthal introgression among all 1000 Genomes Project populations (Sankararaman et al., 2014). We did this analysis for 2634 SNPs spanning a 300kb region near the *DEFA1A3* locus (from hg19 chr8:6,700,000 to 7,000,000) for the 1000 Genomes Project phase1 European (CEU+GBR+TSI), Asian (CHB+CHS+JPT) and African (YRI+LWK+ASW) populations. In total we discovered 38 Neanderthal-derived SNPs in the CNV flanking region. 19.02% (113 out of 594) of European haploid sequences have at least 2

Neanderthal derived SNPs. No such SNPs were discovered in Asian and African populations except 3 exceptions from LWK population.

Analysis of the length of introgression in European populations shows this introgression have clear boundaries between 6800319 and 6932374 (Fig 5.2). Outside there are no Neanderthal derived SNPs identified. The introgression spans the entire *DEFA1A3* CNV flanking region and in the centromeric region also covers the *DEFA5* gene. The full length of the introgression is about 132kb according to the human hg19 reference genome. Given that the average copy number of haplotype 5T/7C is 4, the actual introgression length could be 151kb since human hg19 reference sequence contains 3 repeat copies and each copy is about 19kb in length. Shorter versions also can be seen which could be caused by more recent recombination and gene conversion events after the initial introgression. Sankararaman 2014 showed that the median length of Neanderthal contigs is 129kb when merging all Neanderthal haplotypes in 1000 Genomes Project European and east-Asian populations (Sankararaman et al., 2014). The haplotype 5T/7C 132kb full length introgression is very close to this estimate.

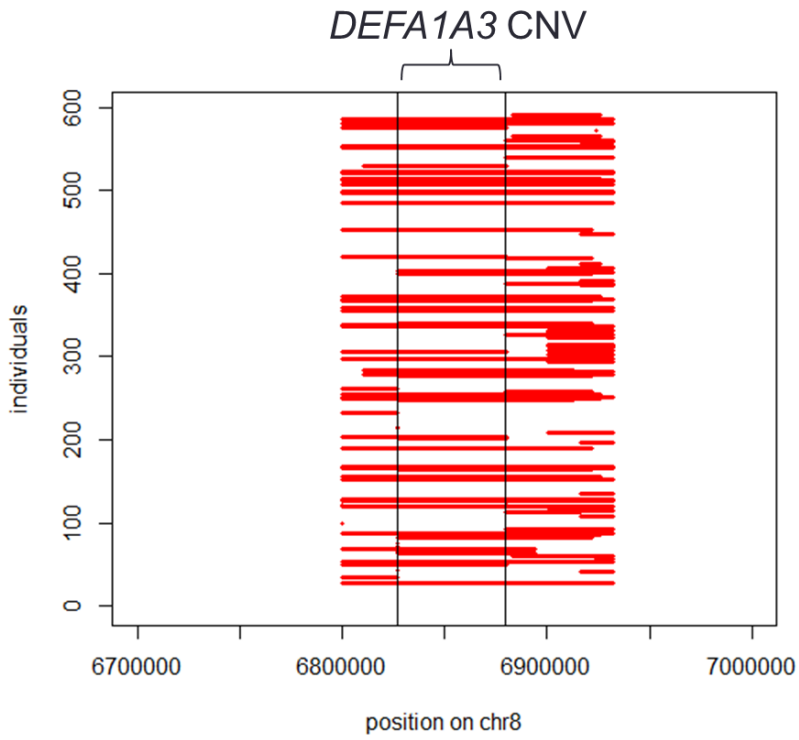


Figure 5.2 Neanderthal derived contigs plotted in 1000 Genomes Project phase 1 European populations (CEU+GBR+TSI) 594 sequences, showing the full introgression length of about 132kb, including the entire *DEFA1A3* CNV and *DEFA5* gene.

5.5 Geographic distribution of 5T/7C haplotype

One of the most important kinds of evidence of archaic introgression is geographic distribution of the introgressed haplotype. Although the majority of non-African human ancestry is shared with Africans, non-Africans also possess a small amount of DNA (1.5-2.1%) from Neanderthals (Green et al., 2010). This is caused by interbreeding

between human and Neanderthals during out of Africa migration in Europe and Asia. If the 5T/7C haplotype was passed from Neanderthals, we should expect that it is not distributed in sub-Saharan Africa. We studied world geographic distribution of 5T/7C haplotype by using 84 populations from the 1000 Genomes Project, HGDP and the ALFRED data set (Table 5.2). For populations from the 1000 Genome dataset we use 5T/7C tag SNP rs62487514 "A" allele frequency to represent 5T/7C haplotype frequency. For populations from HGDP and ALFRED data set we use SNP rs4500099 due to the absence of rs62487514. This SNP is in good LD with rs62487514 (r^2=0.73). Both SNPs are also included in the Neandertal derived SNP list inferred in previous section. The world distribution of 5T/7C haplotype (Fig 5.3) shows it is mainly distributed in Europe, the Middle-East and South Asia. Higher frequency is found in Western Europe, and the Orcadian population has the highest frequency among all populations (21.9%). The frequency is lower in Middle-East, South Asia and North Africa, and may explained by recent migration from or admixture with European population. 5T/7C haplotype is mostly absent from sub-Saharan Africa and East Asia. A few exceptions found in eastern Africa were probably caused by more recent back migration. This world distribution of 5T/7C haplotype is consistent with our Neanderthal introgression hypothesis.

Table. 5.2 Populations used in 5T/7C haplotype geographic distribution study.

population	frequency	total number	source
Orcadian	0.219	32	HGDP
French Basque	0.167	48	HGDP
CEU	0.141	198	1000G
PJL	0.141	192	1000G
Russian	0.14	50	HGDP
IBS	0.126	214	1000G
French	0.121	58	HGDP
Kalash	0.12	50	HGDP
Adygei	0.118	34	HGDP
GBR	0.115	182	1000G
Hungarian	0.1	50	ALFRED
ITU	0.098	204	1000G
CLM	0.09	188	1000G
TSI	0.089	214	1000G
Mozabite	0.083	60	HGDP
Estonian	0.079	1950	ALFRED
PUR	0.077	208	1000G
FIN	0.071	198	1000G
MXL	0.07	128	1000G
Bedouin	0.07	96	ALFRED
Palestinian	0.069	102	HGDP
BEB	0.064	172	1000G
Druze	0.064	94	HGDP
Tuscan	0.063	16	HGDP
Bedouin	0.063	96	HGDP
Makrani	0.06	50	HGDP
STU	0.054	204	1000G
Sardinian	0.054	56	HGDP
GIH	0.044	206	1000G
Pathan	0.043	46	HGDP
Hazara	0.042	48	HGDP
Pashtun	0.04	46	ALFRED
Komi-Zyrian	0.04	48	ALFRED
Burusho	0.04	50	HGDP
Brahui	0.04	50	HGDP
Balochi	0.04	50	HGDP
NAN Melanesian	0.026	38	HGDP
PEL	0.024	170	1000G
Khanty	0.02	46	ALFRED
Chuvash	0.02	50	ALFRED
Sindhi	0.02	50	HGDP
LWK	0.015	198	1000G
ACB	0.01	192	1000G

ASW	0.008	122	1000G
YRI	0	216	1000G
MSL	0	170	1000G
KHV	0	198	1000G
JPT	0	208	1000G
GWD	0	226	1000G
ESN	0	198	1000G
CHS	0	210	1000G
CHB	0	206	1000G
CDX	0	186	1000G
Micronesians	0	48	ALFRED
Koreans	0	106	ALFRED
Gimi	0	30	ALFRED
Atayal	0	50	ALFRED
Ami	0	48	ALFRED
Yoruba	0	48	HGDP
Yizu	0	20	HGDP
Yakut	0	50	HGDP
Xibo	0	18	HGDP
Uygur	0	20	HGDP
Tujia	0	20	HGDP
Tu	0	20	HGDP
She	0	20	HGDP
San	0	12	HGDP
Papuan	0	34	HGDP
Oroqen	0	20	HGDP
Naxi	0	18	HGDP
Mongola	0	20	HGDP
Miaozu	0	20	HGDP
Mbuti Pygmies	0	30	HGDP
Mandenka	0	48	HGDP
Lahu	0	20	HGDP
Japanese	0	58	HGDP
Hezhen	0	18	HGDP
Han	0	88	HGDP
Daur	0	18	HGDP
Dai	0	20	HGDP
Cambodian	0	22	HGDP
Biaka Pygmies	0	64	HGDP
Bantu	0	16	HGDP
Bantu	0	24	HGDP

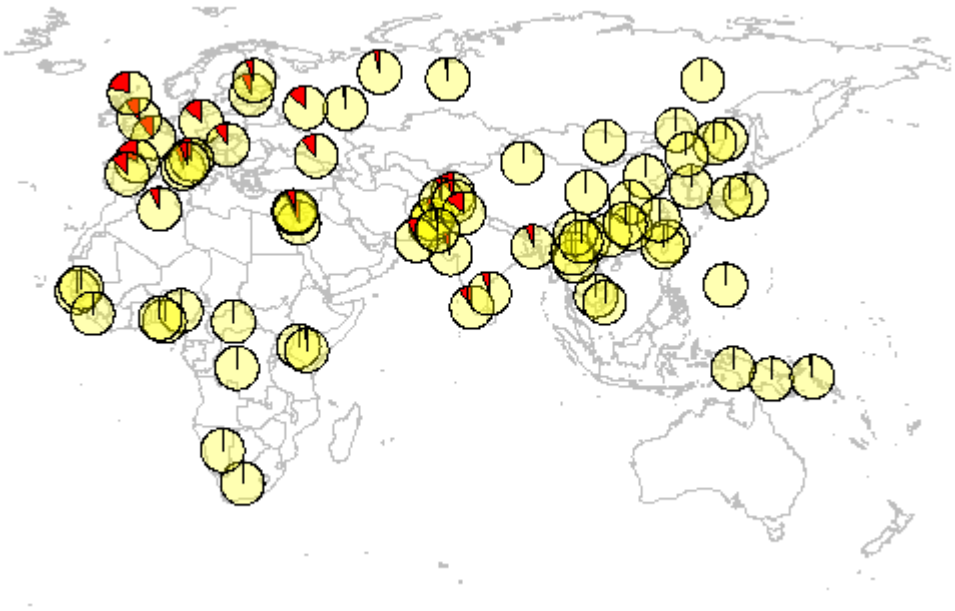


Figure 5.3. Geographic distribution of 5T/7C haplotypes: Most are found in Europe, the Middle-East, North Africa and South Asia. There is a higher frequency in Western Europe. It is absent or very rare in sub-Saharan Africa, East Asia and Melanesia.

5.6 Evidence from S* statistics

The existence of a long introgressed haplotype should increase long range linkage disequilibrium (LD). Therefore, examining the pattern of LD provides an alternative method for establishing introgression (Racimo et al., 2015). The S* statistics were designed to detect such a pattern and were used in our analysis. We chose 5T/7C haplotype homozygous sample NA12273 from the 1000 Genomes Project as a test sample and converted *.bam format to *.fasta format by the GATK tool FastaAlternateReferenceMaker command. We selected the

1000 Genomes Project YRI population as an ancestral population. Then we marked derived SNPs differing from this ancestral population and calculated LD between them. S* values were calculated by a self-made R script following Benjamin et al. (Vernot and Akey, 2014). The detail of the script is described in chapter 2.2.14. This R script uses a sliding window algorithm with window size 6kb and step size 3kb to scan over the entire 300kb region. The S* score shows 5T/7C haplotypes carrying a cluster of SNPs maintained in high LD that were all derived after the separation from the ancestral population. This feature was not seen in other common haplotypes in humans. It suggests that these SNPs were probably derived by introgression after the human out of Africa migration. The S* analysis (fig 5.4) also shows the approximate introgressed area between window 32 and window 80 (6,796,000-6,940,000). This result is also consistent with Neanderthal introgression length analysis discussed in previous section.

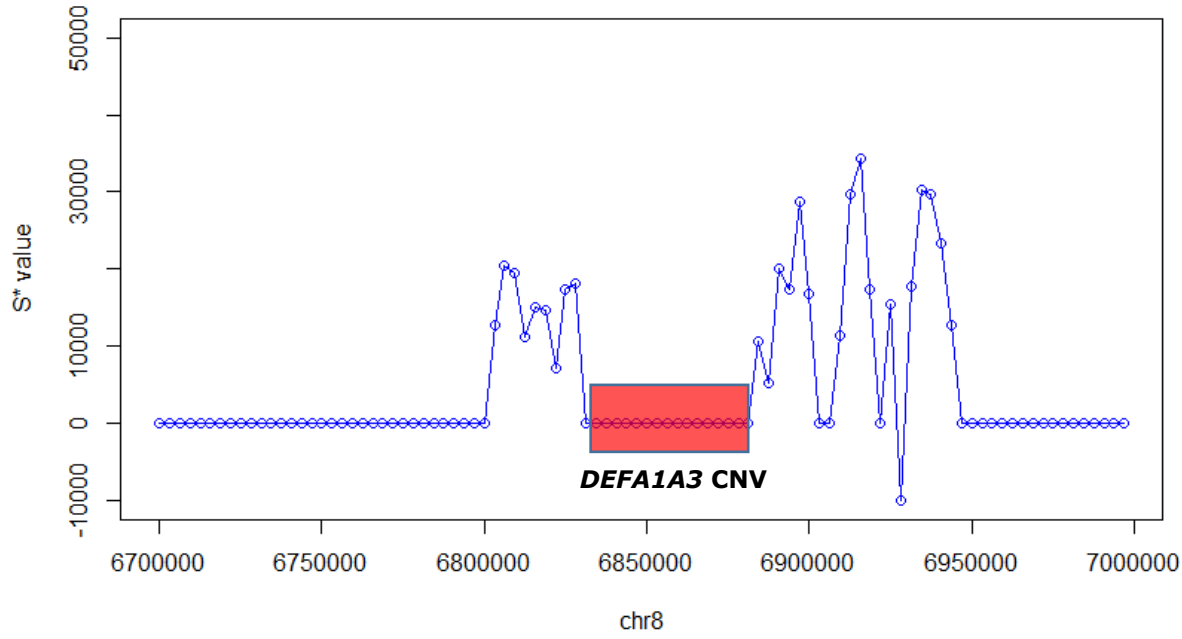


Figure 5.4. S* score plot of a 300kb region near the *DEFA1A3* locus for NA12273, with window size 6kb, and step size 3kb. Variants inside the copy number variable region (marked in red) are excluded from analysis due to poor mapping quality. This plot shows that 5T/7C haplotypes have a high S* in both telomeric and centromeric regions. The size is approximately 144kb.

5.7 Evidence from divergence time estimation

High similarity between human 5T/7C haplotypes and the Altai Neanderthal haplotype can have two possible explanations. 5T/7C haplotype is either an introgression from Neanderthal to humans, or this is explained by incomplete lineage sorting (ILS) (Fig 5.5). In the case of ILS, we would suggest that 5T/7C haplotypes and Neanderthal haplotypes shared a common ancestor in Africa then 5T/7C haplotypes in human lineages were subsequently lost in Africans and Asians due to genetic drift, leaving Europeans as the

only population still possessing this haplotype. In this hypothetical process no introgression of 5T/7C haplotype happened. The difference between the two possibilities is time to the most recent common ancestor (TMRCA). If the TMRCA between the introgressed haplotype and Neanderthal is younger than TMRCA between introgressed haplotype and other human haplotypes then we can reject the possibility of ILS; in a simple world introgression is young, ILS is old. To display the divergence time between haplotype 5T/7C and Neanderthal sequence and distinguish introgression and ILS we used a double Poisson distribution plot (Fig 5.6). If we assume the number of SNP differences between two sequences over a period of time follows a Poisson distribution, by using a calibrator sequence with known divergence time, then we can calculate expected likelihood for possible divergence times between the two sequences (for details: 2.2.15.1). Here we calculate the likelihood of divergence times between 7C haplotype and Neanderthal sequence for 19,522bp of centromeric flanking sequence, and between 7C haplotype and most close related human haplotype 3C. For a given number of SNP differences, we can calculate the most likely time of divergence, and then evaluate the relative probability for other divergence times, giving confidence intervals. The numbers of differences are shown in table 5.3, column 2. To calibrate the time of divergence, we assume generation time is 25 years and human-chimpanzee divergence time

about 6 million years, from which we can calculate the local mutation rate at $4.59X10^{-8}$ per base per generation. This is about 4 times faster than the genomic average of $1.1X10^{-8}$ per base per generation (Roach et al., 2010). We then calculate divergence time for each pair of sequences according to this local mutation rate; results are shown in table 5.3 and fig 5.6.

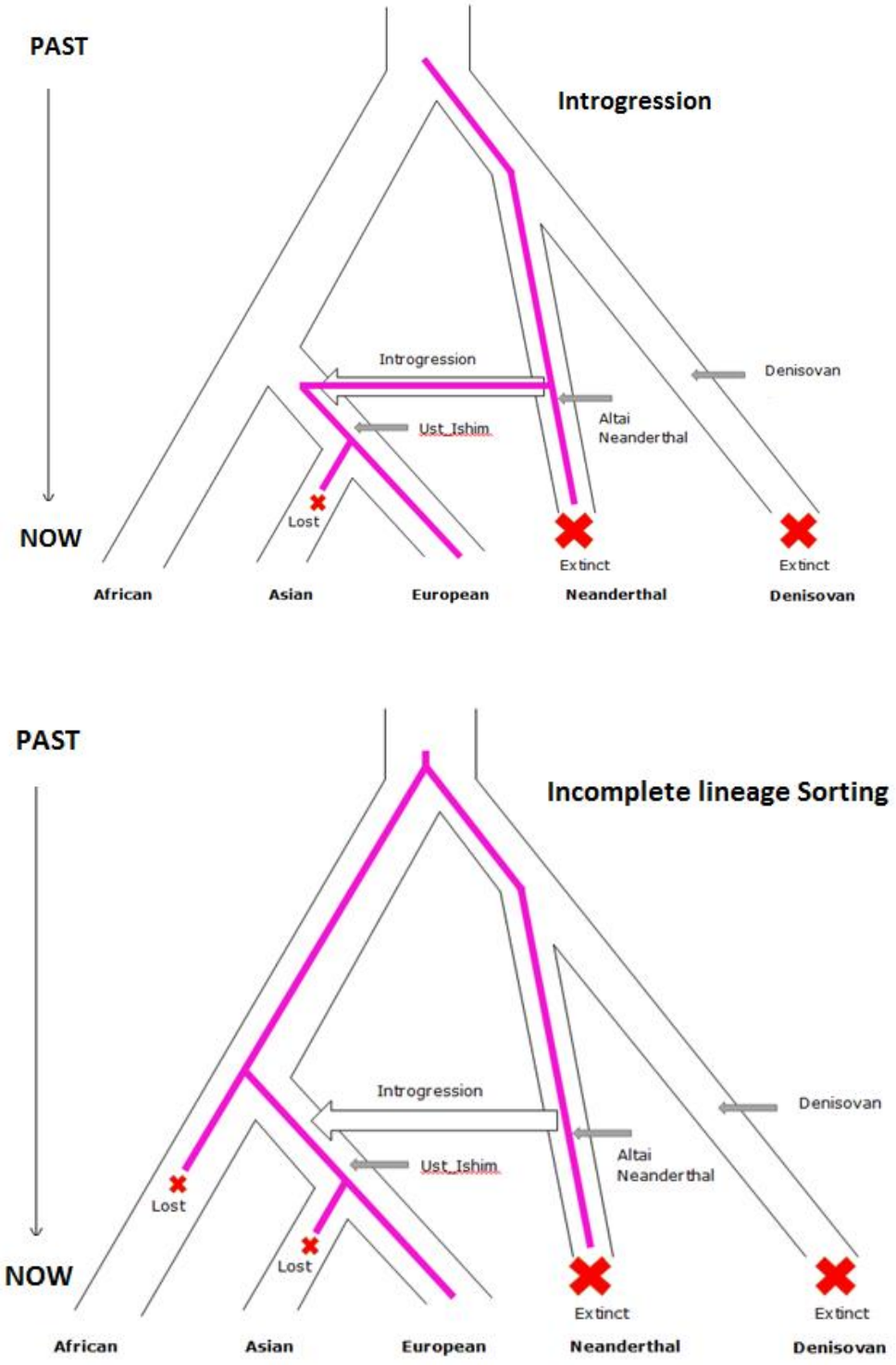


Figure 5.5 Illustration of introgression (above) and incomplete lineage sorting (below) hypothesis of 5T/7C haplotype. Divergence time is a key to distinguish them apart. 5T/7C haplotype in pink colour.

Table 5.3 Divergence time calculation result by double Poisson method.

sequence comparison pair	Number of differences	Most likely divergence time (years)	95% CI divergence time (years)	99% CI divergence time (years)
7C homozygous (NA12273) to Altai Neanderthal sequence	4 SNPs	79250	42500 to 151000	37250 to 182500
7C homozygous (NA12273) to 3C homozygous (NA19003)	35 SNPs	482000	338000 to 656000	302000 to 716000
3C homozygous (NA19003) to chimpanzee (PanTro4)	438 SNPs	Set to 6 million years	Set to 6 million years	Set to 6 million years

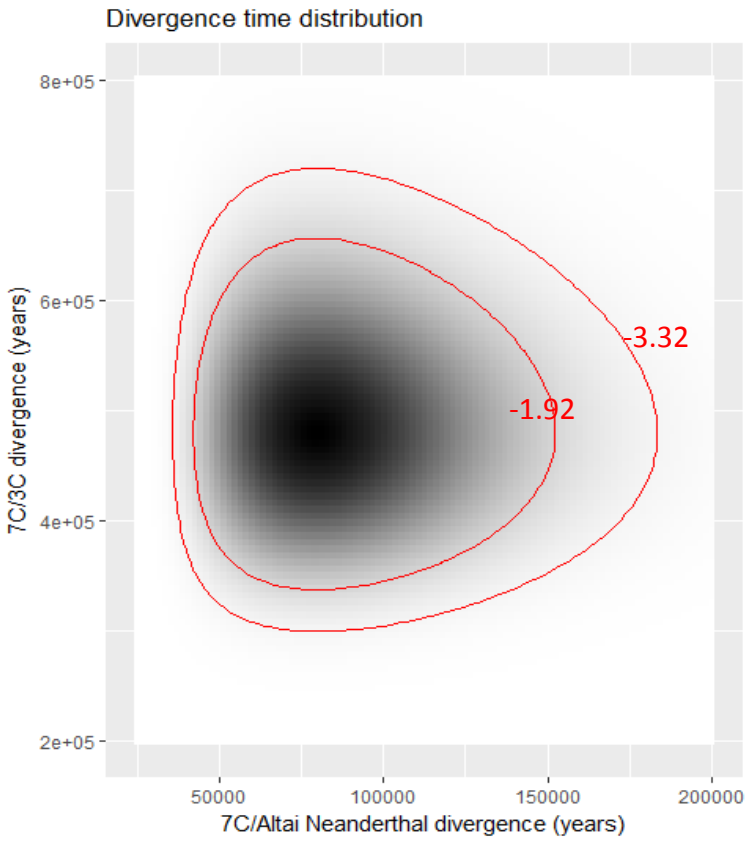


Figure 5.6 Divergence Times estimate between the sequences of haplotype 7C and the 3C; between haplotype 7C and the published Altai Neanderthal Sequence. The -1.92 contour line shows values with log-likelihood value 1.92 lower than the

maximum, giving the 95% confidence interval (CI). Similarly, the -3.32 contour line bounds the 99% CI.

According to calculation, the 95% confidence upper limit divergence time between Altai Neanderthal and 7C haplotypes (TMRCA 151,000 ybp) is younger than the lower limit of the divergence time for 7C haplotype and the 3C haplotype sequence (TMRCA 338,000 ybp). This is also younger than Human Neanderthal population split. This result indicates that the 7C haplotype is more likely introgressed from Neanderthal.

We also used a Bayesian MCMC based method (BEAST software (Drummond et al., 2012)) to calculate TMRCA between 7C haplotype and Altai Neanderthal, and between 7C and the most closely related human haplotype 3C. In this analysis we take the same ~20kb DEFA1A3 CNV centromeric flanking sequence. Samples involved are shown in table 5.4. We used chimpanzee PanTro4 sequence as a calibrator. We set TMRCA between human hg19 reference haplotype and chimpanzee PanTro4 sequence a normal distribution with mean of 6 million years and a standard deviation of 0.5 million years. The details of BEAUti parameter setting are described in chapter 2.2.15.2.

Table 5.4 Sample used in BEAST TMRCA analysis

Name	description	Source
Denisovan	Denisovan	Max Plank Institute for Evolutionary Anthropology (Meyer et al., 2012)
Ust-ishim	Ust-ishim	Max plank institute for evolutionary anthropology (Fu et al., 2014)
Neanderthal	Altai Neanderthal	Max plank institute for evolutionary anthropology (Prufer et al., 2014)
Chimp	Chimpanzee	Chimpanzee reference genome PanTro4
7C NA12273	Modern human 7C haplotype homozygous individual	1000 Genomes Project (The 1000 Genomes Project Consortium, 2015)
1C NA11894	Modern human 1C haplotype homozygous individual	1000 Genomes Project (The 1000 Genomes Project Consortium, 2015)
2C NA06985	Modern human 2C haplotype homozygous individual	1000 Genomes Project (The 1000 Genomes Project Consortium, 2015)
3C NA19003	Modern human 3C haplotype homozygous individual	1000 Genomes Project (The 1000 Genomes Project Consortium, 2015)
4C NA18912	Modern human 4C haplotype homozygous individual	1000 Genomes Project (The 1000 Genomes Project Consortium, 2015)

Analysis results show that the TMRCA distribution between Neanderthal and modern human 7C haplotype mean value is 115,700, 95% highest posterior density (HPD) interval between 12,500 to 271,300 years (fig 5.7). TMRCA between 7C and most closely related human haplotype 3C is 766,200, 95% HPD interval between 296,100 to 1,438,700 years. The upper limit of TMRCA between 7C haplotype and Neanderthal (271,300 ybp) is more recent than the 95% lower limit of the TMRCA between human reference and Neanderthals (296,100 ybp). This is a similar result obtained by

the Poisson method described before, and these observations suggest 7C haplotype is unlikely a result of ILS.

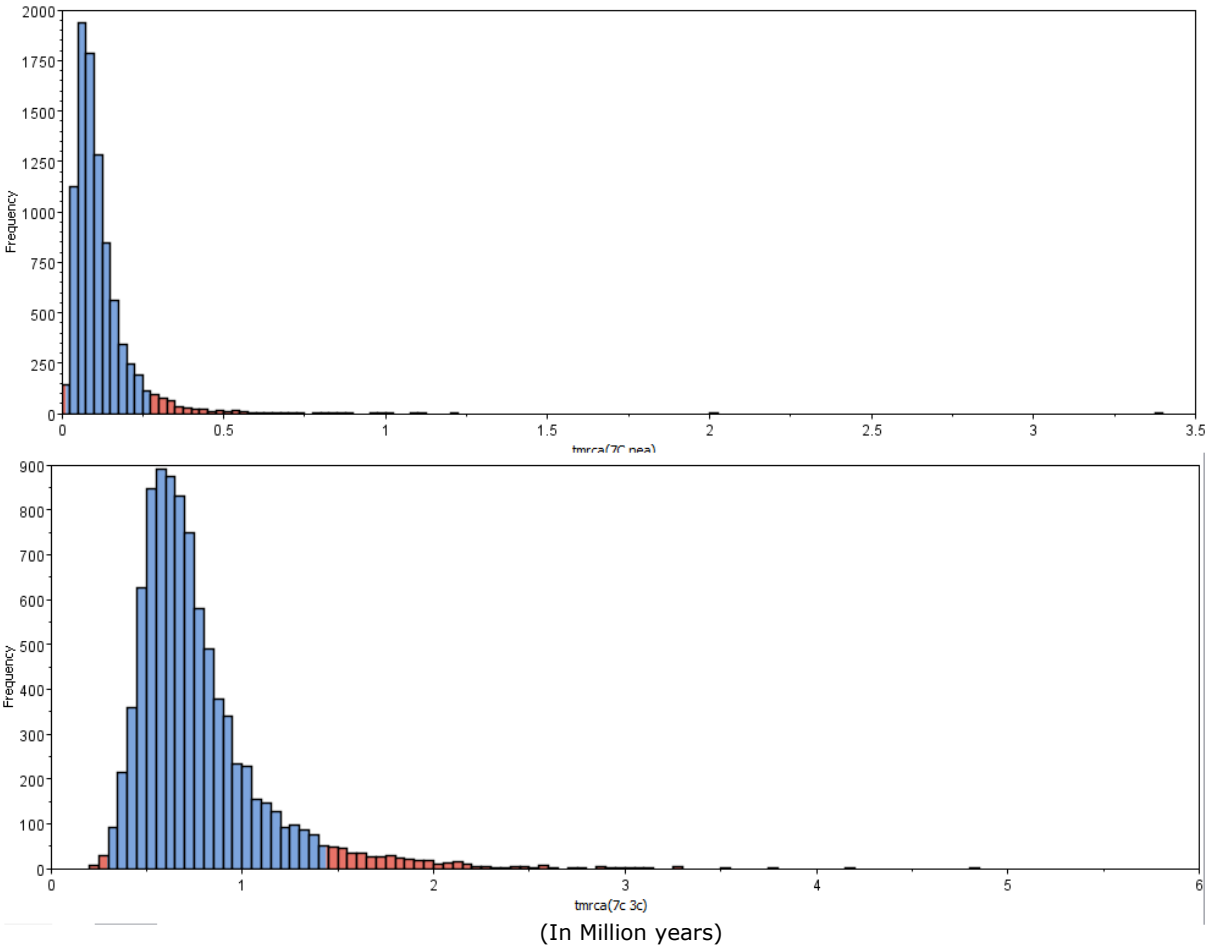


Figure 5.7. Above: TMRCA 95% posterior distribution between human 7C and Altai Neanderthal sequences. Bottom: TMRCA 95% posterior distribution between human 7C and most close related human haplotype 3C.

5.8 Selection test for the 5T/7C haplotype

If a Neanderthal introgression still survives in modern human populations today, the reason is either that it has been maintained by positive selection (i.e., an adaptive introgression) or it was not

removed by negative selection and has been preserved by chance in a process of genetic drift. Adaptive introgression is an important mechanism for an organism to adapt to a new environment in a short amount of time. It is possible that the Neanderthal introgression in this locus helped humans to adapt to a new local environment, and thus the 5T/7C haplotype was positively selected and persisted until today. To address this question we needed to test for positive or balancing selection by statistical methods. There are various selection tests developed so far and they mainly can be classified into three groups. Firstly, there are selection statistics based on allele frequency spectrum, such as Tajima's D (Tajima, 1989). Secondly, some methods are based on linkage disequilibrium structure such as EHH and his (Voight et al., 2006). Thirdly, there are methods based on population differentiation, such as F_{st} (Weir and Cockerham, 1984). We decided to test for positive selection of the 5T/7C haplotype by integrated haplotype score (iHS), because Tajima's D is a method to test all haplotypes in a region, and an F_{st} test examines frequency differences among different populations. In our case the frequency differences of haplotype 5T/7C among populations are mainly caused by migration. iHS can test differences of linkage disequilibrium between ancestral and Neanderthal derived alleles. For calculating iHS first we needed to compute extended haplotype homozygosity (EHH) (Voight et al., 2006) for Neanderthal derived SNPs in the

5T/7C haplotype. We computed EHH for all 38 Neanderthal derived SNPs by the R package rehh. An example of EHH plot for 5T/7C tag SNP rs62487514 is displayed in fig 5.8. It shows the Neanderthal derived allele maintained a much higher level of LD than the ancestral allele. The bifurcation diagram also shows that the Neanderthal derived allele has a slower rate of bifurcation than the ancestral (fig 5.9). We then computed iHH, which is the area under EHH curve. After that we calculated unstandardized iHS which compares iHH area differences between ancestral and derived alleles. Finally, we standardized the iHS value according to whole genome SNP iHS values for SNPs which have similar allele frequency. This measures how unusual an iHS value is in a whole genome level. iHS was calculated for all 38 Neanderthal derived SNPs, and they were normalized by a pre-calculated genome wide standardization file for the 1000 Genome phase3 CEU population obtained from Benjamin F. Voight's lab (Voight et al., 2006). The iHS calculation results for all Neanderthal derived SNPs are listed in table 5.5. The highest |iHS| value is 3.47 for local standardization, 1.47 for genomic standardization, and none of Neanderthal derived SNPs passed the conventional significance threshold of 2 (Voight et al., 2006). This result suggested that although some Neanderthal derived SNPs reached a local iHS significance threshold for positive selection, they are not reaching the threshold on a whole genome level.

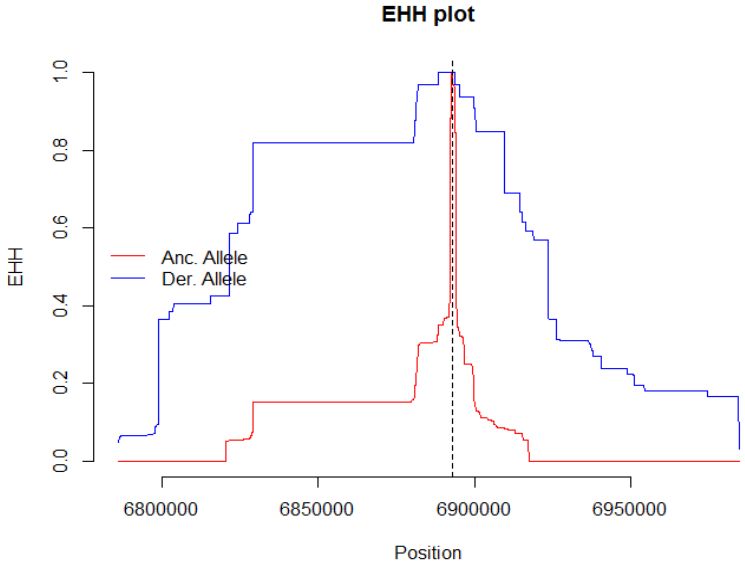


Figure 5.8. EHH plot for rs62487514 of Neanderthal introgression. Derived allele (blue) maintains a higher level of LD than the ancestral allele (red), which suggests that the derived allele may have been under positive selection.

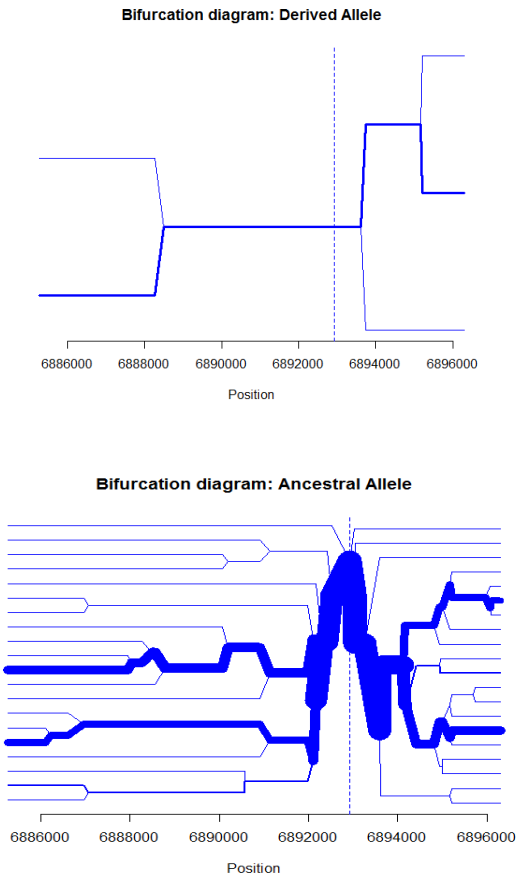


Figure 5.9 Haplotype bifurcation graph for rs62487514 of Neanderthal introgression. The rate of haplotype bifurcation is much slower for the derived allele than the ancestral allele.

Table 5.5 iHS calculation result for all inferred Neanderthal derived SNPs

hg19 chr8 position	iHS	|iHS|	genome wide standardized iHS
6800319	-3.47011	3.470111	-1.465
6810727	-3.04097	3.040969	-1.248
6811380	-3.11301	3.113006	-1.312
6814609	-2.82017	2.820166	-1.256
6819244	-2.46515	2.465154	-1.087
6827143	-2.14674	2.14674	-1.077
6879879	-1.10053	1.100529	-0.925
6880094	-1.20136	1.201356	-1.03
6880521	-1.20136	1.201356	-1.03
6883408	-1.94394	1.943938	-0.674

6883737	-1.9868	1.986801	-0.77
6888488	-2.00178	2.001783	-0.673
6888552	-2.00178	2.001783	-0.673
6890388	-2.05467	2.054672	-0.677
6890854	-2.056	2.055996	-0.684
6892315	-2.07569	2.075686	-0.598
6892915	-2.06709	2.067087	-0.586
6893206	-2.06633	2.06633	-0.586
6894115	-2.07605	2.076047	-0.641
6900678	-2.71363	2.71363	-0.832
6905196	-2.60713	2.607126	-0.849
6909121	-2.61741	2.617406	-0.925
6909886	-2.60222	2.602222	-0.84
6910476	-2.57791	2.577907	-0.823
6911732	-2.55392	2.55392	-0.837
6912787	-2.49387	2.493867	-0.805
6913201	-2.72902	2.729017	-1.079
6916551	-1.98451	1.984515	-0.73
6918858	-1.97171	1.971714	-0.703
6922052	-1.8408	1.840801	-0.571
6923621	-1.8909	1.8909	-0.532
6923700	-2.14314	2.143139	-0.66
6925775	-1.94221	1.942209	-0.503
6926513	-2.25397	2.253969	-0.725
6928549	-2.22161	2.221615	-0.695
6929792	-2.24967	2.249668	-0.712
6930671	-2.28103	2.281034	-0.792
6932374	-2.27904	2.27904	-0.804

There is a potential problem of using iHS as a measure of positive selection in a region involved in Neanderthal introgression. When Neanderthal introgression first occurred, a big block of Neanderthal DNA sequence was passed to human, and then recombination and conversion events cut the introgressed block into smaller pieces. If the introgression is young, it will still maintain a relatively big block, and this may cause a false positive discovery of selection, as iHS is exactly looking for an extended long haplotype as evidence of recent positive selection. To address this problem we plotted empirical iHS distribution for whole genome Neanderthal derived SNPs and

compared it with non-introgressed SNPs to see if there were any differences. Whole genome SNP iHS data was acquired from Benjamin et al 2006 (Voight et al., 2006). This data set contains 8,040,865 SNP iHS scores calculated from the 1000 Genomes Project CEU population. We used whole genome Neanderthal introgressed SNPs as inferred by Sankararaman et al 2014. It use a method based on Conditional Random Field (CRF) that combines information from three features of genetic variations that are signature of Neandertal ancestry. Firstly, common variants between non-Africans and Neanderthal that absent from YRI. Secondly, high sequence divergence to YRI but low divergence to Neanderthal. Thirdly, length of introgression consistent with interbreeding 37-86 thousands years ago (Sankararaman et al., 2014). They concluded that 97,365 SNPs were Neanderthal derived in the 1000 Genomes Project European and east-Asian individuals. Then we combined these two data sets by finding the intersection of SNPs between them. There are 55,044 Neanderthal derived SNPs with an iHS score calculated. We plotted the iHS score density distributions of all Neanderthal derived SNPs and non-introgressed SNPs in the genome (fig 5.10). We found that Neanderthal derived SNPs have a lower mean iHS value of 0.67 compared to the whole genome average SNP value of 0.75. The iHS scores of Neanderthal derived SNPs were not higher on average than the whole genome average, indicating that the Neanderthal

introgression present in the CEU population is old enough that recombination events have enough time to break down the introgressed block. This analysis result suggests that iHS is unlikely to be affected by Neanderthal introgression itself, and thus iHS is not biased and can be directly used in our case. This conclusion is also supported by simulation study carried out by Hu Y. et al 2015 (Hu et al., 2015).

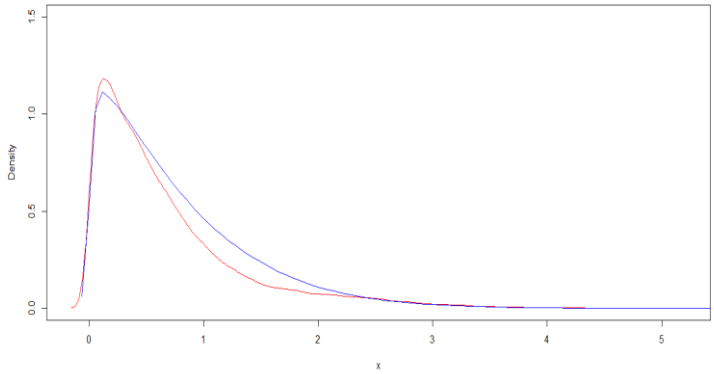


Figure 5.10. Empirical iHS score distribution of whole genome Neanderthal derived SNPs (red) and whole genome SNPs (blue)

5.9 Neanderthal derived variants inside CNV

There is evidence of Neanderthal introgression on either side of *DEFA1A3* flanking sequences and we assume the entire CNV was also introgressed for those haplotypes with introgression on both flanking sides. To validate this point we called variants by a methods described section in 2.2.1 and 2.2.2. We found 79% of the variant

calls in the CNV for 5T/7C can be matched with Altai Neanderthal CNV variant calls.

Another unusual feature of 5T/7C haplotypes is that despite the presence of multiple copies (ranging from 2-7), all copies within the repeat array maintain an identical sequence. This is also a feature when browsing Altai Neanderthal sequences. All other human haplotypes show a diverse inter-copy SNP background. To compare the within-haplotype nucleotide diversity between different haplotype classes, we calculated the average nucleotide diversity per base pair (π) for available samples carrying homozygous haplotypes. We found in total 38 samples from the 1000 Genome Project carrying homozygous haplotypes at both centromeric and telomeric flanking regions. Then we calculated average π per base for a 19kb repeat unit. The result is displayed in fig 5.11; the two 5T/7C homozygous samples NA12273 and NA11919 have lowest average π per base value which means they have least nucleotide diversity and they are 5 times lower than the other samples analysed. The third least diverse sample is the Altai Neanderthal sequence.

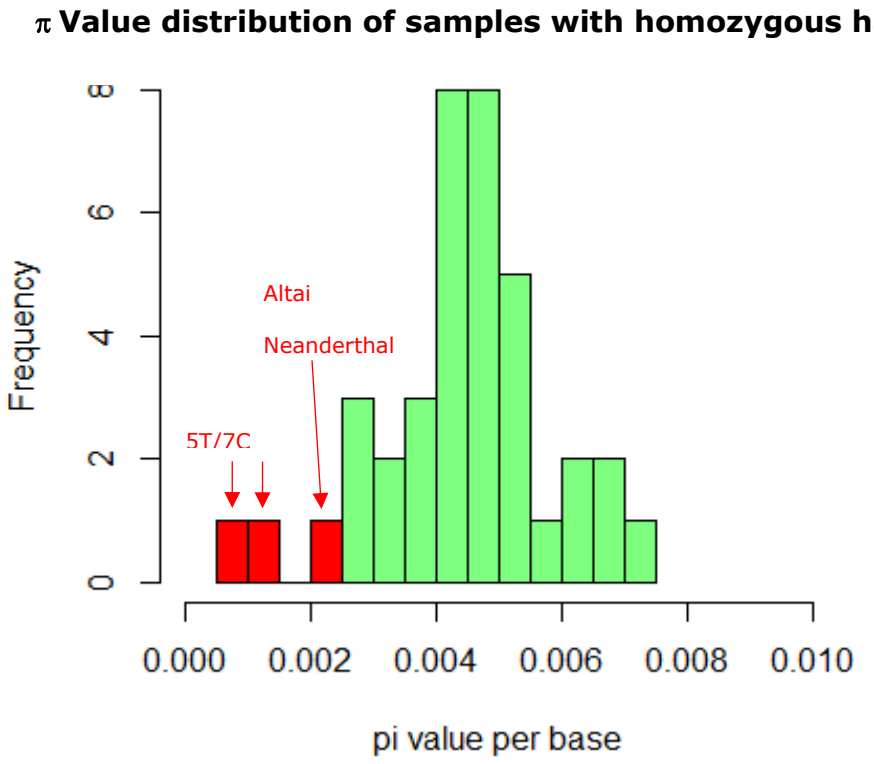


Figure 5.11. Average nucleotide diversity (π) per base. Two 5T/7C haplotype homozygous samples have least CNV internal nucleotide diversity. The third lowest sample is Altai Neanderthal.

5.10 Conclusion

In this chapter we presented evidence that a haplotype 5T/7C in European populations is likely to be introgressed from Neanderthals. We have gathered evidence from sequence similarity on both CNV flanking and CNV repeat internal sequence variation. Geographic distribution of this haplotype is not found in Africans, this is consistent with Neanderthal introgression. Divergence time

estimation suggest it is unlikely caused by ILS. S* calculation detected a signature of introgression. The full introgression length of 5T/7C haplotype (about 140kb) is also similar to the average for Neanderthal introgressions. Together, these lines of evidence confirmed that haplotype 5T/7C was of Neanderthal origin. This is the first reported Neanderthal introgression in a copy number variable region. Its special CNV features may help us to understand evolution pattern in this CNV locus. Finally we evaluated selection on the introgressed haplotype; although iHS analysis shows there is weak evidence of positive selection, it is not significant on a genomic level. With current evidence it is hard to make a conclusion that haplotype 5T/7C is an adaptive introgression unless we find a clear functional target of selection, for example that Neanderthal derived variants may be associated with human diseases.

6 Chapter6: Construction *DEFA1A3* CNV haplotype structure by Oxford Nanopore MinION sequencing

6.1 Introduction to Oxford Nanopore MinION sequencing

Oxford Nanopore company released their third generation sequencing device MinION in 2014. This brought a new revolution to the field of sequencing. Our lab got the opportunity to try this new technology before it was released to market. A general MinION run consists of Library preparation, sequencing and data analysis steps.

6.1.1 Library preparation

Long and double-stranded DNA (>30kb) is usually used as input for the MinION device, so it can generate sequence reads for both strands. Library preparation consists of genomic DNA shearing using a Covaris g-TUBE, then an optional ”PreCR” step to repair damaged DNA, end-repair to create blunt ends in sheared DNA or PCR products and dA-tailing to add “A” base to the 3’ end of DNA fragments; after that two adapters are ligated, the leader adapter and the hairpin

adapter; finally, His-bead purification removes nucleotides and enzymes (Hengyun et al., 2016).

6.1.2 Sequencing process

Currently, the default running time of MinION is 48 hours. Channels on the Nanopore membrane have different activities. Some channels are more active than others and produce the most data, while some channels produce no reads. Each channel contains 4 wells. The most active one is named “g1”, followed by less active “g2”, and so on. After the beginning 24 hours, data are taken from the “g1” well, and in the remaining 24 hours data are equally taken from other three groups (Hengyun et al., 2016).

6.1.3 Data Analysis

The MinION device is connected with a computer installed with MinKNOW software. The main task of MinKNOW is data acquisition, real-time analysis, and feedback. MinION outputs one FAST5 file per read. Currently, the base calling is done by Amazon cloud-based Metrichor. It collects the recorded current values for a molecule and compares with 5-mers using a statistical model trained by Oxford Nanopore Technology. In the library preparation step an adapter was ligated to one side of DNA molecule, which caps the DNA molecule in a hairpin structure and contains easily recognizable characteristic

sequences. During sequencing step, the template strand will first be sequenced by passing through the Nanopore and generating a 1D read, then the adapter sequence and then the complementary strand will produce another 1D sequence. In the base calling step the computer will recognize the position of adapter sequence and combine template and complement strand together to produce 2D reads, if the 1D sequences are of good quality. This step aims to enhance sequencing quality.

LAST and Burrow-Wheeler Aligner (BWA) are popular tools for mapping MinION reads to a reference genome. LAST was initially developed for large-scale sequence comparison. It can map sequence reads from one species to a reference genome of another species (Kiełbasa et al., 2011). It is widely used in MinION sequence mapping because it allows more insertions and deletions. BWA can map DNA sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack (Li and Durbin, 2009), BWA-SW, and BWA-MEM (Li and Durbin, 2010). BWA-backtrack is designed for Illumina short sequences, the other two algorithms supporting long reads and chimeric alignment, suitable for MinION sequence mapping.

6.2 Experiment design to construct *DEFA1A3* CNV haplotype structure

6.2.1 Advantages of using Oxford Nanopore to study CNV

One of the advantages of TGS over SGS is long sequence read length. The Oxford Nanopore MinION device can offer sequence read lengths of tens of kilobases limited only by the length of DNA molecule inputted, while a SGS platform like Illumina can only read ~100bp. This feature allows us to explore highly repeated human genomic regions like the *DEFA1A3* CNV. Long sequence reads would directly cover a large region of repeat unit and greatly simplify the *de novo* assembly process. Each full repeat unit of *DEFA1A3* CNV is ~19kb, and on average each human haploid genome carries 3 copies. We can utilize MinION's long-read ability to read long PCR products from inside the CNV and reconstruct haplotype structure, anchoring with variations in overlapped regions.

6.2.2 Long PCR experiment design

A set of long PCR amplicons was designed by Prof John Armour that overlap with each other to form a tiling path. By *de novo* assembly, we aimed to match different PCR products together and eventually construct the *DEFA1A3* CNV haplotype. The system comprises 3 long

PCRs (fig 6.1): on the hg19 human reference assembly, the predicted PCR1 (chr8:6828471-6838678) length is 10208bp, PCR2 (chr8:6834180-6842187) length is 8101bp, and PCR3 (chr8:6839777-6850357) length is 10574kb. PCR1 and PCR2 have 4kb overlap, PCR2 and PCR3 have 2.4kb overlap and PCR3 and PCR1 have 2.7kb overlap. All overlap regions are expected to contain enough variations to be utilized in *de novo* assembly. Oxford Nanopore MinION have the ability to read single DNA sequences over 10kb in length. Theoretically, we can use this method to reconstruct the two haplotypes carried by each individual, but in practice this method was not very successful. Limitations were identified in the PCR amplification step and sequencing step.

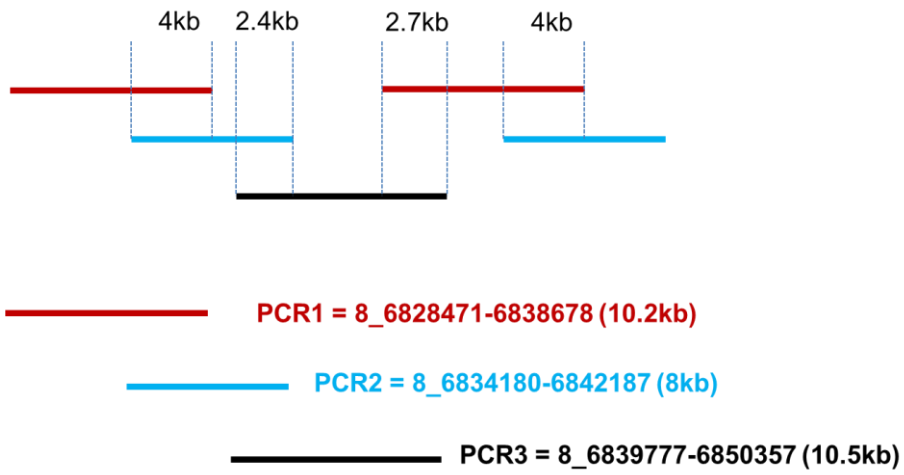


Figure 6.1 Experiment design of *DEFA1A3* CNV haplotype construction by using 3 long PCRs overlapping with each other.

6.3 Problems and limitations for *DEFA1A3* haplotype reconstruction

6.3.1 PCR recombination

We did the Oxford Nanopore MinION sequencing experiment above for sample NA06990. This sample has only 3 copies of *DEFA1A3*. So, we should anticipate a maximum of 3 versions of sequence reads for each long PCR. However, more than three versions of repeat units were observed in Oxford Nanopore MinION reads. We began to suspect that the additional versions of reads might come from PCR amplification steps. Long PCR amplification is more easily influenced by a phenomenon called PCR-mediated recombination. Chimeras (PCR-mediated recombinants) are formed when incompletely extended DNA fragments anneal to matching sequences, generating recombinants between starting templates (Bradley and Hillis, 1997). It can be difficult to differentiate original sequences from chimeras, leading to overestimation of *DEFA1A3* CNV haplotype diversity. To overcome this issue, many studies have determined that chimera formation can be reduced for most DNA polymerases when the cycle number is lowered and extension time increased (Judo et al., 1998).

To determine whether our *DEFA1A3* long PCR was influenced by PCR-mediated recombination, a restriction enzyme digestion experiment

was designed. We use this experiment to examine sample NA06990, who has the simplest diploid constitution (3 copies of the *DEFA1A3* repeat unit) so far found in the human populations studied. This sample is also extensively sequenced in previous studies (Tyson and Armour, 2012), and all variations and their position on CNV array are clear. In our experiment, we used PCR1 primers to amplify this sample and yielded ~10.2kb products from 3 copies. Then we used the restriction enzyme *Sma*I to cut the PCR products; this enzyme has some variable sites that differ between the repeat units in NA06990. Because we know the full sequence of this sample, we can predict the cutting sites and calculate predicted fragment sizes (fig 6.2). Copy 1 has only one *Sma*I cutting site and will create 278bp and 10068bp fragments. Copy 2 has two cutting sites at positions 278 and 3912, and this will create 278bp, 3634bp and 6434bp fragments. Copy 3 has only one cutting site at position 5951, which creates fragment sizes of 5851bp and 4395bp.

We did the above experiment either by using genomic DNA input, or a secondary amplification using PCR amplicons for 3 samples: AF-111, TT-293, and NA06990. The result is shown in figure 6.3 and table 6.1. AF-111 and TT-293 are local samples with a higher copy number of 8 measured by pulsed field gel analysis. All expected bands except 278bp can be seen for NA06990 after *Sma*I enzyme digestion. The 278bp fragment is not shown on the agarose gel image

due to its small size. We also observed recombinant bands at 3912bp and 2039bp. The 3912bp fragment is caused by recombinant between copy 2 and copy 3, bringing a new *Sma*I cutting site into the context of copy 3. 2039bp is caused by the transfer of site 2 to copy 3, and a fragment between site 2 and site 3 has length 2039bp. 5673bp is created by a fragment between site 1 and site 3, caused by recombination between copy 3 and copy 1 or copy 2. In our result, fragment 5673bp is not clear since it is very close to 5951bp; the two bands are not distinguished very well. Nevertheless, the appearance of 3912bp is sufficient to indicate PCR-mediated recombinants created during PCR amplification process. This result suggests that our long PCR approach is troubled by PCR-mediated recombination.

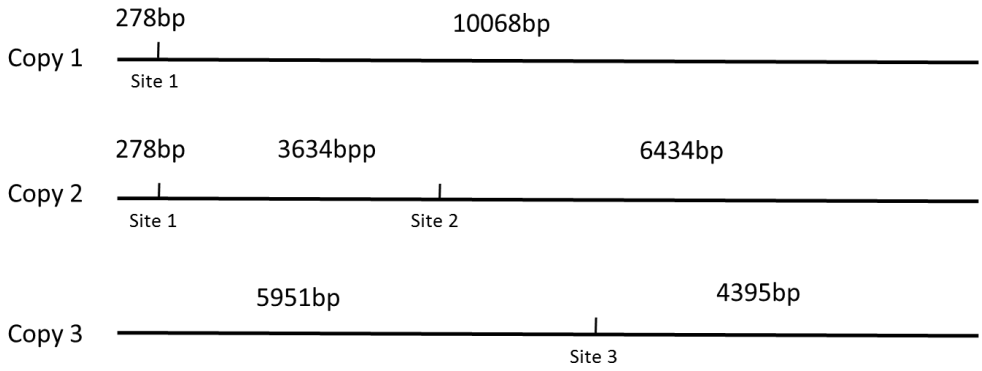
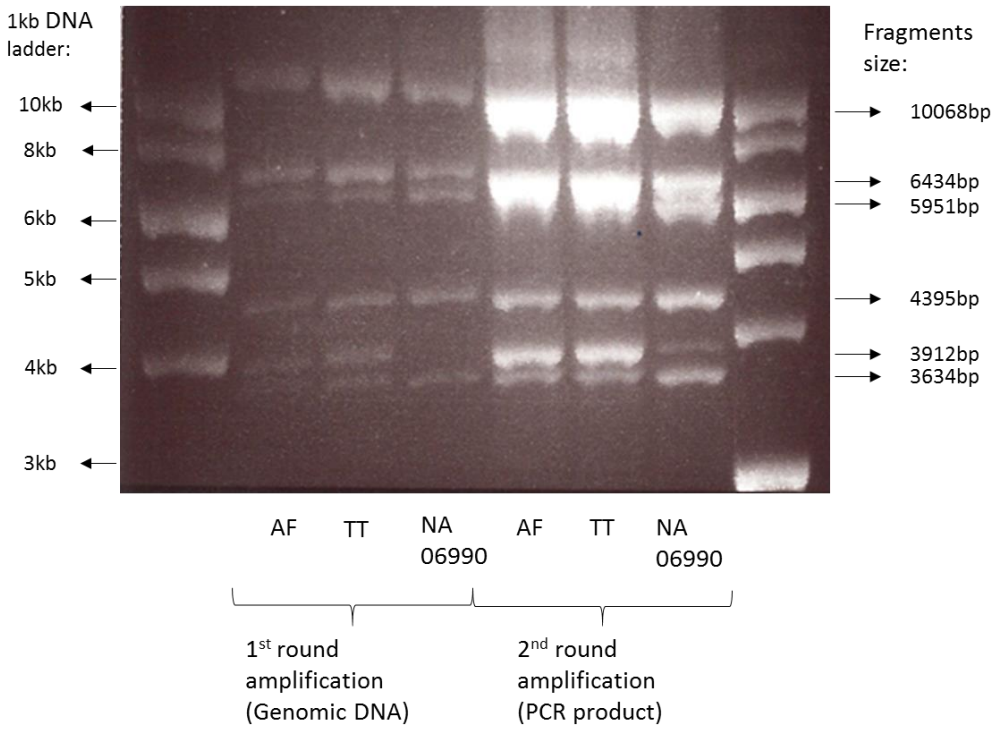


Figure 6.2 *Sma*I restriction enzyme cutting site and fragment size illustration for all 3 non-recombinant copies of sample NA06990



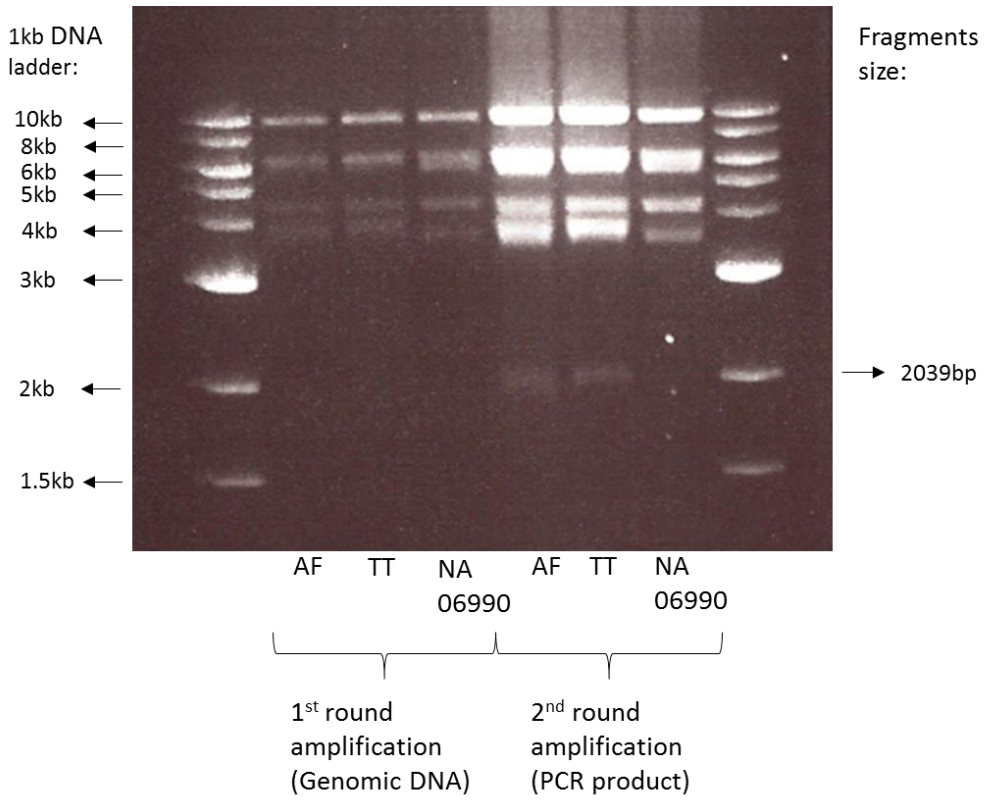


Figure 6.3 Gel electrophoresis image of *Sma*I enzyme digestion experiments. For sample NA06990, recombinant fragments 2039bp and 3912bp are clear on the gel,

indicating our PCR amplification is influenced by a PCR-mediated recombination event. The upper image shows a close-up of fragments greater than 3kb, while the lower image also includes fragments between about 1kb and 3kb.

Table 6.1 List of predicted non-recombinant and recombinant fragments (in red) after *Sma*I enzyme digestion for sample NA06990.

FRAGMENT SIZE		NA06990
278BP	Non-recombinant	?
2039BP	recombinant	√
3634BP	Non-recombinant	√
3912BP	recombinant	√
4395BP	Non-recombinant	√
5673BP	recombinant	?
5951BP	Non-recombinant	√
6434BP	Non-recombinant	√
10068BP	Non-recombinant	√

6.3.2 Efforts to reduce PCR-mediated recombination

Double primer input

PCR-mediated recombination occurred due to incompletely extended amplicons created in a previous PCR cycle acting as a mega-primer and annealing to other templates. So a possible direct method to prevent this happening is to give the PCR master mix sufficient primers. When primers occupy most templates, incompletely extended amplicons will have a lower chance to anneal with templates and thus produce fewer recombinant products.

Double dNTPs input

With the same idea of preventing incompletely extended amplicon formation, another approach is to make sure the PCR reaction is supplied with sufficient dNTPs because the dNTPs shortage is another reason to make extension incomplete. Especially during the last a few cycles, when dNTPs are running out, incomplete extension is more likely to happen.

We used the *Sma*I digestion approach described above to detect PCR-mediated recombination events in NA06690. The appearance and amount of 3912bp and 2039bp fragments indicated how seriously the reaction is influenced by PCR-mediated recombination. We did three sets of reactions using standard PCR master mix, double primer input, and double dNTPs input. Each set of reactions composes a first round amplification and second round re-amplification, during second round of re-amplification PCR product were diluted for 20 times. For each test sets, PCR master mix used were same to first round amplifications (chapter 2.3.7). The gel electrophoresis image is shown in fig 6.4. From the image, in all lanes the recombinant products of 3912bp and 2039bp fragments can be seen. The double primer and double dNTPs master mix have more or less equally strong bands of 3912bp and 2039bp fragments compared with the standard master mix, suggesting that the level of PCR-mediated

recombination events is not greatly reduced by the double primer or double dNTPs approaches. Both recombination bands in second round re-amplification are much stronger than in the first round amplification, indicating that increase of PCR cycle number also increases PCR-mediated recombination events.

Above experiments suggest PCR artefacts cannot be solved by increasing primers and dNTPs input approaches. Oxford Nanopore MinION sequencing result contain sequences generated by PCR recombinations.

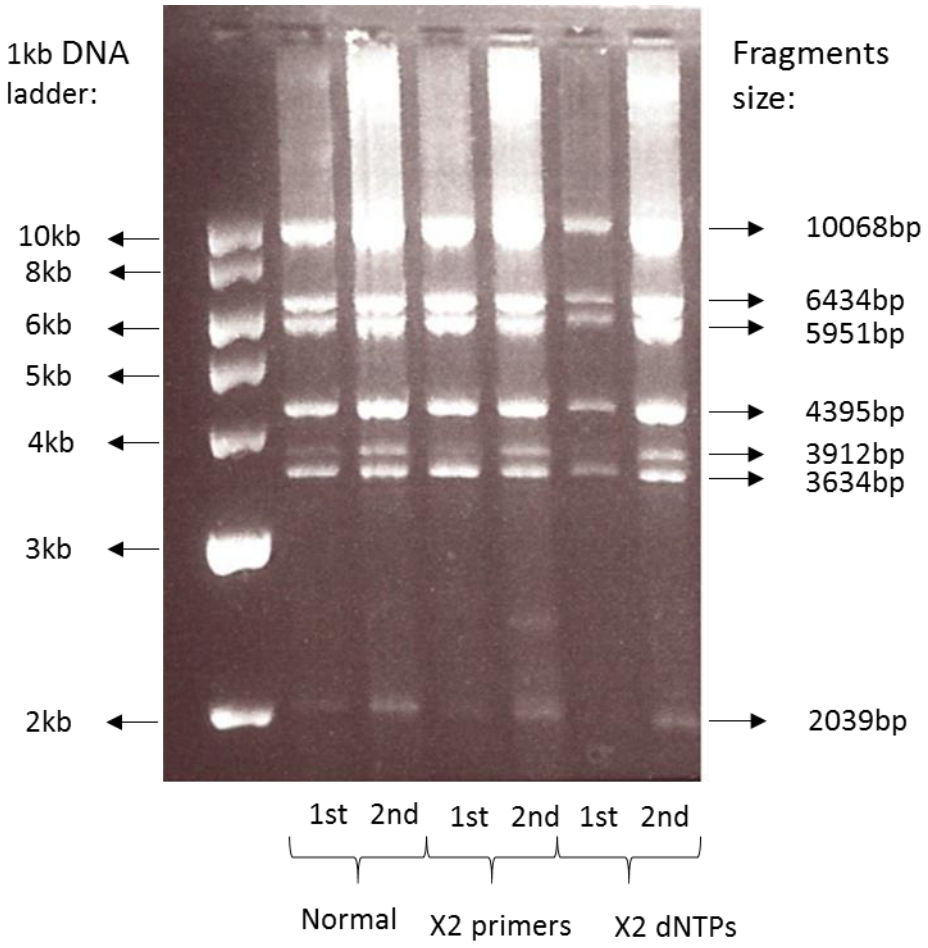


Figure 6.4 Gel electrophoresis image of *SmaI* restriction enzyme digest experiment of NA06990 shows the PCR-mediated recombination events are not reduced by double primer input or double dNTPs input approaches.

6.3.3 Low sequencing accuracy

Another disappointing feature about Oxford Nanopore MinION was the low sequencing accuracy at the time when the experiment was carried out. For our best run, 20,520 reads were generated by the Minion sequencer; 3,454 2D reads can be mapped to reference

sequences, and overall mean coverage is 413.6X. The average mapped read length is 4049bp in length. With this high sequencing depth and long sequence read we should have been able to identify variants and use these variants for *de novo* assembly. However, the R7 flowcells used in this work had an overall accuracy of only 67.4%, with 24.2% mismatched bases, 7.5% insertion and 8.3% deletions according to a report (Norris et al., 2016). Since then, Oxford Nanopore have improved their chemistry and in release version R7.3, the accuracy was enhanced to 86%. The mismatch rate reduced to 9.7%, insertion and deletion rates were down to 4.2% and 4.4% respectively (Norris et al., 2016). Although Oxford Nanopore continues to make efforts to improve sequencing accuracy, it still cannot match the levels of accuracy seen with second and first generation sequencing methods. Sanger sequencing generally provides per-base accuracies as high as 99.999% (Shendure and Ji, 2008). Illumina has a single read error rate around 0.1% (Glenn, 2011). To improve the overall sequencing accuracy repeated sequencing at higher read depth is a strategy to compensate low accuracy in a single read. This approach is widely used in second generation sequencing technologies. However, sequencing at higher depth can only solve random sequencing errors. Oxford Nanopore also produces systematic errors, in which errors appeared repeatedly in certain base positions. This is introduced during the sequencing

process when a DNA molecule passes through a Nanopore. The ionic current amplitude is a function of one or more nucleotides. When more than one nucleotide affects the current (e.g., a triplet), it is impossible for the software to resolve the complex current change, leading to false base calling (O'Donnell et al., 2013); this type of error cannot be mitigated, regardless of how many times the DNA molecule is sequenced.

6.4 Conclusions

In this chapter, we designed a tiling path method to reconstruct *DEFA1A3* CNV haplotype structure by using state of art sequencing technology Oxford Nanopore MinION. Unfortunately, the method was unsuccessful, limited by two main technical difficulties. Firstly, the PCR amplification step was influenced by PCR-mediated recombination, scrambling the variant combinations between different repeat units. So far, we were unable to find an optimum solution to deal with this issue. Secondly, the MinION sequencing accuracy was too low, creating difficulties in *de novo* assembly step later. The Oxford Nanopore company continues to make efforts to improve sequencing technology. Hopefully, we will overcome these difficulties in the future.

7 Chapter 7: Discussion

7.1 Functional variants inside DEFA1A3 CNV

In this research, we detected 1774 variants inside *DEFA1A3* CNV in 9 1000 Genomes Project human populations. This result provided a CNV internal variants list in the *DEFA1A3* CNV for the first time. Variants inside the CNV may cause important changes to alpha defensin gene expressions and functions. For example, coding mutations could alter protein coding, create partially or completely non-functional defensin proteins. Mutation on signal peptide can turn the gene into a pseudo gene just like the cause of theta defensin pseudo genes in human and other primates. Mutation on regulatory elements could dysfunction their regulatory activities. We have discussed details of MSV 2284 and MSV 10479 in chapter4. Apart from that, there are 40 MSVs that have a potential functional effect since they are located in important functional elements (table 4.5).

There are 25 MSVs located in the *DEFA1/A3* coding region. Most of the coding variants appear in low frequency except MSV 1625 (*DEFA1* to *DEFA3* change), MSV 2371 and MSV 2284, which have a frequency greater than 1%. 17 of them change the amino acid coding. Although the change in amino acid coding may not necessarily change gene function, there are still plenty of potential variants that need to be

studied at the functional level. MSV 2353 created a stop codon in exon 2 that stops translation of the alpha defensin gene in at least one repeat unit, and this suggested that pseudo-copies are present in *DEFA1A3* repeat arrays. Given that human individuals have an average of 6 copies of *DEFA1/A3* genes, one or two pseudo-copies may be tolerated in repeat arrays. This indicates that real functional copy number may not be equal to total copy number in some individuals. Previous study suggested that there was no simple relationship between total (DEFA1+DEFA3) mRNA levels and total gene copy number which may be due to the superimposed influence of trans-acting factors (Aldred et al., 2005). Apart from that the presence of non-functional pseudo-copies may also contribute to this observation. Except variants directly leading to amino acid coding changes, there are MSVs found in the signal peptide, promoter region, non-coding (5'UTR) regions in exon 1, and the exon-3 3'UTR.

7.2 Recovering CNV internal variants from next generation sequencing data

In this Ph.D. study, we have developed a new method which can detect variations from inside the *DEFA1A3* CNV, resulting in detection of a large number CNV internal variants for the first time. Some variants were further validated by Sanger sequencing and PCR-RFLP.

The list of *DEFA1A3* CNV internal variation provides good material for further functional studies, and is a guide for disease association and the search for a real causative function variation underlying a GWAS hit. For example, coding variant MSV 2270 is found to tightly associate with 2T/2C flanking haplotypes. It can be used as a tag to study expression of 2T/2C haplotype classes. Currently, our CNV internal variant detection method still lacks a systematic performance evaluation in term of sensitivity and specificity. This evaluation will provide a cut off threshold for the variant-calling quality score and minimize false positive detection. This evaluation cannot be elevated by Sanger sequencing as workload and cost are too high, and a microarray based genotyping method may be more suitable in this case.

For further application, this method can be applied as a tool to detect CNV internal variants for any polymorphic copy number variable regions in the genome. The key point is to develop an algorithm to automatically recognize the CNV region and to generate a single copy reference sequence for realignment of reads. Identifying whole genome CNV internal variants could bring benefits for human disease study, association studies, and evolutionary studies. NGS based sequencing methods are widely used as a tool for finding causative variants for genetic disorders. It is reported that 4.8-9.5% of the human genome is CNV (Zarrei et al., 2015). Due to mapping quality

problems, currently there is no method to recover and utilize variants from inside CNV regions, which ignores a lot of sequence information. The CNV variant detection tool will improve the potential of NGS data to detect disease-causing variants, and enhance the power of disease association studies inside CNV region. In evolutionary studies variants inside CNVs will provide additional information for CNV formation mechanisms and population history.

7.3 Evolutionary history of *DEFA1A3* CNV

Our *DEFA1A3* CNV flanking haplotype diversity analysis classified human population haplotype classes. With additional information on haplotype geographic distribution, we can infer the likely evolutionary history of this locus. *DEFA1A3* CNV diversity in African, Asian and European populations shares lots of common haplotypes which may derive from a common ancestor in East Africa before the human Out of Africa migration. African populations have more diversity and more examples of unique haplotypes compared to other populations. This is consistent with the observation of greater levels of genetic diversity, extensive population substructure, and less linkage disequilibrium in African populations (Campbell and Tishkoff, 2008). Both 1T/1C and 2T/2C haplotypes appear in low frequency in Africa and they became predominant haplotypes in Eurasians by founder effect or selection.

European populations acquired a likely introgressed haplotype 5T/7C from the Neanderthal population, and this haplotype is absent from Asia and sub-Saharan population which indicates that introgression may have happened in Europe after Asian and European population separation. There is a trend of higher introgressed haplotype frequency in Western Europe than in Eastern Europe, Middle East and Indian populations. This may be explained by introgression first happening in Europe, followed by spread to the Indian subcontinent by recent human migration of "Ancestral North Indian" (ANI) populations towards the Indian subcontinent. This is a group of populations which genetically close to Middle Easterners, Central Asians, and Europeans (Reich et al., 2009). The major differences between European and Asian populations are recombinant haplotypes. These recombinant haplotypes are population specific, most of them with clear recombination breakpoints which indicate they are young and formed in Europeans and Asians separately.

7.4 CNV internal features of Neanderthal introgressed haplotype

We have concluded that the 5T/7C haplotype in humans is a Neanderthal introgressed haplotype and discovered that it has a special features different from other human haplotypes. The 5T/7C

haplotype has a very low inter-copy variation diversity; all repeat units in this haplotype seem to carry the same SNP background. This is an unusual phenomenon. Its formation could be explained by three hypotheses.

Firstly, it could be due to the Neanderthal having a relatively small effective population size. Previous studies of Neanderthal DNA sequence have suggested that Neanderthals had lower diversity and thus had a smaller population size than modern humans (Prufer et al., 2014). A small population is more easily influenced by genetic drift and has a lower genetic diversity within species. In 2009, a study estimated variation among Neanderthals was approximately one-third of that estimated for present-day humans worldwide, suggesting a female effective population size of fewer than 3500 individuals (Briggs et al., 2009). It is possible the *DEFA1A3* CNV haplotype in Neanderthal population was reduce to a single haplotype classes and with very low copy number by bottleneck effect and genetic drift. Thus the 5T/7C haplotype could have ended with a very low SNP diversity before it introgressed into the human population. If this is true, it will be an additional evidence of 5T/7C haplotype as a Neanderthal introgressed haplotype.

Secondly, low SNP diversity inside CNV could be a result of a mechanism called concerted evolution. Concerted evolution is a

process that explains the observation that repeated (paralogous) sequences within one species are more closely-related than they are to orthologous sequences in another species, although the gene duplication events happened before the specification events (Liao, 1999). This is thought to be achieved by gene conversion events between paralogous sequences through homologous recombination. In *DEFA1A3* CNV, concerted evolution mechanisms appear to have transferred SNPs present in one copy to other copies. The best example is MSV 10479, which is tightly associated with 1T/1C haplotype and found in every repeat unit of this haplotype. Frequent concerted evolution events lower the CNV internal variation diversity.

Thirdly, this observation could be a result of long-term selection events. NAHR is the main mechanism of copy number polymorphism formation in *DEFA1A3* CNV and intra-allelic rearrangement is thought to be the main pattern (Black et al., 2014). *DEFA1A3* CNV total copy number is frequently changing even within the same flanking haplotype classes. This indicates a faster NAHR rate than flanking SNP mutation rate. However, from chimpanzee to human there is no major difference in total copy number. So, there may be a net balance between duplication and deletion events. If a repeat unit sequence is beneficial, and deletion of that copy is deleterious, the immediate effect will be to raise the frequency of the haplotype carrying this beneficial copy. Inside the CNV, this will create a bias in

which selection will favour the copy being duplicated rather than deleted. In time, this copy may become predominant in the repeat array, finally resulting in a low nucleotide diversity inside *DEFA1A3* CNV. This may represent a past positive selection during Neanderthal evolution, since we failed to detect a significant signal of positive selection for 5T/7C haplotype in modern humans by iHS.

Selection and genetic drift hypotheses can be distinguished if we study whole genome Neanderthal CNV internal variant diversity. If the Neanderthal genome shows low diversity in all CNVs, then we can suggest this is caused by Neanderthal's limited population size, because bottleneck effects act on all CNVs but selection acts on only a few CNVs.

Reference

AI, Z., LI, M., LIU, W., FOO, J. N., MANSOURI, O., YIN, P., ZHOU, Q., TANG, X., DONG, X., FENG, S., et al. 2016. Low alpha-defensin gene copy number increases the risk for IgA nephropathy and renal dysfunction. *Sci Transl Ned,* 8**,** 345ra88.

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., et al. 2006. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature,* 439**,** 851-855.

ALDRED, P. M., HOLLOX, E. J. & ARMOUR, J. A. 2005. Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3. *Hum Mol Genet,* 14**,** 2045 - 2052.

ALLEN, A. C., WILLIS, F. R., BEATTIE, T. J. & FEEHALLY, J. 1998. Abnormal IgA glycosylation in Henoch-Schönlein purpura restricted to patients with clinical nephritis. *Nephrol Dial Transplant,* 13**,** 930-4.

AMARASINGHE, K. C., LI, J. & HALGAMUGE, S. K. 2013. CoNVEX: copy number variation estimation in exome sequencing data using HMM. *BMC Bioinformatics,* 14**,** Suppl 2:S2.

AMID, C., REHAUME, L. M., BROWN, K. L., GILBERT, J. G., DOUGAN, G., HANCOCK, R. E. & HARROW, J. L. 2009. Manual annotation and analysis of the defensin gene cluster in the C57BL/6J mouse reference genome. *BMC Genomics,* 10**,** 606.

ARMITAGE, G. C. 1999. Development of a classification system for periodontal diseases and conditions. *Annals of Periodontology,* 4**,** 1-6.

ARMOUR, J. A. L., PALLA, R., ZEEUWEN, P. L. J. M., DEN HEIJER, M., SCHALKWIJK, J. & HOLLOX, E. J. 2007. Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Res.,* 35**,** e19.

ARSUAGA, J. L., MARTÍNEZ, I., ARNOLD, L. J., ARANBURU, A., GRACIA-TÉLLEZ, A., SHARP, W. D., QUAM, R. M., FALGUÈRES, C., PANTOJA-PÉREZ, A. & BISCHOFF, J. 2014. Neanderthal roots: Cranial and chronological evidence from Sima de los Huesos. *Science,* 344**,** 1358-1363.

AXELSSON, E., RATNAKUMAR, A., ARENDT, M.-L., MAQBOOL, K., WEBSTER, M. T., PERLOSKI, M., LIBERG, O., ARNEMO, J. M., HEDHAMMAR, A. & LINDBLAD-TOH, K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature,* 495**,** 360-364.

BABUSHOK, D. & KAZAZIAN, H. J. 2007. Progress in understanding the biology of the human mutagen LINE-1. *Hum. Mutat,* 28**,** 527–39.

BACKENROTH, D., HOMSY, J., MURILLO, L. R., GLESSNER, J., LIN, E., BRUECKNER, M., LIFTON, R., GOLDMUNTZ, E., CHUNG, W. K. & SHEN, Y. 2014. CANOES:

detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res,* 42**,** e97.

BANDELT, H. J., FORSTER, P. & RÖHL, A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol,* 16**,** 37-48

BARRAT, J. & FEEHALLY, J. 2011. Primary IgA nephropathy: new insight into pathogenesis. *Semin Nephrol,* 31**,** 349–360.

BARRATT, J., SMITH, A. C., MOLYNEUX, K. & FEEHALLY, J. 2007. Immunopathogenesis of IgAN. *Semin Immunopathol,* 29**,** 427-43.

BEDI, T., MAHENDRA, J. & AMBALAVANAN, N. 2015. Defensins in periodontal health. *Indian J Dent Res,* 26**,** 340-4.

BELOV, K., SANDERSON, C. E., DEAKIN, J. E., WONG, E. S., ASSANGE, D., MCCOLL, K. A., GOUT, A., DE BONO, B., BARROW, A. D., SPEED, T. P., et al. 2007. characterization of the opossum immune genome provies insights into the evolution of the mammalian immune system. *Gnome Res,* 17**,** 982-991.

BERGER, J. & HINGLAIS, N. 1968. Intercapillary deposits of IgA-IgG. *J Urol Nephrol (Paris),* 74**,** 694–695.

BERGER, J., YANEVA, H., NABARRA, B. & BARBANEL, C. 1975. Recurrence of mesangial deposition of IgA after renal transplantation. *Kidney Int,* 7**,** 232–241.

BLACK, H., KHAN, F., TYSON, J. & AL ARMOUR, J. 2014. Inferring mechanisms of copy number change from haplotype structures at the human DEFA1A3 locus. *BMC Genomics*.

BORRELL, L. N. & PAPAPANOU, P. N. 2005. Analytical epidemiology of periodontitis. *Journal of Clinical Periodontology,* 32**,** 132-158.

BRADLEY, R. D. & HILLIS, D. M. 1997. Recombinant DNA sequence generated by PCR amplification. *Mol.Biol.Evol,* 14**,** 592-593.

BRIGGS, A. W., GOOD, J. M., GREEN, R. E., KRAUSE, J., MARICIC, T., STENZEL, U., LALUEZA-FOX, C., RUDAN, P., BRAJKOVIC, D., KUCAN, Z., et al. 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science,* 325**,** 318-21.

CAMPBELL, M. C. & TISHKOFF, S. A. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet,* 9**,** 403-33.

CAO, Y. L., QIAO, M., XU, Z. H., ZOU, G. M., MA, L. L., LI, W. G. & XU, B. H. 2016. The clinical study of IgA nephropathy with severe chronic periodontitis and aggressive periodontitis. *National Medical Journal of China,* 96**,** 9-13.

CARPENTER, D., DHAR, S., MITCHELL, L. M., FU, B., TYSON, J., SHWAN, N. A., YANG, F., THOMAS, M. G. & ARMOUR, J. A. 2015. Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes. . *Hum Mol Genet,* 24**,** 3472–3480.

CARPENTER, D., WALKER, S., PRESCOTT, N., SCHWALKWIJK, J. & ARMOUR, J. A. L. 2011. Accuracy and differential bias in copy number measurement of *CCL3L1*. *BMC Genomics,* 12**,** 418.

CARTER, N. P. 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet,* 39**,** S16 - 21.

CHEN, Q. X., HAKIMI, M., WU, S. J., JIN, Y., CHENG, B. L., WANG, H. H., XIE, G. H., GANZ, T., LINZMEIER, R. M. & FANG, X. M. 2010. Increased Genomic Copy Number of DEFA1/DEFA3 Is Associated with Susceptibility to Severe Sepsis in Chinese Han Population. *Anesthesiology,* 112**,** 1428-1434.

CHEN, Y., TANG, Z., WANG, Q., YU, Y., ZENG, C., CHEN, H., LIU, Z. H. & LI, L. S. 2007. Long-term efficacy of tonsillectomy in Chinese patients with IgA nephropathy. *Am J Nephrol,* 27**,** 170–175.

COPPO, R. 2015. The intestine-renal connection in IgA nephropathy. *Nephrol Dial Transplant,* 30**,** 360-6.

COPPO, R., AMORE, A., HOGG, R. & EMANCIPATOR, S. 2000. Idiopathic nephropathy with IgA deposits. *Ped Nephrol,* 15**,** 139-150.

D'AMICO, G. 2000. Natural history of idiopathic IgA nephropathy and factors predictive of disease outcome. *Am J Kidney Dis,* 36**,** 179-196.

DANNEMANN, M., PRUFER, K. & KELSO, J. 2017. Functional implications of Neandertal introgression in modern humans. *Genome Biol,* 18**,** 61.

DAS, S., NIKOLAIDIS, N., GOTO, H., MCCALLISTER, C., LI, J. X., HIRANO, M. & COOPER, M. D. 2010. Comparative Genomics and Evolution of the Alpha-Defensin Multigene Family in Primates. *Molecular Biology and Evolution,* 27**,** 2333-2343.

DE KONING, A. P., GU, W., CASTOE, T. A., BATZER, M. A. & POLLOCK, D. D. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics,* 7**,** e1002384.

DEBSKI, P. R., GEWARTOWSKI, K., BAJER, S. & GARSTECKI, P. 2017. Calibration-free assays on standard real-time PCR devices. *scientific reports,* 7**,** Article Number: 44854.

DIEGO OTTAVIANI, M. L., DENISE SHEER 2014. The role of microhomology in genomic structural variation. *Trends in Genetics,* 30**,** 85-94.

DING, Q., HU, Y., XU, S., WANG, C. C., LI, H., ZHANG, R., YAN, S., WANG, J. & JIN, L. 2014. Neanderthal origin of the haplotypes carrying the functional variant Val92Met in the MC1R in modern humans. *Mol Biol Evol,* 31**,** 1994–2003.

DIVARIS, K., MONDA, K. L., NORTH, K. E., OLSHAN, A. F., REYNOLDS, L. M., HSUEH, W. C., LANGE, E. M., MOSS, K., BARROS, S. P., WEYANT, R. J., et al. 2013. Exploring the genetic basis of chronic periodontitis: a genome-wide association study. *Hum Mol Genet,* 22**,** 2312-24.

DRUMMOND, A. J., SUCHARD, M. A., XIE, D. & RAMBAUT, A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology And Evolution,* 29**,** 1969-1973.

ERSFELD, K. 2004. Fiber-FISH: fluorescence in situ hybridization on stretched DNA. *Methods Mol Biol,* 270**,** 395-402.

ETHERINGTON, G. J., DICKS, J. & ROBERTS, I. N. 2005. Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination. *Bioinformatics,* 21**,** 278-281.

FEEHALLY, J., FARRALL, M., BOLAND, A., GALE, D. P., GUT, I., HEATH, S., KUMAR, A., PEDEN, J. F., MAXWELL, P. H., MORRIS, D. L., et al. 2010. HLA has strongest association with IgA nephropathy in genome-wide analysis. *J Am Soc Nephrol,* 21**,** 1791–1797.

FENG, P., WANG, X., CASADO, P. L., KUCHLER, E. C., DEELEY, K., NOEL, J., KIMM, H., KIM, J. H., HAAS, A. N., QUINELATO, V., et al. 2014. Genome wide association scan for chronic periodontitis implicates novel locus. *BMC Oral Health,* 14**,** 84.

FENG ZHANG, W. G., MATTHEW E. HURLES, AND JAMES R. LUPSKI 2009. Copy Number Variation in Human Health, Disease, and Evolution. *Annu Rev Genomics Hum Genet,* 10**,** 451–481.

FERNANDO, M. M., BOTEVA, L., MORRIS, D. L., ZHOU, B., WU, Y. L., LOKKI, M. L., YU, C. Y., RIOUX, J. D., HOLLOX, E. J. & VYSE, T. J. 2010. Assessment of complement C4 gene copy number using the paralog ratio test. . *Hum Mutat,* 31**,** 866–874.

FEUK, L., CARSON, A. R. & SCHERER, S. W. 2006. Structural variation in the human genome. *Nat. Rev. Genet.,* 7**,** 85-97.

FRANCOIS, B., HITOSHI, S., LISE, T., HIROYUKI, Y., NICOLAS, M., CHRISTOPHE, M., YASUHIKO, T., BRUCE, A. J. & JAN, N. 2012. Autoantibodies targeting galactosedeficient IgA1 associate with progression of IgA nephropathy. *J Am Soc Nephrol,* 23**,** 1579-87.

FRANGIONE, B. & WOLFENSTEIN-TODEL, C. 1972. Partial duplication in the "hinge" region of IgA1 myeloma proteins. *Proc Natl Acad Sci USA,* 69**,** 3673-3676.

FU, Q., LI, H., MOORJANI, P., JAY, F., SLEPCHENKO, S. M., BONDAREV, A. A., JOHNSON, P. L., AXIMU-PETRI, A., PRUFER, K., DE FILIPPO, C., et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature,* 514**,** 445-9.

GAUTIER, M. & VITALIS, R. 2012. rehh: an r package to detect footprints of selection in genome-wide snp data from haplotype structure. *Bioinformatics,* 28**,** 1176–1177.

GHARAVI, A. G., KIRYLUK, K., CHOI, M., LI, Y., HOU, P., XIE, J., SANNA-CHERCHI, S., MEN, C. J., JULIAN, B. A., WYATT, R. J., et al. 2011. Genome-wide association study identifies susceptibility loci for IgA nephropathy. *Nat Genet,* 43**,** 321–327.

GHOSH, D., PORTER, E., SHEN, B., LEE, S. K., WILK, D., DRAZBA, J., YADAV, S. P., CRABB, J. W., GANZ, T. & BEVINS, C. L. 2002. Paneth cell trypsin is the processing enzyme for human defensin-5. *Nat Immunol,* 3**,** 583-90.

GLENN, T. C. 2011. Field guide to next-generation DNA seuencer. *Molecular Ecology Resources,* 11**,** 759-769.

GREEN, R. E., KRAUSE, J., BRIGGS, A. W., MARICIC, T., STENZEL, U., KIRCHER, M., PATTERSON, N., LI, H., ZHAI, W., FRITZ, M. H., et al. 2010. A draft sequence of the Neandertal genome. *Science,* 328**,** 710-722.

GREEN, R. E., MALASPINAS, A. S., KRAUSE, J., BRIGGS, A. W., JOHNSON, P. L., UHLER, C., MEYER, M., GOOD, J. M., MARICIC, T., STENZEL, U., et al. 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell,* 134**,** 416-26.

HAIG, H., KAZAZIAN, J. & JOHN, V. M. 1998. The impact of L1 retrotransposons on the human genome. *Nat. Genet,* 19**,** 19–24.

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. *Nat Genet,* 47**,** 296-303.

HASTINGS, P. J., LUPSKI, J. R., ROSENBERG, S. M. & IRA, G. 2009. Mechanisms of change in gene copy number. *Nature Rev. Genet.,* 10**,** 551-564.

HELIEH S, O. & DAVID A, P. 2011. Animal models for periodontal disease. *J Biomed Biotechnol,* 2011**,** 754857.

HENGYUN, L., FRANCESCA, G. & ZEMIN, N. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics,* 14**,** 265-279.

HIGHAM, T., DOUKA, K., WOOD, R., RAMSEY, C. B., BROCK, F., BASELL, L., CAMPS, M., ARRIZABALAGA, A., BAENA, J., BARROSO-RUIZ, C., et al. 2014. The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature,* 512**,** 306-9.

HOLLOX, E. 2017. Analysis of Copy Number Variation Using the Paralogue Ratio Test (PRT). *Methods Mol Biol,* 1492**,** 127-146.

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**,** 477-484.

HOTTA, O., MIYAZAKI, M., FURUTA, T., TOMIOKA, S., CHIBA, S., HORIGOME, I., ABE, K. & TAGUMA, Y. 2001. Tonsillectomy and steroid pulse therapy significantly impact on clinical remission in patients with IgA nephropathy. *Am J Kidney Dis,* 38**,** 736–743.

HU, Y., DING, Q., HE, Y., XU, S. & JIN, L. 2015. Reintroduction of a Homocysteine Level-Associated Allele into East Asians by Neanderthal Introgression. *Mol Biol Evol,* 32**,** 3108-13.

IAFRATE, A. J. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.,* 36**,** 949-951.

INGASON, A., RUJESCU, D., CICHON, S., SIGURDSSON, E., SIGMUNDSSON, T., PIETILAINEN, O. P., BUIZER-VOSKAMP, J. E., STRENGMAN, E., FRANCKS, C., MUGLIA, P., et al. 2011. Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Mol Psychiatry,* 16**,** 17-25.

JENSEN - SEAMAN, M. I. & HOOPER - BOYD, K. A. 2013. Molecular Clocks: Determining the Age of the Human – Chimpanzee Divergence. *In: eLS. John Wiley & Sons Ltd, Chichester.***,** http://www.els.net.

JESPERSGAARD, C., FODE, P., DYBDAHL, M., VIND, I., NIELSEN, O. H., CSILLAG, C., MUNKHOLM, P., VAINER, B., RIIS, L., ELKJAER, M., et al. 2011. Alpha-Defensin DEFA1A3 Gene Copy Number Elevation in Danish Crohn's Disease Patients. *Dig Dis Sci,* 56**,** 3517-3524.

JUDO, M. S., WEDEL, A. B. & WILSON, C. 1998. Stimulation and supression of PCR-mediated recombination. *Nucleic Acids Res,* 26**,** 1819-1825.

KENJI, Y., YUSUKE, S., HITOSHI, S., KENJI, S., SATOSHI, H., JAN, N. & YASUHIKO, T. 2014. The Kinetics of Glomerular Deposition of Nephritogenic IgA. *PLoS One,* 9**,** e113005.

KERR, M. A. 1990. The structure and function of human IgA. *Biochem J,* 271**,** 285–296.

KHAN, F. F. 2012. *Characterization of the Alpha Defensin Copy Number Variation in Humans.* PhD PhD thesis, University of Nottingham.

KHAN, F. F. & ARMOUR, J. A. L. 2010. Measuring and Characterizing Copy Number Variation of Human Alpha Defensins.

KHAN, F. F., CARPENTER, D., MITCHELL, L., MANSOURI, O., BLACK, H. A., TYSON, J. & ARMOUR, J. A. L. 2013. Accurate measurement of gene copy number for human alpha-defensin DEFA1A3. *BMC Genomics,* 14.

KHRAMEEVA, E. E., BOZEK, K., HE, L., YAN, Z., JIANG, X., WEI, Y., TANG, K., GELFAND, M. S., PRUFER, K., KELSO, J., et al. 2014. Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. *Nat Commun,* 5**,** 3584.

KIDD, J. M., COOPER, G. M., DONAHUE, W. F., HAYDEN, H. S., SAMPAS, N., GRAVES, T., HANSEN, N., TEAGUE, B., ALKAN, C., ANTONACCI, F., et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature,* 453**,** 56 - 64.

KIEŁBASA, S. M., WAN, R., SATO, K., HORTON, P. & FRITH, M. C. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Research,* 23**,** 487-93.

KIRYLUK, K., LI, Y., SANNA-CHERCHI, S., ROHANIZADEGAN, M., SUZUKI, H., EITNER, F., SNYDER, H. J., CHOI, M., HOU, P., SCOLARI, F., et al. 2012. Geographic Differences in Genetic Susceptibility to IgA Nephropathy: GWAS Replication Study and Geospatial Risk Analysis. *PLoS Genet,* 8**,** e1002765.

KIRYLUK, K., LI, Y., SCOLARI, F., SANNA-CHERCHI, S., CHOI, M., VERBITSKY, M., FASEL, D., LATA, S., PRAKASH, S., SHAPIRO, S., et al. 2014. Discovery of new risk loci for IgA nephropathy implicates gene involved in immunity against intestinal pathogens. *Nat Genet,* 46**,** 1187-96.

KIRYLUK, K. & NOVAK, J. 2014. The genetics and immunobiology of IgA nephropathy. *J Clin Invest,* 124**,** 2325-32.

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., et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science,* 318**,** 420-426.

KRUMM, N., SUDMANT, P. H., KO, A., O'ROAK, B. J., MALIG, M., COE, B. P., NHLBI EXOME SEQUENCING PROJECT, N., QUINLAN, A. R., NICKERSON, D. A. & EICHLER, E. E. 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Research,* 22**,** 1525-32.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LANGERGRABER, K. E., PRUFER, K., ROWNEY, C., BOESCH, C., CROCKFORD, C., FAWCETT, K., INOUE, E., INOUE-MURUYAMA, M., MITANI, J. C., MULLER, M. N., et al. 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci U S A,* 109**,** 15716-21.

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**,** 1235-1247.

LEHRER, R. & LU, W. 2012. α-Defensins in human innate immunity. *Immunol Rev,* 245**,** 84-112.

LEVENE, M. J., KORLACH, J., TURNER, S. W., FOQUET, M., CRAIGHEAD, H. G. & WEBB, W. W. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science,* 299**,** 682-6.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics,* 25**,** 1754-60.

LI, H. & DURBIN, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics,* Epub.

LI, J., LUPAT, R., AMARASINGHE, K. C., THOMPSON, E. R., DOYLE, M. A., RYLAND, G. L., TOTHILL, R. W., HALGAMUGE, S. K., CAMPBELL, I. G. & GORRINGE, K. L. 2012. CONTRA: copy number analysis for targeted resequencing. *Bioinformatics,* 28**,** 1307-13.

LI, M., FOO, J. N., WANG, J. Q., LOW, H. Q., TANG, X. Q., TOH, K. Y., YIN, P. R., KHOR, C. C., GOH, Y. F., IRWAN, I. D., et al. 2015. Identification of new susceptibility loci for IgA nephropathy in Han Chinese. *Nat Commun,* 6**,** 7270.

LIAO, D. 1999. Concerted evolution: molecular mechanism and biological implications. *Am J Hum Genet,* 64**,** 24-30.

LIBRADO, P. & ROZAS, J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics,* 25**,** 1451-1452.

LIPSKY, M. S., PASKETT, K. T., GROSS, A. J. & CHEEVER, V. J. 2017. Periodontitis: a global disease and the primary care provider's role. *Postgrad Med J***,** pii: postgradmedj-2017-134801.

MAEDA, I., HAYASHI, T., SATO, K. K., SHIBATA, M. O., HAMADA, M., KISHIDA, M., KITABAYASHI, C., MORIKAWA, T., OKADA, N., OKUMURA, M., et al. 2012. Tonsillectomy has beneficial effects on remission and progression of IgA nephropathy independent of steroid therapy. *Nephrol Dial Transplant,* 27**,** 2806–2813.

MANOLIO, T. A. 2013. Bringing genome-wide association findings into clinical use. *Nature Reviews Genetics,* 14**,** 549–558.

MARTIN, D. P., MURRELL, B., GOLDEN, M., KHOOSAL, A. & MUHIRE, B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution,* 1**,** vev003.

MATTHIAS MUNZ, E. A. 2017. A genome-wide association study identifies nucleotide variants at SIGLEC5 and DEFA1A3 as risk loci for periodontitis. *Human Molecular Genetics,* 26**,** 2577–2588.

MAXAM, A. M. & GILBERT, W. 1977. A new method for sequencing DNA. *Proc Natl Acad Sci U S A,* 74**,** 560-4.

MAXWELL, A. I., MORRISON, G. M. & DORIN, J. R. 2003. Rapid sequence divergence in mammalian beta-defensins by adaptive evolution. *Mol Immunol,* 40**,** 413-21.

MCCARROLL, S. A. & ALTSHULER, D. M. 2007. Copy-number variation and association studies of human diseases. *Nature Genetics,* 39**,** 37–42.

MENDEZ, F. L., WATKINS, J. C. & HAMMER, M. F. 2012. A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *Am J Hum Genet,* 91**,** 265-74.

MEYER, M., KIRCHER, M., GANSAUGE, M. T., LI, H., RACIMO, F., MALLICK, S., SCHRAIBER, J. G., JAY, F., PRUFER, K., DE FILIPPO, C., et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science,* 338**,** 222-6.

MICHALOWICZ, B. S., DIEHL, S. R., GUNSOLLEY, J. C., SPARKS, B. S., BROOKS, C. N., KOERTGE, T. E., CALIFANO, J. V., BURMEISTER, J. A. & SCHENKEIN, H. A. 2000. Evidence of a substantial genetic basis for risk of adult periodontitis. *Journal of Periodontology,* 71**,** 1699–1707.

MICHEAL, C. S., ARTHUR, L. D. & STEVEN, L. S. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research,* 20**,** 1165-1173.

MITCH , M. & SANG , E. L. 2008. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet,* 24**,** 529–538.

MOLDOVEANU, Z., WYATT, R. J., LEE, J. Y., TOMANA, M., JULIAN, B. A., MESTECKY, J., HUANG, W. Q., ANREDDY, S. R., HALL, S., HASTINGS, M. C., et al. 2007. Patients with IgA nephropathy have increased serum galactose-deficient IgA1 levels. *Kidney Int,* 71**,** 1148–1154.

MUNZ, M., WILLENBORG, C., RICHTER, G., JOCKEL-SCHNEIDER, Y., GRAETZ, C., STAUFENBIEL, I., WELLMANN, J., BERGER, K., KRONE, B. & HOFFMANN, P. 2017. A genome-wide association study identifies nucleotide variants at SIGLEC5 and DEFA1A3 as risk loci for periodontitis. *Hum Mol Genet,* 26**,** 2577-2588.

NORD, A. S., LEE, M., KING, M. C. & WALSH, T. 2011. Accurate and exact CNV identification from targeted high-throughput sequence data. *BMC Genomics,* 12**,** 184.

NORRIS, A. L., WORKMAN, R. E., FAN, Y., ESHLEMAN, J. R. & TIMP, W. 2016. Nanopore sequencing detecs structural variants in cancer. *Cancer Biol Ther,* 17**,** 246-53.

O'DONNELL, C. R., WANG, H. & DUNBAR, W. B. 2013. Error analysis of idealized nanopore sequencing. *Electrophoresis,* 34**,** 2137-2144.

PERRY, G. H., DOMINY, N. J., CLAW, K. G., LEE, A. S., FIEGLER, H., REDON, R., WERNER, J., VILLANEA, F. A., MOUNTAIN, J. L., MISRA, R., et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.,* 39**,** 1256-1260.

PIHLSTROM, B. L., MICHALOWICZ, B. S. & JOHNSON, N. W. 2005. Periodontal diseases. *Lancet,* 366**,** 1809–1820.

PRUFER, K., RACIMO, F., PATTERSON, N., JAY, F., SANKARARAMAN, S., SAWYER, S., HEINZE, A., RENAUD, G., SUDMANT, P. H., DE FILIPPO, C., et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *nature,* 2**,** 43-9.

PUKLO, M., GUENTSCH, A., HIEMSTRA, P. S., EICK, S. & POTEMPA, J. 2008. Analysis of neutrophil-derived antimicrobial peptides in gingival crevicular fluid suggests importance of cathelicidin ll-37 in the innate immune response against periodontogenic bacteria. *Oral Microbiology and Immunology,* 23**,** 328–335.

RACIMO, F., SANKARARAMAN, S., NIELSEN, R. & HUERTA-SÁNCHEZ, E. 2015. Evidence for archaic adaptive introgression in humans. *Nat Rev Genet,* 16**,** 359-71.

REICH, D., THANGARAJ, K., PATTERSON, N., PRICE, A. L. & SINGH, L. 2009. Reconstructing Indian population history. *Nature,* 461**,** 489-94.

ROACH, J. C., GLUSMAN, G., SMIT, A. F. A., HUFF, C. D., HUBLEY, R., SHANNON, P. T., ROWEN, L., PANT, K. P., GOODMAN, N., BAMSHAD, M., et al. 2010.

Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science,* 328**,** 636-639.

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., et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature,* 419**,** 832-837.

SANCHEZ-QUINTO, F. & LALUEZA-FOX, C. 2015. Almost 20 years of Neanderthal palaeogenetics: adaptation, admixture, diversity, demography and extinction. *Philos Trans R Soc Lond B Biol Sci,* 370**,** 20130374.

SANDERS, J. T. & WYATT, R. J. 2008. IgA nephropathy and Henoch-Schönlein nephritis. *Curr Opin Pediatr,* 20**,** 163-70.

SANGER, F. & COULSON, A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol,* 94**,** 441-8.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA,* 74**,** 5463-5467.

SANKARARAMAN, S., MALLICK, S., DANNEMANN, M., PRUFER, K., KELSO, J., PAABO, S., PATTERSON, N. & REICH, D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature,* 507**,** 354-7.

SANKARARAMAN, S., MALLICK, S., PATTERSON, N. & REICH, D. 2016. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr Biol,* 26**,** 1241-7.

SANKARARAMAN, S., PATTERSON, N., LI, H., PAABO, S. & REICH, D. 2012. The date of interbreeding between Neandertals and modern humans. *PLoS Genet,* 8**,** e1002947.

SAVAGE, A., EATON, K. A., MOLES, D. R. & NEEDLEMAN, I. 2009. A systematic review of definitions of periodontitis and methods that have been used to identify this disease. *Journal of Clinical Periodontology,* 36**,** 458–67.

SCALLY, A., DUTHEIL, J. Y., HILLIER, L. W., JORDAN, G. E., GOODHEAD, I., HERRERO, J., HOBOLTH, A., LAPPALAINEN, T., MAILUND, T., MARQUES-BONET, T., et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature,* 483**,** 169-75.

SCHAEFER, A. S., JOCHENS, A., DOMMISCH, H., GRAETZ, C., JOCKEL-SCHNEIDER, Y., HARKS, I., STAUFENBIEL, I., MEYLE, J., EICKHOLZ, P., FOLWACZNY, M., et al. 2014. A large candidate-gene association study suggests genetic variants at IRF5 and PRDM1 to be associated with aggressive periodontitis. *J Clin Periodontol,* 41**,** 1122-31.

SCHAEFER, A. S., RICHTER, G. M., NOTHNAGEL, M., MANKE, T., DOMMISCH, H., JACOBS, G., ARLT, A., ROSENSTIEL, P., NOACK, B., GROESSNER-SCHREIBER, B., et al. 2010. A genome-wide association study identifies GLT6D1 as a susceptibility locus for periodontitis. *Hum Mol Genet,* 19**,** 553-62.

SCHENA, F. P. 1990. A retrospective analysis of the natural history of primary IgA nephropathy worldwide. *Am J Med,* 89**,** 209–215.

SCHOUTEN, J. P., MCELGUNN, C. J., WAAIJER, R., ZWIJNENBURG, D., DIEPVENS, F. & PALS, G. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.,* 30**,** e57.

SHENDURE, J. & JI, H. 2008. Next-generation DNA sequencing. *Nature Biotechnology,* 26**,** 1135-1145.

SHIMIZU, S., MOMOZAWA, Y., TAKAHASHI, A., NAGASAWA, T., ASHIKAWA, K., TERADA, Y., IZUMI, Y., KOBAYASHI, H., TSUJI, M., KUBO, M., et al. 2015. A genome-wide association study of periodontitis in a Japanese population. *J Dent Res,* 94**,** 555-61.

SOARES, P., ALSHAMALI, F., PEREIRA, J. B., FERNANDES, V., SILVA, N. M., AFONSO, C., COSTA, M. D., MUSILOVA, E., MACAULAY, V., RICHARDS, M. B., et al. 2012. The Expansion of mtDNA Haplogroup L3 within and out of Africa. *Mol Biol Evol,* 29**,** 915-27.

SUZUKI, K., HONDA, K., TANABE, K., TOMA, H., NIHEI, H. & YAMAGUCHI, Y. 2003. Incidence of latent mesangial IgA deposition in renal allograft donors in Japan. *Kidney Int,* 63**,** 2286–2294.

TAJIMA, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* 123**,** 585–95.

TAKAHASHI, K., SMITH, A. D., POULSEN, K., KILIAN, M., JULIAN, B. A., MESTECKY, J., NOVAK, J. & RENFROW, M. B. 2012. Naturally occurring structural isomers in serum IgA1 O-glycosylation. *J Proteome Res,* 11**,** 692-702.

TANG, Y. Q., YUAN, J., OSAPAY, G., OSAPAY, K., TRAN, D., MILLER, C. J., OUELLETTE, A. J. & SELSTED, M. E. 1999. A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated alpha-defensins. *Science,* 286**,** 498-502.

TEUMER, A., HOLTFRETER, B., VÖLKER, U., PETERSMANN, A., NAUCK, M., BIFFAR, R., VÖLZKE, H., KROEMER, H. K., MEISEL, P., HOMUTH, G., et al. 2013. Genome-wide association study of chronic periodontitis in a general German population. *J Clin Periodontol,* 40**,** 977-85.

THE 1000 GENOMES PROJECT CONSORTIUM 2010. A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-1073.

THE 1000 GENOMES PROJECT CONSORTIUM 2015. A global reference for human genetic variation. *Nature,* 526**,** 68–74.

THE INTERNATIONAL HAPMAP CONSORTIUM 2003. The International HapMap Project. *Nature,* 426**,** 789-796.

TOMANA, M., NOVAK, J., JULIAN, B. A., MATOUSOVIC, K., KONECNY, K. & MESTECKY, J. 1999. Circulating immune complexes in IgA nephropathy consist of IgA1 with galactose-deficient hinge region and antiglycan antibodies. *J Clin Invest,* 104**,** 73-81.

TSUCHIDA, S., SATOH, M., TAKIWAKI, M. & NOMURA, F. 2017. Ubiquitination in Periodontal Disease: A Review. *Int J Mol Sci,* 18.

TURNER, D. J. & HURLES, M. E. 2009. High-throughput haplotype determination over long distances by haplotype fusion PCR and ligation haplotyping. *Nat Protoc,* 4**,** 1771–1783.

TYSON, J. & ARMOUR, J. A. L. 2012. Determination of haplotypes at structurally complex regions using emulsion haplotype fusion PCR. *BMC Genomics,* 13.

ULLMANN, R., TURNER, G., KIRCHHOFF, M., CHEN, W., TONGE, B., ROSENBERG, C. & ET AL 2007. Array CGH Identifies reciprocal 16p13.1 duplications and deletions that predispose to autism and /or mental retardation. *Hum Mutat,* 28**,** 674–682.

UNTERGASSER, A., CUTCUTACHE, I., KORESSAAR, T., YE, J., FAIRCLOTH, B. C., REMM, M. & ROZEN, S. G. 2012. Primer3–new capabilities and interfaces. *Nucleic Acids Res,* 40**,** e115.

VARIS J, R. I., PASTERNACK A, OKSA H, JANTTI M, ET AL 1993. Immunoglobulin and complement deposition in glomeruli of 756 subjects who had committed suicide or met with a violent death. *J Clin Pathol,* 46**,** 607–610.

VERNOT, B. & AKEY, J. M. 2014. Resurrecting surviving Neandertal lineages from modern human genomes. *science,* 343**,** 1017-21.

VERNOT, B., TUCCI, S., KELSO, J., SCHRAIBER, J. G., WOLF, A. B., GITTELMAN, R. M., DANNEMANN, M., GROTE, S., MCCOY, R. C., NORTON, H., et al. 2016. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science,* 352**,** 235-9.

VOIGHT, B. F., KUDARAVALLI, S., WEN, X. & PRITCHARD, J. K. 2006. A map of recent positive selection in the human genome. *PLoS Genet,* 4**,** e72.

VOS, T., FLAXMAN, A. D., NAGHAVI, M., LOZANO, R., MICHAUD, C., EZZATI, M., SHIBUYA, K., SALOMON, J. A., ABDALLA, S., ABOYANS, V., et al. 2012. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet,* 380**,** 2163–96.

WEIR, B. S. & COCKERHAM, C. C. 1984. Estimating F-statistics for the analysis of population structure. *Evolution,* 38**,** 1358-1370.

WHITE, S. H., WIMLEY, W. C. & SELSTED, M. E. 1995. Structure, Function, and Membrane Integration of Defensins. *Current Opinion in Structural Biology,* 5**,** 521-527.

WYATT, R. & JULIAN, B. 2013. IgA nephropathy. *N Engl J Med,* 368**,** 2402-2414.

XU, R., FENG, S., LI, Z., FU, Y., YIN, P., AI, Z., LIU, W., YU, X. & LI, M. 2014. Polymorphism of DEFA in Chinese Han population with IgA nephropathy. *Hum Genet,* 133**,** 1299-309.

YU, X. Q., LI, M., ZHANG, H., LOW, H. Q., WEI, X., WANG, J. Q., SUN, L. D., SIM, K. S., LI, Y. & FOO, J. N. 2012. A genome-wide association study in Han

Chinese identifies multiple susceptibility loci for IgA nephropathy. *Nat Genet,* 44**,** 178-182.

ZARREI, M., MACDONALD, J. R., MERICO, D. & SCHERER, S. W. 2015. A copy number variation map of the human genome. *Nature Reviews Genetics,* 16**,** 172–183.