Allelic structures and mechanisms of copy number change at the human *DEFA1A3* copy number variable locus

Holly Ann Black, BSc.

Thesis submitted to The University of Nottingham

for the degree of Doctor of Philosophy

December 2014

Abstract

The *DEFA1A3* locus on human chromosome 8p23.1 exhibits extensive copy number variation; individuals have between 3-16 copies of *DEFA1A3*. The region has additional complexity in that each repeat unit contains a gene locus that can be occupied by one of two different genes, *DEFA1* or *DEFA3*. These encode the human neutrophil peptides (HNPs) 1-3, antimicrobial peptides involved in the innate immune response. In order to understand the mutational processes and evolutionary history of a complex locus like *DEFA1A3*, spatial information is essential. Whilst haplotype *DEFA1A3* copy numbers and haplotype ratios of *DEFA1 vs. DEFA3* have been determined, little is known about the features shared by, and the structures of, related haplotypes.

In this study, flanking sequence variation has been used to identify five classes of *DEFA1A3* haplotype, which are tagged by four SNPs. Haplotypes within each class share similar features, such as *DEFA1A3* copy number, but the associations differ between-class and between-population. Emulsion haplotype fusion-PCR has been used to determine the spatial arrangement of the *DEFA1* and *DEFA3* genes, as well as additional internal variants, across haplotypes of European ancestry. A comparison of the structures of related haplotypes suggests that the predominant mechanism of copy number change at the *DEFA1A3* locus is intra-allelic rearrangements (i.e. between haplotypes from the same class), facilitated by the high sequence similarity of repeat units within each class. This explains the preservation of linkage disequilibrium across

the *DEFA1A3* locus.

The relationship between *DEFA1A3* copy number and gene expression is unclear. A comparison between *DEFA1A3* haplotype class and HNP1-3 expression in a UK cohort suggests that *DEFA1A3* haplotype structure does not influence gene expression. However, the identification of four SNPs which tag *DEFA1A3* haplotype class and, in turn, haplotype structure in haplotypes of European ancestry, will aid further studies in this area.

Publications

Black, H. A., Khan, F. F., Tyson, J., and Armour, J. A. L. (2014). Inferring mechanisms of copy number change from haplotype structures at the human *DEFA1A3* locus. Submitted to BMC Genomics.

Khan, F. F., Carpenter, D., Mitchell, L., Mansouri, O., Black, H. A., Tyson, J., and Armour, J. A. L. (2013). Accurate measurement of gene copy number for human alpha-defensin *DEFA1A3*. BMC Genomics 14, 719.

Acknowledgements

Firstly, I would like to thank my supervisor, John Armour, for giving me the opportunity to do my PhD in his research group. His guidance and support have enabled me to complete a project I am very proud of, which has been thoroughly enjoyable and informative. I am also grateful to the BBSRC and the University of Nottingham for my studentship, without which it would not have been possible to complete my PhD.

I would like to thank everyone I have worked with in the JALA lab- Jess Tyson, Danielle Carpenter, Laura Mitchell, Fayeza Khan, Omniah Mansouri, Tamsin Majerus, Somwang Janyakhantikul, Raquel Palla, Sugandha Dhar, Dibo Pughikumo, Ivan Stetsenko, Nzar Shwan and Xiao Xu. A special thanks goes to Jess, for always being there to listen and advise. Thank you to all of my friends who have supported me, in particular those in the Nottingham Wind Ensemble and at 1^{st} Trowell Guides.

Last, but by no means least, I would like to thank my family, especially my parents John and Shirley. Words cannot describe the love, support, advice and help they have given me and it is this which has enabled me to be successful in completing my studies.

I draw inspiration from many sources and would like to end my acknowledgements by thanking: Delta Goodrem, Andrew Lloyd Webber, Michael Ball, Jonny Wilkinson, Roger Federer, Leicester Tigers, England Rugby, the British and Irish Lions, Preston North End FC, the 2012 Ryder Cup Team and Team GB.

Contents

List of Figures

List of Tables

List of Abbreviations

BIR	Break-induced replication
BSA	Bovine serum albumin
CEPH	Centre d'Etude du Polymorphisme Humain
CN	Copy number
CNV	Copy number variable
CNVR	Copy number variable region
dH_2O	Distilled H_2O
dNTP	Deoxynucleotide Triphosphate
DSB	Double-strand break
ECACC	European collection of animal cell cultures
EHF-PCR	Emulsion haplotype fusion-PCR
EHH	Extended haplotype homozygosity
FISH	Fluorescent *in situ* hybridisation
FoSTeS	Fork stalling and template switching
FSHD	Facioscapulohumeral muscular dystrophy
GWAS	Genome-wide association study
HD	Human defensin
HNP	Human neutrophil peptide
HR	Homologous recombination
HRC	Human random control
iHH	Integrated extended haplotype homozygosity
iHS	Integrated haplotype score

Indel	Insertion-deletion polymorphism
LD	Linkage disequilibrium
MAPH	Multiplex amplifiable probe hybridisation
MAF	Minor allele frequency
MLCN	Maximum likelihood copy number
MLPA	Multiplex ligation-dependent probe amplification
MMBIR	Microhomology-mediated break-induced replication
MMEJ	Microhomology-mediated end joining
NAHR	Non-allelic homologous recombination
NGS	Next-generation sequencing
NHEJ	Non-homologous end joining
PCR	Polymerase chain reaction
PFGE	Pulsed-field gel electrophoresis
PRT	Paralogue ratio test
PSV	Partial repeat-specific variant
RFLP	Restriction fragment length polymorphism
SD	Segmental duplication
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
TBE	Tris-Borate-EDTA

1 Introduction

1.1 Human Genetic Variation

The completion of the Human Genome Project in 2001 provided the first insight into the sequence of the human genome and the location of the genes within it [1]. Since then, the focus has turned to the identification and characterisation of variants found within the sequence [2]. Variation in the genome has been found to contribute to a wide range of phenotypic variability, including susceptibility to rare and common disease, as well as providing insight into evolutionary processes and population histories [3–9].

Types of genetic variation

There are many different types of genetic variation, usually categorised by their size. Single nucleotide variants (SNVs) involve the substitution of a single base of DNA with an alternative base; SNVs with a minor allele frequency of greater than 1% are referred to as single nucleotide polymorphisms (SNPs) [10]. The effect of SNVs on phenotype can be variable. The SNV may fall within non-coding regions, where they are expected to have little or no effect on phenotype, although it is possible that they could disrupt promoter or enhancer elements or splice sites. If the SNV falls within a coding region, it is possible that the mutation will be synonymous- i.e. will not change the amino acid coded for- due to the

redundancy of the genetic code. Alternatively, the mutation may be non-synonymous. This would result in either a missense mutation, where the amino acid coded for is changed, or a nonsense mutation, where a premature stop codon is introduced [3]. SNVs are well-characterised, with over 44 million SNPs now listed in dbSNP [11]. Due to the ease of genotyping SNPs on a large scale, genome-wide association studies (GWAS) have been able to genotype millions of SNPs in cases and controls to identify regions of the genome associated with disease [4, 5]. However, many of the reported associations do not identify the causative variant; therefore, it is difficult to determine the mechanistic basis of the association. In addition, the associations are sometimes weak and do not account for the entire heritability of the disease [4, 12]

Structural variants encompass mutations involving more than a single base pair of DNA. Insertion-deletion polymorphisms (Indels) involve the deletion, insertion or duplication of a region of DNA less than 1kb in length [13]. They are expected to have more serious consequences on phenotype than SNVs, especially if the Indel falls within a coding region, where there will always be a change in the resulting protein, which will be more severe if the mutation leads to a frame shift [14]. At the larger end of the scale are copy number variants, defined as the deletion, duplication or insertion of a region of DNA at least 1kb in length [9]. Microsatellites are a form of tandem repeat, in which a sequence of 2-10bp is repeated a variable number of times. They are widely distributed in the human genome. Microsatellites can lead to frameshift mutations and alter coding or regulatory sequences. However, they are also thought to perform a regulatory function in the human genome [15–17].

The International HapMap project

The International HapMap Project aimed to determine the common patterns of sequence variation in the human genome [10]. Over 1 million SNPs were genotyped in individuals from four different populations with ancestry in Europe, Asia and Africa. This allowed the identification of associations between variants as well as the differences in these associations between populations. As many SNPs will have evolved on the same haplotype background and because of the loss of intermediates via drift, they will display linkage disequilibrium (LD), meaning that they are inherited together and display population-level allelic association. Therefore, very few SNPs are needed to tag the different haplotypes found across a LD block, limiting the number of SNPs that need to be genotyped to allow an inference of the genome-wide variation. This information is useful for disease association studies and for investigating population histories [18].

1000 Genomes Project

The 1000 Genomes Project aims to characterise the majority of polymorphic variants in the human genome across different worldwide populations. Using next-generation sequencing (NGS) technologies to sequence whole genomes and whole exomes from 2500 individuals, it aims to identify at least 95% of sequence variants present at a frequency of at least 1% and coding variants present a frequency of at least 0.1% [19]. The project includes individuals with European, Asian, African and American ancestry. The pilot project involved the sequencing of 1092 individuals and identified 38 million SNPs, 1.4 million Indels and 14,000

copy number variants [20]. Once complete, the 1000 Genomes Project will provide a comprehensive catalogue of worldwide variation, allowing within- and between-population comparisons of variation.

1.2 Copy Number Variation

Copy number variation is defined as the deletion, insertion or duplication of a region of DNA $\geq$ 1kb in length (figure 1.1) [13, 21, 22]. Over 100,000 copy number variants have been identified in the human genome [23], affecting a greater number of base pairs than those covered by SNPs and including many genes and functional elements [6, 24]. However, the place of minisatellites and mobile elements in this definition is difficult to clarify [22]. Some loci are subject to non-recurrent changes in copy number. The presence of an allele in the population carrying a deletion will lead to the affected region being present in 0-2 copies per diploid genome, whereas the presence of an allele carrying a duplication will lead to the affected region being present in 2-4 copies. These are termed simple copy number variants. However, some loci experience multiple deletion and duplication events, leading to the affected region being present in a highly variable number of copies (figure 1.2). These are known as multiallelic copy number variants and are usually mediated by segmental duplications (SDs), which are regions of DNA $\geq$ 1kb in length with $\geq$ 95% sequence identity [22, 25]. The mutation rate at copy number variable (CNV) loci is estimated to be higher than that for SNPs. The mutation rate per CNV locus per generation is estimated to range from 1.7x10^{-6} - 1x10^{-4} [8, 26], whereas for point mutations the rate is 1.8-2.5 x 10^{-8} per base per generation [27, 28]. For this reason, multiallelic copy number variants are rarely tagged by SNPs [29].

Deletions in the genome tend to encompass gene-free regions, due to negative selection on a variant which reduces gene dosage. However, duplication events frequently include genes, suggesting an evolutionary advantage in which gene duplication may allow the development of novel functions [24]. In addition, certain gene families are enriched for copy number variants. These include genes involved in the immune response, metabolism, cell adhesion, sensory perception and neurotransmission [21, 24, 30–32], which are presumably less sensitive to copy number change and for which variation in gene copy number may confer an advantage.

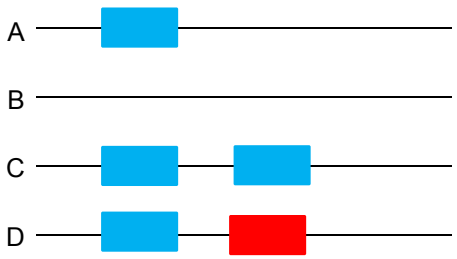


Figure 1.1: There are various types of copy number variation. Relative to a reference sequence shown in A, copy number variation can involve a deletion (B), duplication (C) or insertion (D) of a region of DNA.

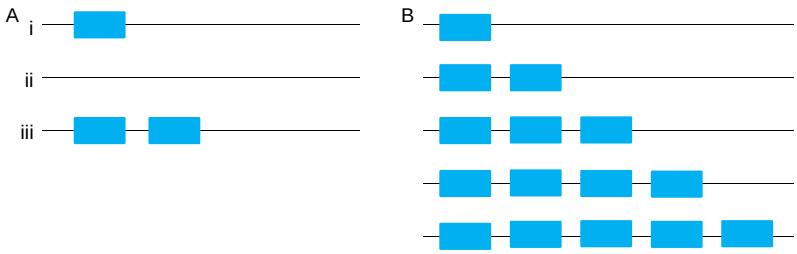


Figure 1.2: A: A simple copy number variant involves either a single deletion or duplication event. With respect to the reference shown in i, ii shows a deletion and iii a duplication of the region. B: At multiallelic CNV loci, the variable region can be present in a wide range of different copy numbers; for example, the figure shows a locus with between 1-5 copies of a CNV region per haplotype.

Mechanisms of Copy Number Variant Formation

Various mechanisms of copy number variant formation have been proposed. The mechanism of formation usually depends on whether the copy number variant is simple or multiallelic, with simple copy number variants formed through non-recurrent mechanisms and multiallelic copy number variants through recurrent mechanisms [33].

Non-homologous end joining

Non-homologous end joining (NHEJ) is a non-recurrent mechanism through which simple copy number variants are formed and is one pathway in the double strand break (DSB) repair process [33]. It does not require a homologous template to repair the break; instead the break is repaired via the insertion or deletion of nucleotides to create microhomologies, which join the broken strands [25]. Occasionally, the repair is accurate, but in most cases, it leads to the loss or gain of a region of DNA [8, 33]. Another mechanism, microhomology-mediated end joining (MMEJ) has recently been proposed as an alternative form of NHEJ, in which the ends of a double strand break become exposed as single-strand sequences, revealing 5-25bp of sequence, which allow for microhomologies to anneal strands and repair across the break. In this case, there is always a deletion of the sequence between the regions of microhomology [25]. If the broken strands join with non-homologous sequence, NHEJ and MMEJ can lead to chromosomal rearrangements, such as deletions, duplications and translocations [25].

Non-allelic homologous recombination

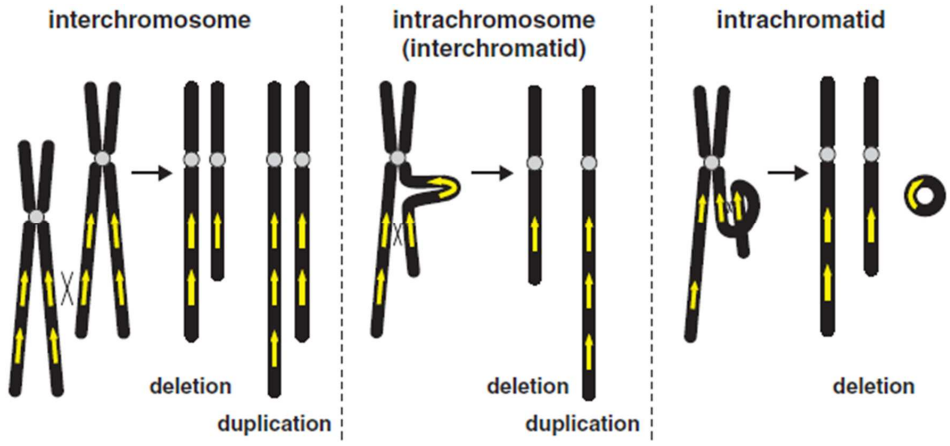


Figure 1.3: NAHR can occur interchromasomally between homologous chromosomes, resulting in a reciprocal deletion and duplication. It can also occur intrachromasomally, either between sister chromatids, leading to a deletion and a duplication, or intrachromatid, resulting in a deletion event only. Figure from Liu *et al.* [34].

Homologous recombination (HR) is another DSB repair pathway. In this pathway, the nucleotides from the 5' ends of a DSB are removed to leave 3' overhangs, which can invade a homologue. DNA synthesis is then able to repair the gap. Depending on the resolution of the complex, it can lead to either gene conversion or crossover [25]. Non-allelic homologous recombination (NAHR) is the process by which HR results in a change in copy number (figure 1.3). If the invasion occurs between non-allelic copies of a sequence (i.e. SDs), it leads to either the inversion of the intervening sequence if the repeats were in an inverted orientation or deletions and duplications if the repeats were in direct orientation [8, 34]. Therefore, this process leads to recurrent changes in copy number. Break induced replication (BIR) is another form of HR which repairs DSBs at replication forks. This can again result in deletion and duplication events if the invading strand joins to a homologue in a different

chromosomal position [25].

Gene Conversion

Gene conversion is not a mechanism which creates copy number variation, but can maintain the ability of a locus to undergo recurrent mutations. NAHR relies on SDs having a high sequence identity and the process of gene conversion is able to maintain this similarity [35]. It does so via the non-reciprocal copying of DNA from one copy of a region to another over a length of 200bp-1kb, with no change in the donor sequence, but a loss of the recipient sequence (figure 1.4) [36]. This leads to the homogenisation of repeat units. Examples of gene conversion have been observed previously at CNV loci. At the *FCGR3B* locus, there is evidence of ancestral sequence exchanges between two regions containing different paralagous sequence variants [37]. At the *DEFA1A3* locus, a flanking region, highly similar in sequence to the variable repeats, has been shown to exchange sequence with the repeats in an event termed the telomeric replacement polymorphism ([38], Fayeza Khan, personal communication).

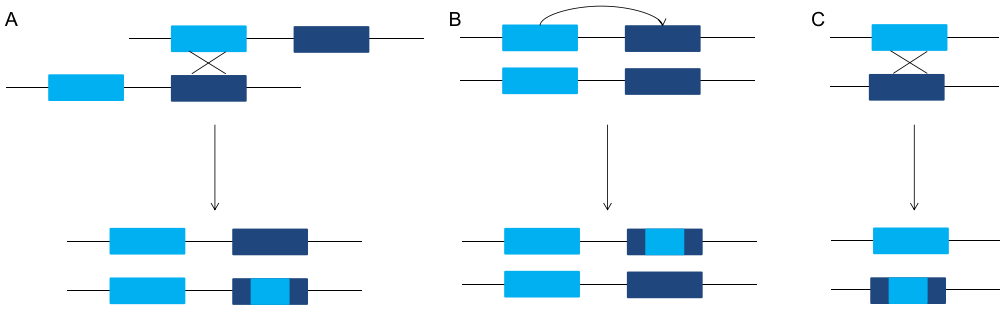


Figure 1.4: Gene conversion involves the non-reciprocal exchange of DNA between two loci. This can be between two non-allelic gene copies either on different chromatids (A) or on the same chromatid (B) or between allelic copies on homologous chromosomes (C). Adapted from Chen *et al.* [36].

Replicative mechanisms of copy number variant formation

Some copy number variants are formed not through the repair of DSBs, but during DNA replication. One such mechanism is replication slippage; when a replication fork encounters short duplicated regions with high sequence identity, replication slippage can occur, leading to a duplication or deletion of the intervening region [25].

Fork stalling and template switching (FoSTeS) is another replicative mechanism through which non-recurrent copy number variants can be formed. During DNA replication, the replication fork can stall and the lagging strand can disengage and invade nearby templates sharing microhomology, which are short stretches of complementary sequence [39]. This can occur multiple times in series until the replication fork proceeds normally [8]. Depending on the direction and orientation of invasion, this can lead to duplications, deletions and inversions, as well as complex rearrangements if it occurs several times in series [8]. A similar mechanism has been proposed which is based upon the BIR mechanism of HR. This is microhomology mediated break-induced replication (MMBIR) and can also result in complex rearrangements. It involves strand invasion to sequences with microhomology, which may not be the homologous copy of that sequence [40].

Measuring Copy Number

Various different methods have been applied to the measurement of copy number variants, which is technically challenging at multiallelic loci. Multiplex Amplifiable Probe Hybridisation (MAPH) and Multiplex Ligation-dependent Probe Amplification (MLPA) allow the simultaneous measure-

ment of up to 100 loci [41–43]. Both involve the binding of sequence-specific probes to DNA and subsequent amplification of the bound probes, allowing relative quantification of the number of probes bound to each locus. Hence, this allows the quantification of the regions of DNA to which the probes were bound [41, 42, 44].

Quantitative PCR (qPCR) is technique that has been widely used to measure the copy number of a single locus. This compares the relative intensity of amplification from a two-copy reference locus to a CNV test locus, each amplified using a different pair of primers, in a process suitable for high-throughput analysis [45, 46]. However, the two primer pairs used for qPCR will have different amplification efficiencies and, although repeat measurements for the same sample can be consistent, it appears that measurement is influenced by factors which lead to an incorrect assignment of copy number [45–47]. The paralogue ratio test (PRT) appears to overcome these issues, in that it takes advantage of dispersed repeat sequences in the genome to amplify a two-copy reference locus and the CNV test locus using a single pair of primers, reducing the differences in amplification properties of the two loci [48]. This has been successfully applied to the measurement of several multiallelic copy number variants: *CCL3L1/CCL4L1* [49], *DEFB4* [48], *DEFA1A3* [50], *FCGR3B* [51] and *AMY1* [52]. Another method for measuring copy number is the use of read depth from NGS data. In theory, the number of sequence reads mapping to a locus should be proportional to its copy number, allowing the detection of deletion and duplication events [46, 53].

At multiallelic CNV loci, it may be necessary to not only know the total number of copies of a region, but the number of copies per haplotype. Fiber fluorescent *in situ* hybridisation (FISH) has been used to achieve

this at the *AMY1* locus [52, 54], but the technique requires the use of specialist equipment and is low-throughput [46]. Pulsed-field gel electrophoresis (PFGE) followed by southern blotting has been used to measure the haplotype copy numbers of *DEFA1A3*, following restriction digestion with the enzyme *Hpa*I, which cuts the entire *DEFA1A3* locus out in a single restriction fragment [55]. Again, this is a low-throughput technique. Segregation in three-generation families has been successfully used to deduce haplotype copy numbers at multiallelic CNV loci [50, 56]; however, a family resource is not always available.

Phenotypic consequences of copy number variation

Copy number variation can have a wide range of effects on gene expression (figure 1.5). If a gene is CNV, it would be expected that expression from the gene would increase as gene copy number increased. However, this is not always the case and there are examples of a negative correlation between gene copy number and expression in lymphoblastoid cell lines [57]. In addition, even in cases where expression increases with copy number, the increase is not always proportional to the number of copies [6]- i.e. an increase from 2 to 3 copies does not always result in 50% more protein. It is important to note that an increase in mRNA expression will not always result in an increase in protein, due to additional regulation at the translation stage of gene expression [6]. An excess of protein may lead to an increase or decrease in pathway activity or result in off-target effects, for example [58]. However, not all genes will be dosage sensitive and a change in activity may not result in a phenotypic change [13].

Various copy number variants have been associated with human dis-

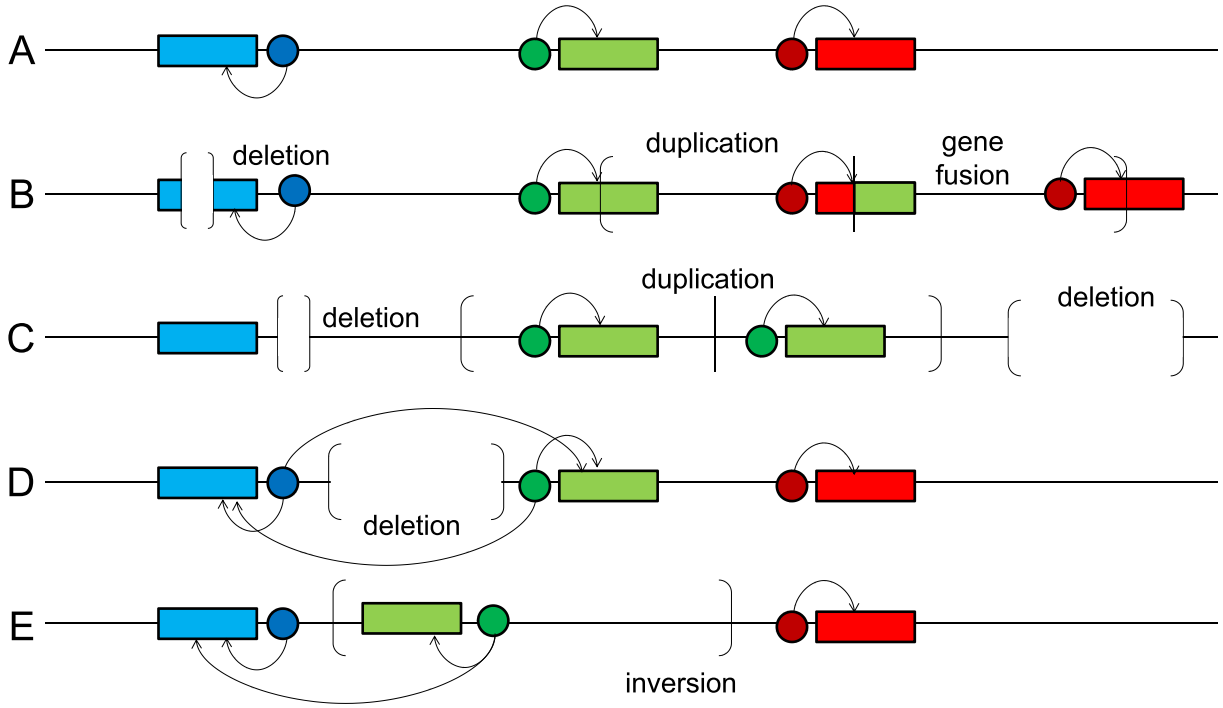


Figure 1.5: There are many possible effects of copy number variation. A: A genomic region without structural variants containing three genes (boxes) each controlled by a different enhancer element (circles). B: Intragenic rearrangements could lead to the deletion or duplication of all or part of a gene, leading to possible gene fusions and resulting in aberrant expression. C: Deletion or duplication events can involve a whole gene, potentially leading to increased or decreased expression. Deletions may also remove enhancer elements, leading to a change in the location or timing of expression. Deletions of either a gene or an enhancer could also unmask recessive mutations in the unaffected haplotype. D and E: Copy number variants may result in positional effects. For example, a deletion may bring together two enhancer elements, changing the location, timing or level of gene expression (D) as may an inversion (E). Adapted from a figure by Weischenfeldt *et al.* [6].

ease, ranging from rare, Mendelian to common, complex diseases [6]. For example, a duplication of the *PMP22* gene has been associated with Charcot-Marie-Tooth Type 1A. Duplication results in increased protein production, which causes the neuropathy phenotype [59]. Another dosage-sensitive gene is *RAI1*, which, when duplicated causes Potocki-Lupski syndrome, yet when deleted causes Smith-Magenis syndrome, due to haploinsufficiency [9]. An increased dosage of the *SLC2A3* gene has been shown to delay the age of onset of Huntington's disease, by increasing glucose uptake in the brain, which is usually reduced in Huntington's disease [60]. Associations with disease have also been observed at multiallelic copy number variable regions (CNVRs). For example, a low copy number of *FCGR3B* has been associated with systemic lupus erythematosus (SLE) [61]. However, in this instance, it is the aberrant transcription of the chimeric *FCGR2B'* gene in natural killer cells, resulting from a zero-copy *FCGR3B* haplotype, that may be the key factor in SLE risk [62]. A low copy number of the *CCL3L1* gene has been associated with an increased susceptibility to HIV/AIDS [63] and a high copy number associated with a susceptibility to Rheumatoid arthritis [64], although both associations have been disputed, due to inaccuracies in copy number measurement in the initial studies [65–67]. A high β-defensin copy number has been associated with an increased risk of Psoriasis [68, 69]. Two independent studies found associations between both a high [70] or low [71] β-defensin copy number and Crohn's disease risk, neither of which were replicated in further studies [45, 72]. Not all copy number changes resulting in an altered phenotype will have a negative effect and some copy number variants are thought to confer an adaptive advantage [58]. For example, an increase in *AMY1* copy number, which correlates with an increase in salivary amylase expression,

has been observed in populations with high starch diets [73]. However, many studies looking for associations between multiallelic CNV loci and human disease fail to combine accurate copy number measurement with information about the resulting change in expression; hence, few reproducible associations have been reported.

1.3 Defensins

There are three families of mammalian defensin genes, α-, β- and θ-defensins, which evolved from a single precursor gene and all encode antimicrobial peptides [74]. Defensin-like proteins are found throughout animal and plant species [75], with the β-defensins found in all vertebrate species [76] and the α-defensins, which evolved from two different β-defensin genes [76], found only in mammals [77]. The θ-defensins are specific to primates [74] and evolved from α-defensins [76, 78, 79], although in humans, all θ-defensin genes contain a premature stop codon [79]. The variability in the number of genes between species and the high number of pseudogenes observed within the defensin family suggests frequent duplication events allow the defensins to adapt to a changing exposure to pathogens [79–82].

α-defensins

In humans, there are five different α-defensin genes, found in a cluster on chromosome 8p23.1 (figure 1.6) [83], one of which, *DEFA3*, is human-specific [55]. The genes *DEFA1* and *DEFA3* are CNV [50, 55, 84, 85] and for each copy of these genes, there is also a copy of either *DEFT1P*, a θ-defensin pseudogene, or *DEFA10P*, a α-defensin pseudogene, which

have a high sequence similarity, suggesting they evolved in an ancient duplication event [76, 78]. The five genes express six different antimicrobial peptides [86]. *DEFA5* and *DEFA6* express the human defensins, HD-5 and HD-6 respectively, in the Paneth cells of the small intestine [87, 88]. *DEFA1*, *DEFA3* and *DEFA4* produce the human neutrophil peptides, HNP1-4 [89]. These are expressed in promyelocytes [90], but are stored in the azurophil granules of neutrophils. HNP-2 does not have a corresponding gene, but is thought to represent a cleavage product of either *DEFA1* and/or *DEFA3* [84]. HNP1-3 are highly similar in their sequence, differing only in the N-terminal amino acid of the mature peptide, which is alanine in HNP-1, aspartic acid in HNP-3 and absent in HNP-2 [91, 92].

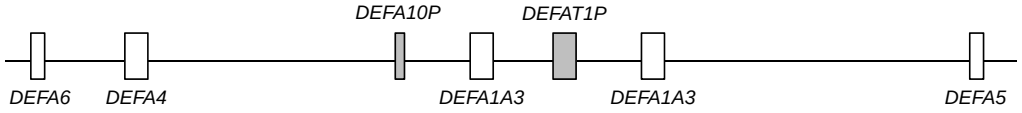


Figure 1.6: The human α-defensins are found in a cluster on chromosome 8p23.1. A haplotype is shown with two copies of the *DEFA1A3* locus, a CNV region that can be occupied by either the *DEFA1* or *DEFA3* genes; the other three genes are copy number constant [93]. The gene *DEFA10P* is an α-defensin pseudogene, which shares high sequence similarity with *DEFT1P*, the θ-defensin pseudogene found within the same cluster [76, 78].

Human Neutrophil Peptides in the Immune Response

The HNPs are expressed as prepropeptides, which are neutral in charge [90]. Subsequent cleavage of a signal peptide and an anionic prosegment produce an active 29-30 amino acid cationic antimicrobial peptide [92]. These cleavage steps are important in ensuring the peptides are not toxic to host cells, as it is the cationic nature of the HNPs which allow

them to exert their antimicrobial properties [94].

The HNPs are initially involved in the innate immune response, where they function directly as antimicrobial peptides [86]. They are stored at high concentrations in the azurophil granules of neutrophils (50% azurophil granule protein content [95]), which are phagocytic cells that engulf invading microorganisms. The azurophil granules fuse with vesicles containing the invading pathogen, exposing them to high concentrations of HNPs. The HNPs can create pores in the membranes of invading microorganisms, due to their cationic and hydrophobic properties, leading to leakage of cellular contents, disruption of membrane potential and, ultimately, cell death [75, 95–97]. This allows HNPs to actively degrade gram positive and gram negative bacteria, fungi and enveloped viruses; however, they are inactive against non-enveloped viruses [89, 95, 98, 99]. The HNPs are one of many antimicrobial components of neutrophils and form part of a multi-faceted innate immune response.

Secondly, the HNPs activate the adaptive immune response. They have been shown to regulate the complement system, a process which helps antibodies and phagocytic cells to remove pathogens [86, 100]. In addition, they are chemotactic for mast cells, naive dendritic cells, naive T cells and macrophages [86, 100–103]. Mast cells induce inflammation, which is part of the wound healing process, protecting the body from infection [86, 100, 104]. Dendritic cells and macrophages are antigen presenting cells which, through T cells, activate an adaptive immune response [105]. Therefore, the HNPs are involved in many process of the immune response.

Copy number variation at the α-defensin locus

The α-defensins are subject to extensive copy number variation, which is independent of the copy number observed at the β-defensin locus [50, 55, 84, 85]. The structure of the *DEFA1A3* CNVR is shown in figure 1.7. Each repeat can be occupied by either *DEFA1* or *DEFA3*, leading to the designation of the locus as *DEFA1A3* [55]; therefore, both genes vary in copy number. Between 10-37% of individuals, depending on the population studied, do not possess a copy of the *DEFA3* gene [55, 78]; the functional consequences of lacking the *DEFA3* gene and, in turn, HNP-3, are currently unknown [55].

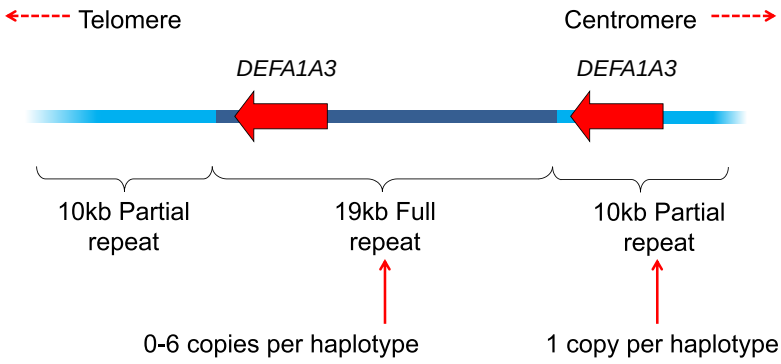


Figure 1.7: The *DEFA1A3* locus consists of two single-copy 10kb partial repeats surrounding a variable number of 19kb full repeats. Each full repeat and the centromeric-most partial repeat contain a gene locus that can be occupied by either *DEFA1* or *DEFA3*; therefore, the locus is referred to as *DEFA1A3*. The full and partial repeats are highly similar in their sequence, but are diverged at particular positions.

Whilst several studies have used qPCR to measure *DEFA1A3* copy number [85, 106, 107], a combination of two PRTs and three allelic ratio assays has provided an accurate and reliable method for copy number measurement of the locus [50]. The diploid copy number range for *DEFA1A3* ranges from 3-16 copies, with a reduced range in the Chinese population, of 3-11 copies [50, 55, 85, 108, 109]. Segregation in

three-generation families of European origin has allowed a haplotype copy number range of 1-7 copies to be determined, with the majority of haplotypes having between 2-5 copies [50]. SNPs tagging copy number at multiallelic CNV loci are rare, due to the fast mutation rate at CNV loci which results in a high variability in copy number status [29]. However, in the European population, the SNP rs4300027 has been identified as a tag of *DEFA1A3* haplotype copy number ($p=1.3x10^{-45}$), but this association has not been replicated in other populations [50].

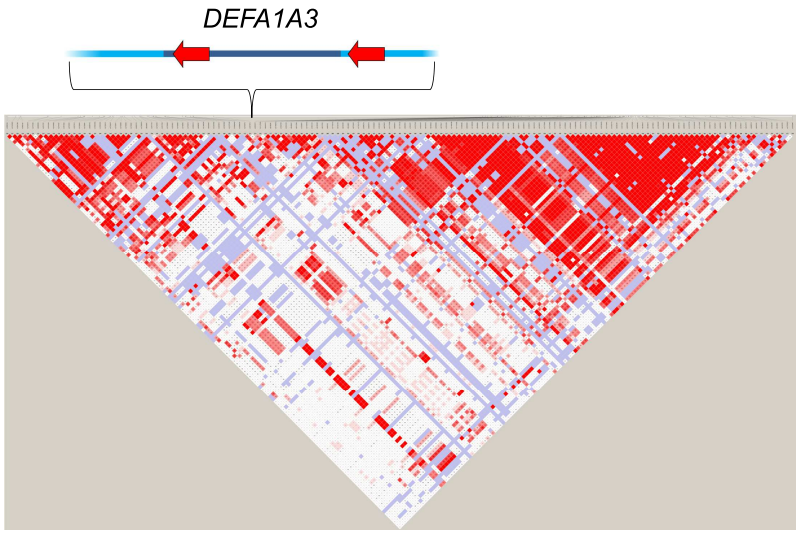


Figure 1.8: SNPs either side of the *DEFA1A3* locus display strong LD ($D'=1$), thus suggesting that recombination events resulting in a crossover are infrequent across the locus. HapMap CEU SNP data downloaded from the International HapMap Project [10].

The *DEFA1A3* locus is intriguing, because there are SNPs either side of the locus displaying strong LD (figure 1.8) [10], suggesting that crossover events within the locus are rare. However, the locus also exhibits extensive copy number variation and, at a locus where the repeat units have such high sequence similarity, it would be expected that mutations changing the *DEFA1A3* copy number of haplotypes would be occurring frequently. Therefore, the mechanism by which copy number change

occurs at the *DEFA1A3* locus is unclear.

The relationship between *DEFA1A3* copy number and HNP1-3 expression remains unclear. Whilst one small-scale study found that *DEFA1A3* copy number was proportional to the amount of HNP1-3 expressed in neutrophils [85], no relationship was observed in a larger second study (Danielle Carpenter, personal communication). Another study looked at mRNA expression in relation to *DEFA1A3* copy number in neutrophils, showing there was not a relationship between total mRNA and *DEFA1A3* copy number, but that the *DEFA1* vs. *DEFA3* ratio was maintained in mRNA expression, suggesting all copies of the array are expressed [55]. However, as *DEFA1A3* mRNA expression occurs in promyelocytes and not neutrophils, it is possible this is not an accurate representation of the profile of mRNA expression from the *DEFA1A3* locus [90]. Several studies have associated *DEFA1A3* copy number with disease, for example Crohn's disease susceptibility and risk of severe sepsis [106, 107]; however, copy number measurement in these studies usually involves qPCR, which has been shown previously to distort associations at CNV loci [45–47]. A recent GWAS in the Han Chinese population identified a SNP within the *DEFA1A3* locus as a risk for IgA nephropathy [110]; however, it is unclear if *DEFA1A3* copy number is responsible for this association.

1.4 Phasing the human genome

In recent years, there has been a huge advance in DNA sequencing technologies, allowing large datasets to be generated at an ever-decreasing cost. However, NGS technologies generally ignore the diploid nature of

the human genome. By obtaining short read sequencing data, the ability to phase the genome post-sequencing is lost, as reads will vary rarely contain more than one heterozygous sequence variant [111]. Therefore, many different approaches have been taken to generate both localised and whole-genome phased sequence information.

There are many reasons why obtaining diplotype information is important. For example, there are many studies aiming to identify disease-causing mutations. If an individual has two loss of function mutations in the same gene, it is important to know whether these are found in *cis* or in *trans*, in which one situation would retain a functional copy of the gene and one would not [111]. In addition, there are examples of where the combination of variants on a haplotype can influence disease risk, disease progression and response to therapeutics [62, 112–115]. Also, it is haplotypes that are the unit of inheritance and, as such, phased information is necessary for evolution and population history studies. For example, the HapMap project has enabled a worldwide study of SNP genotype frequencies, LD patterns, haplotype sharing and diversity and recombination rate variation [18, 116–118], in addition to allowing a genome-wide assessment of evidence for selection [119].

Phasing of the human genome becomes more complex at CNV loci, where phasing involves not only knowing the number of copies of a region and the variants found on each haplotype, but the positions of these variants relative to each other along the chromosome- i.e. the structure. This information is important when trying to determine the relationship between copy number variation and gene expression, which will involve a combination of the copy number of each haplotype, the phase of variants which may influence gene expression and the mechanism of expression

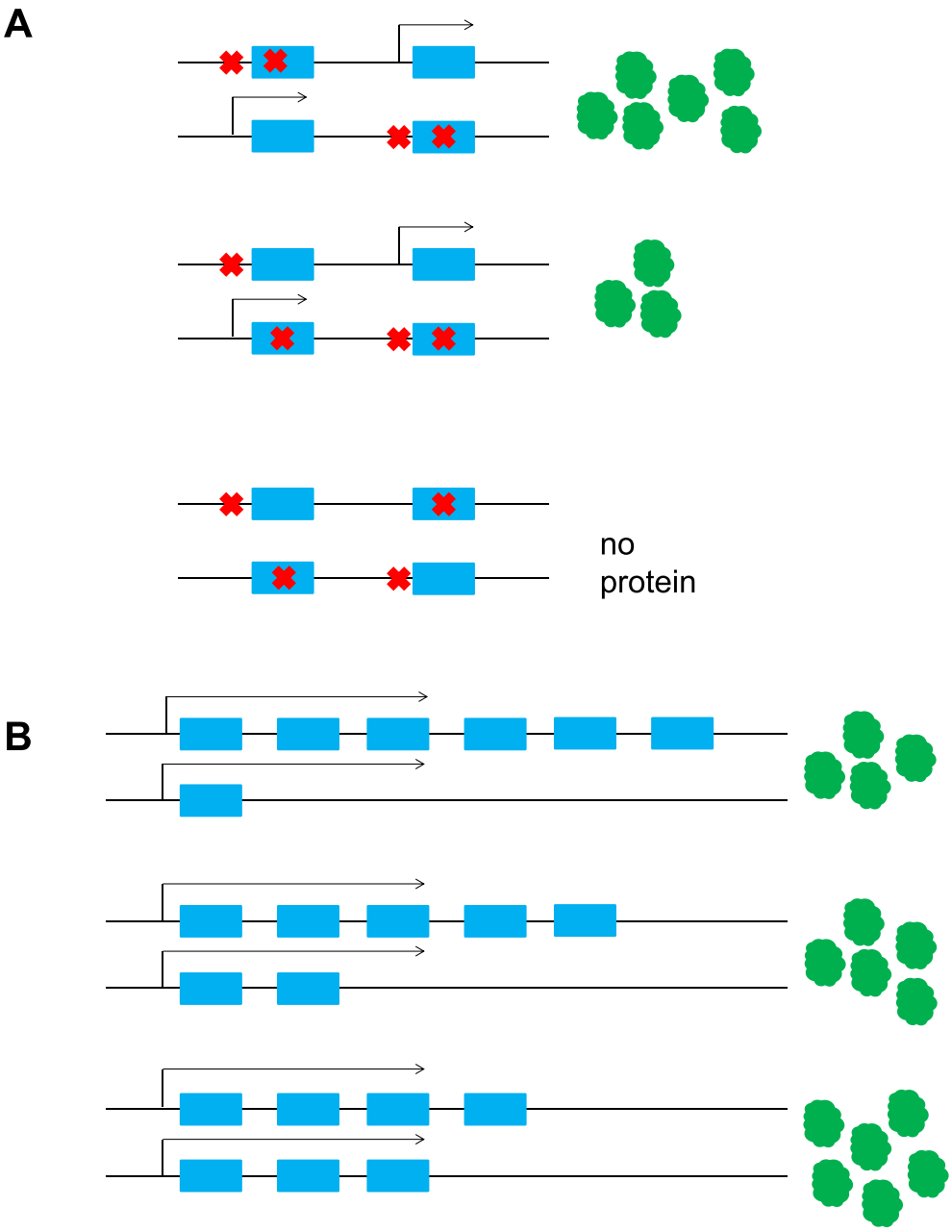


Figure 1.9: Phasing and structural information is important in understanding the relationship between copy number variation and gene expression. A: The blue gene is copy number variable, but there are loss of function mutations disrupting the promoter and gene regions (red cross). Different structures of four-copy diplotypes could result in differential gene expression of the protein (green), depending on the phase of the two mutations [111]. B: It is possible that gene expression regulatory mechanisms limit the number of copies of a gene expressed in a repeat array [115]. For an individual with six copies of a gene (blue), the expression level will depend on the phasing of those copies, if, for example, only the first three copies in the array are expressed. Adapted from a figure by Tewhey *et al.* [111].

at the locus (figure 1.9). As shown in section 1.2, the relationship between copy number and expression is not straightforward. Therefore, different approaches have been taken to phase these complex regions of the human genome.

Methods for phasing

Many different approaches to phasing the human genome have been used. They vary in scale, phasing variants across either targeted regions or the whole genome. The type of variants phased also varies; whilst some methods are able to phase CNV regions of genomes, others can only phase SNVs and Indels. Some methods require specialised equipment, whereas others are PCR-based and are widely applicable.

Methods used to phase variants across the whole genome include fluorescence-activated cell sorting [120] and microdissection [121], which separate out whole chromosomes prior to genotyping. However, both of these procedures require the use of specialist equipment. Segregation in families is an effective method of phasing variants across an entire genome [116], but is a problem if a variant is heterozygous in all members of a trio and a family resource is not always available. Computational phasing is often used to infer the phase of variants based on the common haplotypes in the population. However, this will miss novel and rare haplotypes, which may be of most interest [122]. Fosmid sequencing has been applied to the phasing of both SNVs and Indels across a whole genome [123, 124]. The sequencing of sperm cells has been used to generate haploid data [125] , but this method is not widely applicable. A recently developed method is whole-genome proximity ligation, in which a digest and ligation process aims to link regions of DNA from the

same molecule that were once far apart, allowing recovery of phase over long distances using short-read sequencing [126]. Dilution-based methods of phasing have also recently been used. The principle of these is that, by diluting genomic DNA to sub-haploid quantities before barcoding and NGS of each pool, the chance of two molecules covering the same region from different haplotypes being in same pool is very small. Therefore, the sequence data from each pool must have originated from the same starting molecule. This has been used for both whole-genome and targeted phasing [127–129]. Long-range allele-specific PCR can also be used to phase variants in the genome, although this is only applicable to targeted regions [130].

However, whilst these methods are all able to achieve phasing of SNPs and, in some cases, Indels, they are unsuitable for CNV regions of the genome. This is because CNVRs are incredibly complex and cannot be assembled using short-read sequencing data; even if longer read sequence data is available, the software to assemble this sequence is unavailable. Segregation in three-generation families has been used to successfully determine the number of copies of a CNVR on each haplotype [50], as well as the basic structure of the complex β-defensin locus [56]. However, a three-generation family resource is rarely available. Emulsion haplotype fusion-PCR (EHF-PCR) has been used to obtain structural information in relation to the relative positions of the *DEFA1* and *DEFA3* genes at the *DEFA1A3* CNV locus [131]. However, whilst this method can capture sequence variants across a region of up to 1kb across the CNVR, it is unable to provide complete phased sequence across the entire locus. Single-molecule sequencing may address this. Pacific Biosciences long read single-molecule real-time sequencing can generate reads of up to 30kb, with an average read length of 8.5kb. If

targeted to a specific region, this can allow reconstruction of complex regions of genomes [132, 133]. The advantage of this method is that the long reads mean a single read should contain multiple heterozygous positions, making phasing possible. However, this technology has yet to be applied to multiallelic CNV loci.

The importance of phase information

There are many examples where knowing the phase of sequence variants in the human genome is essential. One example is in the matching of host and donor recipients for transplants. In trying to identify whether a donor is suitable, markers across the major histocompatibility complex of both host and donor are genotyped [134]. However, it has been shown that the phase of these markers and not just the diploid composition is important in transplant success, with haplotype mismatches increasing the risk of graft versus host disease [112]. Another example is found at the locus associated with facioscapulohumeral muscular dystrophy (FSHD). It initially appeared that a contraction of the D4Z4 repeat was associated with FSHD, but there were examples where a D4Z4 contraction was not sufficient to cause FSHD. Further analysis indicated that the contraction of D4Z4 must occur on a haplotype background also containing a mutation in the Dux4 transcript [113]. From a diploid genotype, it would not be possible to tell if the mutations occur on the same haplotype, but diagnosis of FSHD as opposed to other muscular dystrophies is key for appropriate treatment of the disease. A study by Lupski *et al.* [114] showed that different mutations in the *SH3TC2* gene, associated with Charcot-Marie-Tooth disease, segregate with different clinical phenotypes when in compound heterozygous form; however, the phe-

notype would differ if both variants were found in the same copy of the gene. Therefore, in order to provide an accurate diagnosis and, as such, appropriate treatment, phased information was essential.

Phased sequence information at CNV regions can also be necessary to understand the phenotypic effect of mutations. For example, a low copy number of the gene *FCGR3B* has been associated with an increased risk of SLE [135]. However, determining the structure of each haplotype at the locus was essential for understanding the mechanism of the disease. It is the presence of a zero-copy *FCGR3B* haplotype that increases the risk of SLE, by bringing together the regulatory region from *FCGR2C* with the coding region of *FCGR2B*, leading to aberrant expression of a chimeric gene, *FCGR2B'*, in natural killer cells [62]. A second example is found at the red and green cone photopigment genes. Due to their high sequence similarity, the two genes are prone to NAHR, leading to CNV of both genes. However, it is thought that only the first two genes in the array are expressed, which can result in red-green colour blindness if there is not expression from both a red and a green cone photopigment gene. This could be due to the first two genes both encoding the same cone photopigment, one of the genes containing an inactivating mutation or one of the genes consisting of a fusion of the red and green photopigment genes [115].

1.5 Remaining questions at the *DEFA1A3* locus

There has been extensive work to define both the diploid and, in the case of the European population, haploid copy numbers at the *DEFA1A3* locus [50, 55, 84, 85]. In addition, information is known about vari-

ants within the *DEFA1A3* repeats; the ratios of *DEFA1* vs. *DEFA3*, inserted to deleted form of a 5bp Indel upstream of each copy of *DEFA1A3* and unduplicated to duplicated form of a 7bp duplication in intron 1 of *DEFA1A3* are known for 151 independent haplotypes with European ancestry [50]. However, the relationship between *DEFA1A3* haplotypes is unknown; haplotypes with the same copy number but different composition of internal variants have been observed. It is possible that haplotypes with the same copy number are more closely related than haplotypes with different copy numbers and that it is the internal composition of the locus that changes. Alternatively, related haplotypes may have different copy numbers. Each scenario would suggest the occurrence of different mutational processes at the *DEFA1A3* locus. This is of particular interest, given that the locus falls within a region of high LD [10], suggesting crossover mechanisms are not responsible for changes in *DEFA1A3* copy number. In order to address this question, it will be necessary to define related classes of *DEFA1A3* haplotypes.

The SNP rs4300027 has previously been identified as a tag of *DEFA1A3* copy number in the European population [50]. However, the basis of this association and the inability to tag copy number in non-European populations is unknown. It is also possible that additional variants may be able to further partition this association or tag additional variants across the locus, such as the presence of *DEFA3*. However, these regions usually lack SNP annotations, due to the CNV nature of the locus. Therefore, the identification of variants flanking the *DEFA1A3* locus could identify additional tags of features of the *DEFA1A3* locus.

In addition, very little information is known about the structures of *DEFA1A3* haplotypes- i.e. the relative positions of the *DEFA1* and *DEFA3* genes

and additional variants across the CNVR. EHF-PCR has been used to achieve this across a small number of haplotypes [131]. The application of this method to a larger number of haplotypes could provide detailed structural information for haplotypes across the *DEFA1A3* locus. Combined with information about the relatedness of haplotypes, this could identify the mechanisms responsible for changes in *DEFA1A3* copy number. Also, the relationship between *DEFA1A3* copy number and HNP1-3 expression is unclear. Whilst one small-scale study found a positive correlation between *DEFA1A3* copy number and HNP1-3 expression [85], this has not been replicated in a second study (Danielle Carpenter, personal communication). However, there is a wide variability in HNP1-3 expression (Danielle Carpenter, personal communication). In addition, a SNP within the same LD block as the *DEFA1A3* locus has been associated with a decreased risk of IgA nephropathy [110]. Therefore, it is likely that variation at the *DEFA1A3* locus contributes to this risk. It is possible that the structure of the *DEFA1A3* locus may contribute to differences in expression. Therefore, structural information at the *DEFA1A3* locus is needed to both infer mechanisms of copy number change and investigate how variation at the locus contributes to differences in HNP1-3 expression.

2 Methods

2.1 General Reagents and Methods

DNA samples

CEPH Families

The Centre d'Etude du Polymorphisme Humain (CEPH) DNA samples represent three-generation families with European ancestry. Some of the first and second-generation individuals are represented on the HapMap CEU panels. For this project, DNA extracted from lymphoblastoid cell lines of individuals from 24 families were used: 12, 66, 104, 884, 1331, 1332, 1333, 1334, 1340, 1341, 1344, 1345, 1346, 1347, 1350, 1362, 1375, 1408, 1416, 1420, 1421, 1424, 1454 and 13292. The DNA samples can be purchased from Coriell (http://ccr.coriell.org).

HapMap

The HapMap phase 1 samples represent 360 individuals from four populations [116]. 180 are from individuals living in Utah, USA, with northern and western European ancestry (CEU) and were collected by the CEPH. These consist of 56 trios, 5 duos and 2 singletons. 90 samples are from the Yoruba people of Ibadan, Nigeria (YRI) and these consist of 30 trios. 45 samples are from unrelated individuals from the Japanese popula-

tion of Tokyo (JPT) and 45 are from unrelated individuals from the Han Chinese population of Beijing (CHB). These DNA samples can be purchased from Coriell (http://ccr.coriell.org).

ECACC

The European Collection of Animal Cell Cultures (ECACC) human random control (HRC) samples are 480 randomly selected, unrelated Caucasians from the United Kingdom. The DNA samples can be purchased from ECACC (http://www.hpacultures.org.uk/collections/ecacc.aspx).

BBSRC Project BB/I006370/1 Volunteers

120 unrelated individuals of UK ancestry, defined as having at least 3 grandparents born in the UK, were recruited from the University of Nottingham as part of BBSRC Project BB/I006370/1. The volunteers gave informed, written consent in a project approved by the local ethical board and had no known clinical phenotype. This study aims to identify whether there is an association between *DEFA1A3* copy number and the amount of HNP1-3 expression in neutrophils. Blood collection was performed by John Armour, with DNA and neutrophil extraction, *DEFA1A3* copy number typing and HNP1-3 quantification performed by Danielle Carpenter and Laura Mitchell.

PRT reference samples

Seven PRT reference samples were used as calibrators for samples with an unknown *DEFA1A3* copy number. Three were from the CEPH families and their *DEFA1A3* copy number had been confirmed via segregation; 1340-03 (5 copies *DEFA1A3*), 1420-04 (6 copies) and 1340-05 (7 copies). Four reference samples were taken from the ECACC HRC-1 panel and these were selected based the consistency of the *DEFA1A3* copy number estimated from repeated measurements using different copy number typing methods; C0007 (7 copies), C0075 (6 copies), C0150 (8 copies) and C0877 (9 copies).

Non-human primate DNAs

Four Gorilla DNA samples were used to obtain additional information about the ancestral state of sequence variants in the flanking region of the *DEFA1A3* locus. Three of these, Guy, Sylvia and J79, were obtained from Gorillas at Twycross Zoo by Alec Jeffreys (University of Leicester). The fourth, EBJC, was taken from a Gorilla cell line prepared by John Clegg (IMM, Oxford).

Human Reference Assembly

All genome coordinates used refer to the March 2006 NCBI36/hg18 human reference assembly, available via the UCSC genome browser [93], unless otherwise stated. This assembly was used for sequence alignments and primer design.

Primer design

PCR primers were designed using Primer3 [136]. The target DNA sequence was obtained from the UCSC genome browser [93]. Primer pairs were chosen to have similar annealing temperatures, a GC content of approximately 50% and a length of approximately 20bp. Allele-specific primers were designed to have an annealing temperature 3-4°C lower than the corresponding primer in the pair, in order to generate allele-specific PCR conditions. All primers were checked using the UCSC genome browser *in silico* PCR tool [93], to ensure specificity to the target location, and dbSNP [11] and the trace archive [137], to ensure the primers did not overlap known sequence variants.

PCR buffers

10x LD PCR buffer

10x LD buffer was used for all PCRs with the following reagent concentrations unless otherwise stated. Each 10µl reaction contained 1xLD buffer, 1µM Forward primer, 1µM Reverse primer, 0.5 Units Taq DNA polymerase (New England Biolabs) and 10ng genomic DNA, made up to 10µl with distilled H_2O (dH_2O). 10x LD buffer contains: 500mM Tris HCl pH8.8, 125mM Ammonium Sulphate, 14mM $MgCl_2$, 75mM 2-mercaptoethanol, 2mM each dNTP and 1.25mg/ml BSA.

One*Taq*

One*Taq* (New England Biolabs) was used for the amplification of the 4.1kb region flanking the *DEFA1A3* locus. Each 20µl reaction contained 1x One*Taq* buffer (1.8mM Mg^{2+}), 1µM Forward primer, 1µM Reverse primer and 20ng genomic DNA, made up to 20µl with dH_2O.

Phusion

Phusion DNA polymerase (New England Biolabs) was used for the first stage of the emulsion haplotype fusion-PCRs (EHF-PCRs). Each 25µl reaction contained 1x GC buffer (1.5mM Mg^{2+}), 0.2mM each dNTP, 1µM F1 primer, 25nM F2′R1/F2R1′ primer, 1µM R2 primer, 2 Units Phusion DNA polymerase and 50ng genomic DNA, made up to 25µl with dH_2O.

BIOTAQ

BIOTAQ DNA polymerase (Bioline) was used for all allele-specific PCRs, with the exception of the reamplification of the Centromeric Indel5 and all three Microsatellite EHF-PCR product. Each 20µl reaction contained 1x NH_4 buffer, 2mM $MgCl_2$, 0.2mM each dNTP, 0.5µM Forward primer, 0.5µM Reverse primer, 1 Unit BIOTAQ and 1µl PCR template, made up to 20µl with dH_2O.

Agarose gel electrophoresis

Agarose gels of between 0.8%- 2.5% were used, depending on the sizes of the products being separated. Gels were prepared using 0.5x TBE

buffer containing 0.5µg/ml ethidium bromide. 10x TBE was prepared using 109g Tris Base, 55g Boric acid and 9.3g EDTA in 1l of dH_2O. 5x loading buffer was added to each sample prior to loading; 5x loading buffer was prepared with 0.5x TBE, 40% sucrose and 0.02% bromophenol blue. 1µl of either 100bp or 1kb DNA ladder (New England Biolabs) was run alongside the samples. Gels were run at 80-120V in 0.5x TBE, containing 0.5µg/ml ethidium bromide, and visualised under UV light.

Capillary electrophoresis

For each set of 16 samples, 2µl of 500 ROX size standard (Applied Biosystems) was added to 170µl of HiDi formamide (Applied Biosystems). 10µl of the HiDi/ROX mixture was aliquoted per sample, to which the fluorescent PCR products were added. The samples were denatured at 96°C for 3 minutes and then cooled on ice for 2 minutes. The volume of product added, as well as the injection time and voltage used, are listed for each assay. Capillary electrophoresis data was analysed and peak areas extracted using GeneMapper v4.1 (Applied Biosystems). For the allelic ratio assays, including the *DEFA1A3* microsatellite, the ratios between the peaks were calculated. For the PRTs, the ratios of test *vs.* reference peak areas were plotted against copy number for the seven reference samples. This graph was then used to infer the copy number for the test samples, based on their test *vs.* reference peak ratio.

Sanger DNA sequencing

PCR products were purified using AmpureXP (Agencourt), according to the manufacturer's protocol. Approximately 20ng of purified PCR prod-

uct was sequenced in a 10µl reaction containing 1µl Big Dye (Invitrogen), 3µl 5x sequencing buffer and 0.5µM primer. 5x sequencing buffer contains 10mM $MgCl_2$ and 250mM Tris pH9. The cycling conditions were:

96°C	30 seconds	
50°C	15 seconds	25 cycles
60°C	4 minutes	

Sequenced products were cleaned using CleanSeq (Agencourt), according to the manufacturer's protocol. Samples were sent to DBS genomics (https://www.dur.ac.uk/biosciences/services/dna/) for electrophoresis using an ABI 3730. BioEdit [138] was used to view sequence traces and score variants, with sequence alignments created using ClustalW [139].

HapMap and 1000 Genome phased SNP data

Phased SNP genotype data for the HapMap CEU1 samples was obtained from the HapMap project release #24, phase 1 and 2 (http://hapmap.ncbi.nlm.nih.gov/) [18]. Phased SNP genotype data for the 1000 Genomes samples was obtained from the 1000 Genomes project (http://www.1000genomes.org/) [20].

2.2 *DEFA1A3* copy number typing

The *DEFA1A3* copy number was determined for the third-generation individuals from three CEPH families (1340, 1420 and 1454) using two PRTs (MLT1A0 and DEFA4) and two allelic ratio assays (DefHae3 and Indel5), as described by Khan *et al.* [50]. The 7bp duplication ratios

Assay	Primers 5′ to 3′	Details
MLT1A0	FAM/NED-CCCAGAGAGCTCCTTC GTGACTTATAAACAACAAAAA	Uses the MLT1A0 dispersed repeat within the *DEFA1A3* full repeats as the test locus and a similar repeat on chromosome 1 as the reference locus
DEFA4	TGCTCCTGCTCTCCCTCCT HEX-TTGGAATCAAGTCTTTGGAGAAA	Uses the *DEFA4* gene as a reference locus and an *Msp*I digest to distinguish test and reference
DefHae 3	TGTCCCAGGCCCAAGGAAAA FAM-TCCCTGTAGCTCTCAAAGCA	Using a *Hae*III digest, this assay determines the ratio of *DEFA1* to *DEFA3*
Indel5	HEX-CTGTCCAGGAAGAGGGAGAG CAGCTGGAGGGTCTCTGTTC	Provides the ratio of inserted to deleted form of a 5bp Indel located upstream of each *DEFA1A3* gene
7bp dup	HEX-AGCAAAAATCAAACAACCTGA GCTATGCCTCCAATCTGACC	Provides the ratio of non-duplicated to duplicated form of a 7bp duplication located in intron 1 of each *DEFA1A3* gene

Table 2.1: The primer sequences for the two PRTs and three allelic ratio assays used to determine *DEFA1A3* copy number, as described by Khan *et al*[50].

for these samples were provided by Omniah Mansouri. The *DEFA1A3* copy number typing of the 21 additional CEPH families and the HapMap, ECACC and BBSRC Project BB/I006370/1 volunteer samples was performed by Fayeza Khan, Danielle Carpenter, Laura Mitchell and Omniah Mansouri. The primers used for each of the five assays are shown in table 2.1.

The PCR products were analysed using capillary electrophoresis, as described in section 2.1, using the conditions shown in table 2.2. Additional details can be found in Khan *et al.* [50].

Assay	Voltage	Injection time
1µl MLT1A0 NED 1µl MLT1A0 FAM 1µl Indel5	1kV	30 seconds
4µl DEFA4 4µl DefHae3	2kV	45 seconds
1µl 7bp dup	1.2kV	23 seconds

Table 2.2: The capillary electrophoresis conditions used to quantify the PCR products from the two PRT and three allelic ratio assays. The MLT1A0 and Indel5 products were multiplexed, as were the DEFA4 and DefHae3 products.

Assigning the maximum likelihood copy number values

A maximum likelihood copy number (MLCN) was assigned to each sample typed for *DEFA1A3* copy number using the ”DEFAML” program, written by John Armour. This takes the unrounded copy number values predicted by the two PRTs and the ratio values obtained from the DefHae3, Indel5 and 7bp duplication assays. It evaluates the probability of obtaining these results assuming a *DEFA1A3* copy number of 2 to 16. For each possible copy number, the program assigns a probability, with the most likely copy number assigned a probability of 1 and all other probabilities

assigned relative to this; the program assumes flat priors. A minimum ratio value is also determined, which shows how many times more likely the assigned copy number value is compared to the next most likely copy number. Therefore, a high minimum ratio value suggests a high confidence in the assigned copy number. The program takes into account the standard deviations for each PRT measurement for each copy number, assuming a normal distribution. The program also applies correction factors to the DefHae3 and Indel5 ratio values. The DefHae3 is a digest-based assay which leads to a consistent over-representation of the larger *DEFA3* peak due to heteroduplex formation. The deleted Indel5 allele is consistently overrepresented in the Indel5 assay, due to an amplification advantage for the smaller product during the PCR.

Segregation in three-generation families

The *DEFA1A3* diploid copy number and frequency of the *DEFA3* and Indel5 insertion alleles were deduced from segregation to give the constituent haplotype values for individuals in the three-generation CEPH families 1340, 1420 and 1454. The transmission of haplotypes from the first to third-generation individuals was inferred using linkage data, extracted from the CEPH genotype database (www.cephb.fr/cephdb), using the CHROMPIC option of CRIMAP [140] by John Armour.

Read depth analysis

Next-generation sequencing data for 1062 samples in the 1000 Genomes project [20] were downloaded by John Armour; 15 of these samples were not part of the 1000 Genomes Phase 1 data release. Read counts

were obtained for reads mapping to the *DEFA1A3* locus (GRCh37/hg19 chr8: 6829298-6837591, 6848458-6856701 and 6867561-6875800) and to the two-copy flanking regions (GRCh37/hg19 chr8: 6700000-6830000 and 6900000-7000000), using Samtools, with the command ”samtools view -c” [141]. Flanking regions were selected to have a similar GC content to the CNVR selected from the *DEFA1A3* locus. The *DEFA1A3* regions were selected as they represent the region shared by all three copies of *DEFA1A3* (2 full and 1 partial), as shown on the human reference assembly. For each flanking region and the *DEFA1A3* locus, the reads per base was calculated. A weighted mean was obtained for the reads per base for the two flanking regions. The ratio of reads per base for the *DEFA1A3* locus to the reads per base for the flanking regions was obtained and multiplied by two; this represented the diploid *DEFA1A3* copy number estimation. The *DEFA1A3* copy number estimates for the 98 HapMap CEU1 samples were compared to those obtained by Khan *et al.* [50].

The ratio of reads mapping to *DEFA1* compared to *DEFA3* was calculated by counting the number of reads carrying each allele for the 45 HapMap CEU1 samples included within the 1000 Genomes dataset [20]. This was achieved using Samtools with the command ”samtools tview” [141]. The reads with either the G allele (*DEFA1*) or the T allele (*DEFA3*) were counted and a ratio was obtained. This was compared to the ratio obtained by Khan *et al.* [50].

2.3 *DEFA1A3* microsatellite

The *DEFA1A3* locus contains a highly variable microsatellite, formed from the poly-A tail of an Alu element (figure 2.1). A *DEFA1A3* microsatellite locus is found downstream of each copy of the *DEFA1A3* gene, with a copy in each full repeat and the telomeric partial repeat. This allows the *DEFA1A3* microsatellite to act as a measure of *DEFA1A3* copy number, as well as acting as an internal variant distinguishing different copies of *DEFA1A3* from each other.

$(YCTTT)_{1\text{-}2}\ (CTTTT)_{2\text{-}3}\ (CTTT)_{2\text{-}14}\ (CTTTCTTTTTTT)_{1\text{-}2}\ (Y)_{0\text{-}1}\ (CTTTT)_{1\text{-}2}\ (Y)_{0\text{-}1}\ (C)\ (T)_{7\text{-}21}$

Figure 2.1: Sequence of the *DEFA1A3* microsatellite. Trace archive searches show a high variability in length, with repeats of varying length each found with a variable number of copies.

An allelic ratio assay was designed to measure the *DEFA1A3* microsatellite allele sizes and frequencies of each allele. The primers HEX-5′-GGATCCAGGTGGAGTCTCA-3′ and 5′-CAGCGTGGGTGACAGATCTA-3′ were used to amplify the *DEFA1A3* microsatellite, using the PCR conditions:

95°C	1 minute	1 cycle
95°C	30 seconds	
60°C	1 minute	24 cycles
70°C	1 minute	
70°C	10 minutes	1 cycle

1µl of the PCR product was analysed using capillary electrophoresis with an injection time of 30 seconds at 1.5kV.

Assigning maximum likelihood copy number values

The *DEFA1A3* microsatellite peak areas were used to estimate the MLCN of the sample, using the microsatellite peak analysis program, written by John Armour. This program takes the peak areas of up to six peaks per sample and finds the best fit for the given peak ratios to an integer copy number value of between 2 and 12. The program assumes that the observed ratio/expected ratio values fit a normal distribution with a mean of 1 and a standard deviation of 0.135 and also assumes flat priors. As for the DEFAML program, a probability of 1 is assigned to the MLCN, with the probabilities for the other copy numbers assigned relative to this. A minimum ratio value is generated, indicating how many times more likely the assigned copy number is than the next most likely copy number, with a high minimum ratio value indicating a high confidence in the assigned copy number.

Segregation in three-generation families

The *DEFA1A3* microsatellite allele sizes could be partitioned into the constituent haplotype compositions using segregation in three-generation families, as described for *DEFA1A3* copy number and the Indel5 and *DEFA3* variants in section 2.2.

2.4 Pulsed-field Gel Electrophoresis

A *Hpa*I digest cuts out the entire *DEFA1A3* region onto a single restriction fragment. This allows the haplotype copy numbers to be determined for a sample using Pulsed-field Gel Electrophoresis (PFGE), as previ-

ously shown by Aldred *et al.* [55].

Sample preparation

2µg of each genomic DNA sample was digested overnight at 37°C using 5 units of *Hpa*I. This cuts either side of the *DEFA1A3* repeat array, leaving the *DEFA1A3* genes on a fragment 30 + 19(n-1)kb in length, where n is the *DEFA1A3* haplotype copy number.

Gel preparation and electrophoresis

A 1% agarose gel was prepared using 0.5x TBE buffer. 1ml of the molten agarose was saved and used to partially fill each well, before the digested DNA samples were added. The Midrange I PFG marker (New England Biolabs) was added to one lane. The wells were filled with the molten agarose and left to set. The gel was run using a CHEF Mapper XA (BioRad) in 0.5x TBE cooled to 14°C for 24 hours, with a switch time ramped from 1-8 seconds at 6V/cm. These were the conditions used by Aldred *et al.* [55].

Staining and visualisation

The gel was stained for 1 hour using 5µl of 5mg/ml ethidium bromide in enough water to cover the gel. The Midrange I PFG marker lane was cut from the gel and viewed under UV light, allowing a cut to be made at the position of each marker band. This allowed a gel block to be cut from each of the positions expected to be occupied by a *DEFA1A3* haplotype with between 1-7 copies, with the regions in between these

blocks sampled to act as controls. A control block was also cut from an unloaded lane. Three gel volumes of distilled water were added to each gel block.

Testing for presence of *DEFA1A3* DNA

The gel block samples were heated to 95°C to allow the agarose gel block to melt. 2.5µl of this was added to the "DEFA1A3F/R" PCR, designed to detect the presence of the *DEFA1A3* genes. The primers for the assay are 5′-TGGGAAATCAGTGGTGTCA-3′ and 5′-CCTTCTGCTCAATTTCCTTTCC-3′, with the cycling conditions:

95°C	1 minute	1 cycle
95°C	30 seconds	
65°C	30 seconds	35 cycles
70°C	1 minute	

Products were run on a 1% agarose gel. This confirmed the presence of the *DEFA1A3* genes within the DNA lane, but not blank lanes, of the gel.

Allelic ratio assays

In order to determine the haplotype copy numbers of a sample following PFGE, the DefHae3, Indel5 and 7bp duplication allelic ratio assays (section 2.2) were performed, using 2µl of each melted gel block as input in place of genomic DNA. The cycle number for each assay was increased-DefHae3 to 28 cycles, Indel5 to 26 cycles and 7bp duplication to 26 cycles. The PCR products were analysed using capillary electrophoresis. 5µl of the digested DefHae3 PCR product was run at 1.5kV with an injection time of 45 seconds; 2µl of the Indel5 PCR product was run at 1kV

with an injection time of 30 seconds; and 2µl of the 7bp duplication PCR product was run at 1.2kV with an injection time of 23 seconds.

2.5 Sequencing the *DEFA1A3* centromeric flanking region

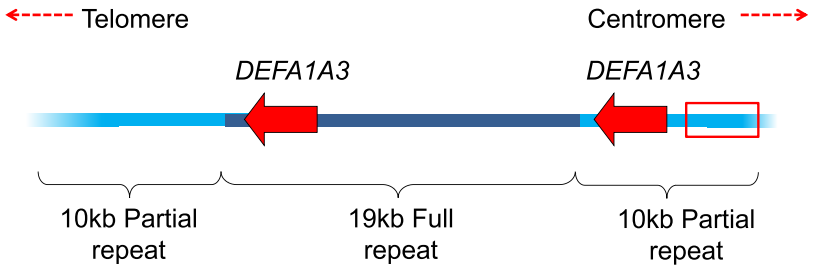


Figure 2.2: A 4.1kb region within the *DEFA1A3* centromeric partial repeat was selected for sequencing across 30 HapMap trios (chr8: 6864188-6868287; red box). The reverse primers used to amplify the region were placed in a region with reduced sequence similarity to the *DEFA1A3* full repeats, ensuring specific amplification from the *DEFA1A3* partial repeat.

A 4.1kb region flanking the *DEFA1A3* locus (chr8: 6864188-6868287; figure 2.2) was sequenced across the 90 HapMap CEU 1 samples, which represent 30 two-generation trios. This region was selected by John Armour based on its proximity to the *DEFA1A3* locus, whilst enabling amplification specifically from the partial repeat copy only, due to the specificity of the reverse primers to the partial repeat location. Two overlapping fragments covering the 4.1kb region were amplified with One*Taq*, using the primers shown in table 2.3. The PCR conditions used were:

94°C	30 seconds	1 cycle
94°C	30 seconds	
68°C	1 minute	30 cycles
68°C	4 minutes	
68°C	5 minutes	1 cycle

The PCR products were sequenced as described in section 2.1, using twelve different primers to cover the entire 4.1kb region (table 2.4).

Primer name	Sequence 5′ to 3′	Region amplified
F+P_F_2	TGGGAGGCATAGGAGTTTCCA	6864055-
3078R	CCCAGAAAACAGCATGGCATC	6867265
RF4_2	GCCTCCCCATGAAACTCAGAACC	6865535-
RR	TCTCACCAAGACCGTGCCTTCTG	6868287

Table 2.3: The two pairs of primers used to amplify overlapping fragments of the region chr8:6864188-6868287 in preparation for sequencing.

Primer name	Sequence 5′ to 3′
F+P_F_seq	TGCAAKGCTCCAACTCTTCAG
F+P_R_seq	CCAGGTTTCTGCAGGACACACT
F+P_R_seq2	CCTCACTACCGTCCACCACAA
F+P_R_seq3	AGCAGGACCACRAGGCTTTT
RF4_2	GCCTCCCCATGAAACTCAGAACC
RR5	CACTGATTGTCTACACTGGCTGCAA
RR4	CAGGGACTTGGAGCTCCTACCTGT
RR3	TCTACAGGGGCACTCATTCCATTCA
RR2	TCCTCCTCCAAGCATGGTATCTGG
RR_2	GGACTGTGCGAAAAGACACCACA
RR	TCTCACCAAGACCGTGCCTTCTG
FE	CACCTGCAAGGATGGGCTAGAGA

Table 2.4: The twelve primers used to sequence the region chr8: 6864188-6868287 flanking the *DEFA1A3* locus. K= mixed primer made with either a G or T nucleotide at this position.

Phasing sequence variants across the *DEFA1A3* flanking region

As the samples sequenced were part of two-generation trios and the region amplified was present as a single copy per haplotype, the majority of variants could be phased across the entire 4.1kb region using segregation. However, for variants that were heterozygous in all three individuals within the trio, allele-specific PCR was used to confirm the phasing. The fragment of the 4.1kb region containing the variant to be

Position	Primers 5′ to 3′	PCR conditions
6864445 rs62487509	GAGGGCTGCTGGAACAA or GAGGGCTGCTGGAACAG with AGCCCATTTGGATTGAAGCCT	95°C for 30 seconds 68.1°C for 30 seconds 70°C for 1 minute for 30 cycles
6865667 rs4284061	CCCACTTGTCCATTCTGCA or CCCACTTGTCCATTCTGCT with Fseq (table 2.4)	95°C for 30 seconds 69.4°C for 30 seconds 70°C for 2 minutes for 26 cycles
6864878 rs11137085	GTGAAATGGAGAGGTGTGGTC or GTGAAATGGAGAGGTGTGGTG with Rseq (table 2.4)	95°C for 30 seconds 69.4°C for 30 seconds 70°C for 90 seconds for 25 cycles
6868004 rs4512398	CAGATCAGGCCAGCTCATGAGA or AGATCAGGCCAGCTGATGAGG with TGGGATCAGGCGCTAGTGAA	95°C for 30 seconds 68.1°C for 30 seconds 70°C for 1 minute for 30 cycles

Table 2.5: The primers and cycling conditions used for the allele-specific reamplification of sections of the 4.1kb region flanking the *DEFA1A3* locus.

phased was reamplified; 1µl of a 1:10 dilution of this product was used as the PCR template. Allele-specific primers were designed to amplify from four positions across the 4.1kb region (table 2.5), ensuring each product contained an additional variable site to allow phasing of the variant. The PCR products were sequenced as described in section 2.1.

Classifying sequence variants

In order to determine the ancestral and derived alleles for each variant, sequence alignments were made using human and non-human primate *DEFA1A3* sequences. A partial repeat alignment was made using Human, Chimpanzee (CSAC2.1.4/panTro4; Feb 2011) and Orang-utan (WUGSC2.0.2/pomAbe2; July 2007) sequence, obtained from the UCSC genome browser [93], and Gorilla sequence, obtained from sequencing a region of the partial repeat from four Gorilla samples. A full repeat alignment was made using Human, Chimpanzee and Gorilla (gorGor3.1/gorGor3; May 2011) sequence from the UCSC genome browser [93]. Due to the incomplete alignments of the *DEFA1A3* region in the non-human primate reference assemblies, it was not possible to identify the *DEFA1A3* partial repeat sequence for the Gorilla or full repeat sequence for the Orang-utan. The ancestral and derived alleles were identified for each variant and were then used to categorise the variants as either partial repeat-specific variants or a full repeat sequence replacement of the partial repeat. The rs numbers were obtained from dbSNP [11].

2.6 Exchange 1 sequence replacement typing assays

A sequence replacement event, termed Exchange 1, was observed across the 4.1kb region flanking the *DEFA1A3* locus, in which regions of the partial repeat were replaced with sequence from the equivalent region of the full repeat. A three-primer assay was designed to test for the presence of this sequence replacement (figure 2.3).

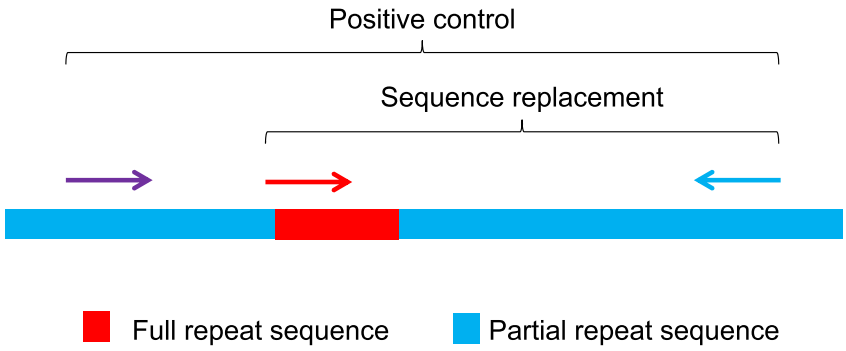


Figure 2.3: The Exchange 1 sequence replacement assay uses three primers. The reverse primer (blue) is specific to the partial repeat, whilst the leftmost forward primer (purple) matches both full and partial sequence. These produce a product from all individuals, acting as a positive control. The second forward primer (red) is specific to the sequence replacement junction and as such a product is only produced with this primer if the sequence replacement has occurred.

The reverse primer E1_P_R (5′-TGTAAGCCCTGTTAGAGGGGCTGT-3′) was designed to bind specifically to the partial repeat, ensuring amplification from only this location. The two forward primers used were E1A1_F+P_F (5′-TCTGGGAGGCATAGGAGTTTCCA-3′), which with the reverse primer produced a 341bp product from all samples, and E1_E_F (5′-TGTGTGCTGCSGCATCA-3′), which bound specifically to the Exchange 1 sequence replacement, only producing the 212bp product if this replacement had occurred. The E1A1_F+P_F primer was used at a final reaction concentration of 0.2µM, whilst the E1_E_F and E1_P_R

primers were used at a final reaction concentration of 0.5μM. The PCR cycling conditions were:

95°C	1 minute	1 cycle
95°C	30 seconds	
67°C	1 minute	30 cycles
70°C	90 seconds	

The products were run on a 1% agarose gel. This assay was used to type the HapMap CEU1, CHB, JPT and YRI samples.

2.7 Drawing phylogenetic trees

Two different phylogenetic trees were constructed using sequences from the flanking region of the *DEFA1A3* locus (chr8: 6864188-6868287). One used sequence across the entire 4.1kb region from haplotypes free of gene conversion events- Reference Sequence, Class 1 and Class 2. The second used sequence from only the final 1kb of sequence across each of 21 different haplotypes observed (chr8: 6867142-6868287). The sequences were aligned using ClustalW [139]. The sequence alignment was used to generate an unrooted phylogenetic tree, using the dnapars option of PHYLIP [142]. The drawtree option of PHYLIP was used to obtain a graphical output of the tree [142].

2.8 Assessing extended haplotype homozygosity

Extended haplotype homozygosity (EHH) is defined as the probability that two randomly chosen chromosomes carrying the allele of interest are identical by descent (i.e. homozygous) across a defined interval. It

can indicate selection, as a selected allele will take the flanking variants along with it during the selective sweep, leading to regions of EHH [143]. The integrated haplotype score (iHS) is a measure described by Voight *et al.* [144], which uses EHH to identify selective sweeps in the genome. Chromosome position (x-axis) is plotted against EHH (y-axis) for the derived and ancestral alleles of a SNP of interest. The area under the curve for each allele (integrated EHH; iHH) is calculated and the area for the ancestral allele is divided by the area for the derived allele. Taking the natural logarithm of this gives an unstandardised iHS. The minor allele frequency of the SNP of interest will affect the iHH value generated, as rare alleles would be expected to have arisen more recently and, as such, display higher levels of EHH, due to the fact there has been less time for new mutations to arise. Therefore, to standardise the iHS value based on minor allele frequency, the following formula is used:

$$standardised\ iHS = \frac{ln(\frac{iHHa}{iHHd}) - E[ln(\frac{iHHa}{iHHd})]}{SD[ln(\frac{iHHa}{iHHd})]}$$

where E= expected iHS for SNPs of same minor allele frequency and SD= standard deviation of iHS for SNPs of same minor allele frequency.

To calculate iHS for the SNPs tagging the five *DEFA1A3* haplotype classes and the SNP rs4300027, phased SNP genotype data across a region approximately 100kb in length surrounding the *DEFA1A3* locus was downloaded for HapMap CEU1 samples from the HapMap project [10] and for 1000 Genomes individuals from the 1000 Genome project [20]. For the HapMap CEU1 individuals, this was combined with the genotypes for variants across the region chr8:6864188-6868287, which were obtained through sequencing, as described in section 2.5. The ancestral allele for

each SNP was determined using data from Voight *et al.* [144]. The R package rehh [145] was used to calculate the iHH (i.e. area under the curve) for the derived and ancestral alleles for each SNP of interest, calculated to the point where EHH $<$0.05. Due to the obvious lack of SNP data within the *DEFA1A3* repeat itself, this region was removed and chromosome coordinates adjusted accordingly, to prevent an apparent large region of EHH due only to the lack of informative markers.

To standardise the iHS scores, phased SNP data was downloaded for the same 120 HapMap CEU1 haplotypes for 15,000 SNPs from chromosome 8 (chr8:21312570-39999923). The ancestral allele for each SNP was determined using data from Voight *et al.* [144]. The rehh package was used to calculate the mean and SD of the iHS statistic for each different minor allele frequency category [145]. In order to assess the significance of the iHS values obtained for rs4300027 and the five SNPs tagging *DEFA1A3* haplotype classes, iHS scores were downloaded from Voight *et al.* [144] for all SNPs assessed across the genome (n=748345) and the position of the iHS values obtained in this distribution was determined. Large negative iHS values indicate selection on a derived allele, whilst large positive values indicate selection on the ancestral allele.

2.9 Assigning *DEFA1A3* haplotype class

The sequence data obtained from the 4.1kb region flanking the *DEFA1A3* locus identified five common haplotypes, the identity of which could be determined using a combination of four SNP genotypes. A restriction fragment length polymorphism (RFLP) assay was used to genotype each of these four SNPs.

rs4300027

A RFLP assay was designed by Fayeza Khan to genotype the SNP rs4300027 (chr8: 6867985) [50]. This SNP creates a variable *Hin*fI site. The region around the SNP was amplified as a 1043bp product, in which the uncut T allele produces a restriction fragment of 613bp, whilst the cut C allele produces two restriction fragments of 439bp and 174bp (figure 2.4). The PCR conditions used were:

95°C	2 minutes	1 cycle
95°C	30 seconds	
56.8°C	30 seconds	36 cycles
70°C	30 seconds	

chr8: 6867808-6868850
AGATACCATGCTTGGAGGAAgactaagcatcccacagggagaggaactgagcccacctgcaa
ggatgggctagagaacactgagcaaccagctttctaggaaaaaagaaaactctgatttgcaa
tgtttgtaaatttctgtggttaaaatgctcccagctatagacagtttaagaattatcacaca
aaaactcctccctcatgagctggcctgatctgaccccagcacatcacagggtctcatccttc
agctttctcagagtttccagctgagccaacaccacctgccacctgtgcacgagtgtcctggc
cctgaaattttcagatctcagcagaacctctcctcttatgcccgtggaaggatccaaacccc
aattgcaaatgtgtgagtgaagacgtgatcatgctgtttcaatccactactttctgtggtgt
cttttcgcacagtcctagatgaacagaaggcacggtcttggtgagaagttgaatgtgtgcat
tttttgtgtgtgtaaatctcagcctctctataatattgttgaagtaggacagaaccctctca
ccttatttccaaagtgtcacaaagagcccatcttaatggcagcgtggaattgtggactctt
ggagtgactgaagaacccccgtcacccattcttagtttaaattcttcctgttcagagcaggg
gtggtgtgggagccaggtggagtgtcaacctctccccacagtgacagactcagaggaggcca
cgggacttgggggttggctggaacaacatgggaagaagtagggattttctccaggagattag
ctacaaaagtcatagagagatgacgatgattcacatggttgtcacctggggagagagcccaa
agtaagttcaagagaaacaaagggaaaaaaaaggtgaagacaacagtgtaactcatgctttg
atttcttggatcaagcaattcctgaagctagaattcctcctcaattccaagatataggagtc
aaaatgttaacttagcttgttgttggtttcTGACATTTGAATTCAAGACCC

Figure 2.4: Sequence of the PCR product amplified for typing of rs4300027. Upper case= primer sequences; red= rs4300027; underlined= *Hin*fI restriction sites.

6µl of PCR product was digested overnight at 37°C, using 2 units of *Hin*fI. The products were run on a 1.5% agarose gel. This assay was used to genotype individuals with missing data, with the majority of HapMap, CEPH, ECACC HRC and BBSRC Project BB/I006370/1 individuals typed

by Fayeza Khan, Danielle Carpenter and Laura Mitchell.

rs7826487

chr8: 6867476-6868496
CCTTCATTCCTTCCACCAGAtgaagacacaataagaagatgttgtatatgaaccaaaaagct
ggttctcaccagacactgattctgctgccaccttgatttggacttcttagcctccacagggg
attccgaatggtggcagaagttagggtgtggccatgcggtggagggttggaacaaagaagat
gaagggtcactgttgtcccggaggacagagcgctgagaggtagtgagggtgaccttccaggc
aatattagagtaacttagttcacgacattttgaatttttttagcagtccacatgtgcttag
agggctggtgcctttggagcccagataccatgcttggaggaagactaagcatcccacaggga
gaggaactgagcccacctgcaaggatgggctagagaacactgagcaaccagctttctaggaa
aaaagaaaactctgatttgcaatgtttgtaaatttctgtggttaaaatgctcccagctatag
acagtttaagaattatcacacaaaaactcctccctcatgagctggcctgatctgaccccagc
acatcacagggtctcatccttcagctttctcagagtttccagctgagccaacaccacctgcc
acctgtgcacgagtgtcctggccctgaaattttcagatctcagcagaacctctcctcttatg
cccgtggaaggatccaaaccccaattgcaaatgtgtgagtgaagacgtgatcatgctgtttc
aatccactactttctgtggtgtcttttcgcacagtcctagatgaacagaaggcacggtcttg
gtgagaagttgaatgtgtgcatttttgtgtgtgtaaatctcagcctctctataatattgtt
gaagtaggacagaaccctctcaccttatttccaaagtgtcacaaagagcccatcttaatggc
agcgtggaattgtggactctttggagtgactgaagaacccccgtcacccattcttagtttaa
attcttcctGTTCAGAGCAGGGGTGGTGT

Figure 2.5: Sequence of the PCR product amplified for typing of rs7826487. Upper case= primer sequences; red= rs7826487; underlined= *Ssp*I restriction sites.

A restriction digest assay was designed to genotype the SNP rs7826487 (chr8: 6868335), which acts as a tag of the ”Class 1” haplotype. This SNP forms a variable *Ssp*I restriction site. The region around this SNP was amplified as a 1021bp product, in which the uncut G allele produces a restriction fragment of 770bp and the cut A allele produces two restriction fragments of 611bp and 159bp (fig 2.5). The PCR conditions used were:

95°C	1 minute	1 cycle
95°C	30 seconds	
64.5°C	1 minute	35 cycles
70°C	90 seconds	

1µl of PCR product was digested overnight at 37°C, using 1 unit of *Ssp*I. The products were run on a 1.5% agarose gel. This assay was used to genotype all HapMap CEU, ECACC HRC and BBSRC project BB/I006370/1 samples.

rs7825750

A restriction digest assay was designed to genotype the SNP rs7825750 (chr8: 6867213), which acts as a tag of the ”Class 2” haplotype. This SNP does not form a variable restriction site, but via the introduction of a mismatch in the reverse primer, the SNP creates a variable *Rsa*I site. A mismatch is also added to the forward primer, allowing the creation of a control site. The region around this SNP is amplified as a 165bp product, in which the uncut C allele produces a restriction fragment of 144bp and the cut T allele produces two restriction fragments of 119bp and 25bp (fig 2.6). The PCR conditions used were:

95°C	1 minute	1 cycle
95°C	30 seconds	
56.5°C	30 seconds	33 cycles
70°C	30 seconds	
70°C	5 minutes	1 cycle

chr8: 6867074-6867238
GGATTGCAGCAGGTTTATTGTACaataaaccataaggaaatagcttccttgagctgttttaa
aatatgttccctaaattctttattatttttatagaaagaaaggtttctgttctctcctgt
taggtctaggatgggtACCATTTTGACCAACAGACTACATG

Figure 2.6: Sequence of the PCR product amplified for typing of rs7825750. Upper case= primer sequences; red= rs7825750; underlined= *Rsa*I restriction sites; blue= introduced mismatches.

10µl of PCR product was digested overnight at 37°C, using 2 units of *Rsa*I. The products were run on a 2.5% agarose gel. This assay was used to genotype all HapMap CEU, ECACC HRC and BBSRC project BB/I006370/1 samples.

rs62487514

A restriction digest assay was designed to genotype the SNP rs62487514 (chr8: 6867504), which acts as a tag of the "Exchange 1" haplotype. This SNP does not form a variable restriction site, but via the introduction of a mismatch by the reverse primer, the SNP creates a variable *Tsp*509I site. The region around this SNP is amplified as a 233bp product, in which the uncut C allele produces a restriction fragment of 188bp and the cut A allele produces two restriction fragments of 162bp and 26bp (fig 2.7). The PCR conditions used were:

95°C	1 minute	1 cycle
95°C	30 seconds	
50°C	30 seconds	32 cycles
70°C	1 minute	
70°C	5 minutes	1 cycle

chr8: 6867297-6867529
ATCTCTCTTTGGATGGTGctgtggtctgagacatttttcctctgaaattcatatgttgaagt
cttagaccctaggatgatagcattaaaagaaagggtctttttggaaatgaggaggtcagcat
ggtggagtcctcatgaatggaatgagtgcccctgtagaagaggtcccagagagctccttcat
tccttccaccagatgaagacaCATTAAGAAGATGTTGTATATGAACC

Figure 2.7: Sequence of the PCR product amplified for typing of rs62487514. Upper case= primer sequences; red= rs62487514; underlined= *Tsp*509I restriction sites; blue= introduced mismatches.

5µl of PCR product was digested overnight at 65°C, using 1 unit of *Tsp*509I. The products were run on a 2% agarose gel. This assay was used to genotype all HapMap CEU, ECACC HRC and BBSRC project BB/I006370/1 samples.

1000 Genome Samples

DEFA1A3 haplotype class was assigned to the 2184 haplotypes within the 1000 Genomes dataset using the same four SNPs described above. Phased SNP genotype data was downloaded from the 1000 Genomes project (http://www.1000genomes.org/) [20].

2.10 Separation of *DEFA1A3* full and partial repeats

Sample preparation

2µg of genomic DNA was digested overnight at 37°C using 10 Units of *Kpn*I. This allowed size-separation of the centromeric-most copy of *DEFA1A3* (the ”partial” repeat) from all other copies (the ”full” repeats).

Gel electrophoresis and DNA recovery

The *Kpn*I-digested genomic DNA was run on a 0.8% agarose gel. A dark reader was used to view the gel, allowing blocks of gel to be cut from the genomic DNA lane at the positions corresponding to lengths of 4kb and $\geq$10kb in size, as well as a control block from a blank lane of the gel. Each gel block was added to 3 gel volumes of distilled water.

Testing for presence of *DEFA1A3* DNA

Each gel block was heated to 96°C and 2µl was added to the "DEFA1A3F/R" PCR, as described in section 2.4, but with 7 fewer cycles (28 cycles). This confirmed the presence of the *DEFA1A3* genes in the 4kb and $\geq$10kb fragments, but not in blank lanes of the gel.

Allelic ratio assays

The gel block samples were heated to 95°C to allow the agarose gel block to melt. 2µl of this was used in place of genomic DNA in the DefHae3 allelic ratio assay (section 2.2), with an additional two cycles (total 27 cycles), allowing the ratio of *DEFA1*:*DEFA3* to be determined for the partial and full repeats separately. The PCR products were digested with *Hae*III overnight at 37° and 5µl of this was analysed using capillary electrophoresis, with an injection time of 45 seconds at 1.5kV. 2µl of the melted gel block was also used in place of genomic DNA in the 7bp dup allelic ratio assay (section 2.2), with a total of 27 cycles, allowing the ratio of unduplicated to duplicated form of the 7bp duplication to be determined for the partial and full repeats separately. 1.5µl of the PCR

product was analysed using capillary electrophoresis, with the conditions shown in section 2.2.

2.11 Emulsion haplotype fusion-PCR

Emulsion haplotype fusion-PCR (EHF-PCR) fuses two regions of DNA separated by a distance too long to be amplified efficiently using standard PCR. The EHF-PCR design used was developed by Jess Tyson [131] (figure 2.8) and the emulsion preparation was as described by Turner and Hurles [146]. These were based on an original idea of linking emulsion PCR [147, 148].

Seven EHF-PCR systems were designed, each of which fused a non-copy number variable region telomeric or centromeric to the *DEFA1A3* locus to a region of interest from within the *DEFA1A3* CNV region. The non-CNV regions were selected in conjunction with Jess Tyson and were the same for all assays - a 684bp region telomeric to the *DEFA1A3* locus (chr8: 6810649-6811332) and a 196bp region centromeric to the *DEFA1A3* locus (chr8: 6867863-6868058). The exception was for the Telomeric Microsatellite rs2738046 system, which fused a shorter, 409bp region (chr8: 6810924-6811332) to the *DEFA1A3* microsatellite locus. Each of the flanking regions contained multiple SNPs to allow allele-specific reamplification of the fused products. The flanking regions were then fused to either a 128-132bp region surrounding the Indel5 variant (chr8: 6827670-6827801; 6846780-6846906; 6865876-6866002), a 579-582bp region surrounding the variant distinguishing *DEFA1* from *DEFA3* (chr8: 6822765-6823346; 6841882-6842463; 6860985-6861562) or a 444-521bp region surrounding the *DEFA1A3* microsatellite (chr8:

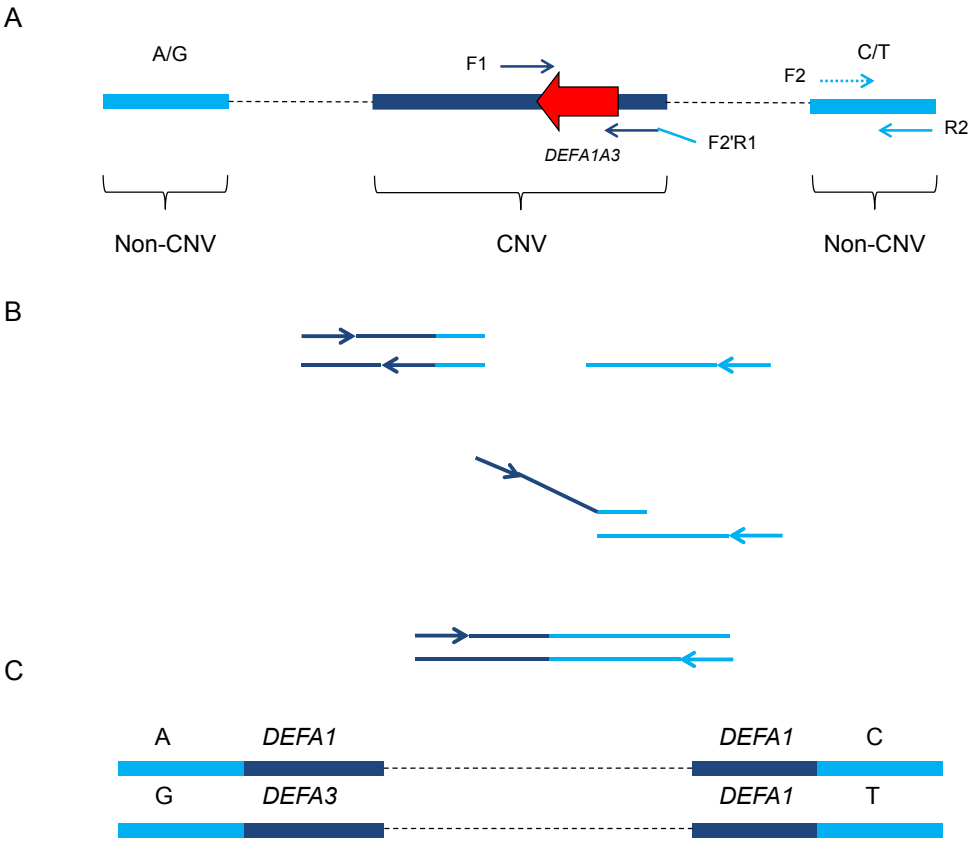


Figure 2.8: A: In EHF-PCR, the primers F1 and R1 are designed to amplify a region of interest from a CNV locus; here, this is the variant distinguishing *DEFA1* from *DEFA3*. This is fused to a non-CNV flanking region containing an informative variant- the primers F2 and R2 are designed for this region. The F2 primer is not used in the reaction, but the reverse complement is added as a tail to the R1 primer. B: Over the first few cycles of the PCR, the CNV region is exponentially amplified and linear amplification is achieved at the flanking region. The top strand of the CNV product can act as a primer with the product from the non-CNV region, creating a fused product. The tailed primer is used at a low concentration, such that the outer primers drive amplification of the fused product. C: Post-fusion allele-specific PCR and sequencing can allow the gene at the centromeric- or telomeric-most *DEFA1A3* repeat to be identified.

System	Primers 5′ to 3′	Regions fused
Telomeric Indel5	F1: TTGTGAAAACATGGCCCACCT F2′R1: GCTCTCCCTCTTCCTGGACAGTGAAAGCACCCATGTGCCTAGT R2: CCAGCTGGAGGGTCTCTGTTCT	chr8:6810649-6811322 to chr8: 6827670-6827801*
Centromeric Indel5	F1: CACTGTCCAGGAAGAGGGAGAGC F2′R1: TTCTCTAGCCCATCCTTGCAGGCCAGCTGGAGGGTCTCTGTTCT R2: TGGTGTTGGCTCAGCTGGAA	chr8: 6865876-6866002* to chr8: 6867863-6868058
Telomeric Gene	F1: TTGTGAAAACATGGCCCACCT F2′R1: TTGCAGAATACCAGCGTGCATTAAAGCACCCATGTGCCTAGT R2: CGGGAGAGAGGTTCCAGAGTT	chr8:6810649-6811322 to chr8: 6822765-6823346*
Centromeric Gene	F1: GCAGTGGGTGGGGAAATCAG F2′R1: TTCTCTAGCCCATCCTTGCAGGCGGGAGAGAGGTTCCAGAGTTG R2: TGGTGTTGGCTCAGCTGGAA	chr8: 6860985-6861562* to chr8: 6867863-6868058
Telomeric Microsatellite	F1: GGCATGGAATTGTGGACTCT F2R1′: TAGATCTGTCACCCACGCTGTTGTGAAAACATGGCCCACCT R2: AAAGCACCCATGTGCCTAGT	chr8:6810649-6811322 to chr8: 6818759-6819257*
Centromeric Microsatellite	F1: GGCATGGAATTGTGGACTCT F2R1′: TAGATCTGTCACCCACGCTGCCTGCAAGGATGGGCTAGAGAA R2: TGGTGTTGGCTCAGCTGGAA	chr8: 6867863-6868058 to chr8: 6857017-6857477*
Telomeric Microsatellite rs2738046	F1: GGCATGGAATTGTGGACTCT F2R1′: TAGATCTGTCACCCACGCTGGATGCTCCGTCCTCATCT R2: AAAGCACCCATGTGCCTAGT	chr8: 6810924-6811332 to chr8: 6818759-6819257*

Table 2.6: Primers used for the seven EHF-PCR assays and coordinates of the two regions fused together. * indicates the telomeric or centromeric-most matches to the primers, for the CNV regions fused, but fusion can occur to any copies of *DEFA1A3* .

6818759-6819257; 6837914-6838374; 6857017-6857477). This gave seven assays: Telomeric Indel5, Centromeric Indel5, Telomeric Gene, Centromeric Gene, Telomeric Microsatellite, Centromeric Microsatellite and Telomeric Microsatellite rs2738046. This alternative Telomeric Microsatellite system was required for allele-specific reamplification at position rs2738046, in order to give products short enough to be sized using capillary electrophoresis. The Telomeric Gene and Centromeric Gene systems were designed by Jess Tyson. An example is shown in figure 2.9, with all primers used shown in table 2.6.

Figure 2.9: The sequence of the fusion product obtained from the Telomeric Indel5 EHF-PCR. Upper case= primers; / =fusion boundary; yellow= sequence variants; underlined= reamplification primers.

Emulsion preparation and PCR

The aqueous PCR component was made as described in section 2.1. This was aliquoted into 0.5ml PCR tubes, to which 50µl of silicone oil was added. The silicone oil used was as described by Turner and Hurles [146] and contained, per 100ml, 40ml silicone polyether/ cyclopentasiloxane (Dow Corning), 30ml cyclopentasiloxane/ trimethylsiloxysilicate (Dow

Corning) and 30ml AR20 silicone oil (Aldrich). A 3mm tungsten carbide bead (Qiagen) was placed in the lid of the PCR tube and the tube was closed so that it remained in the inverted position. The inverted tube was vortexed at speed 5 for 1 minute 30 seconds using a Vortex Genie 2, to give aqueous droplets approximately 5μM in diameter, as determined by light microscopy (Jess Tyson). The tubes were briefly centrifuged at low speed to collect the emulsion at the bottom of the tube.

The PCR cycling conditions for the Centromeric Indel5 system were:

98°C	30 seconds	1 cycle
98°C	30 seconds	
71°C	30 seconds	40 cycles
72°C	15 seconds	
72°C	5 minutes	1 cycle
4°C	Hold	

The PCR cycling conditions for the Telomeric Indel5, Centromeric Gene and Telomeric Gene systems were:

98°C	30 seconds	1 cycle
98°C	30 seconds	
71°C	30 seconds	40 cycles
72°C	1 minute	
72°C	5 minutes	1 cycle
4°C	Hold	

The PCR cycling conditions for the Telomeric Microsatellite, Telomeric Microsatellite rs2738046 and Centromeric Microsatellite systems were:

98°C	30 seconds	1 cycle
98°C	30 seconds	
69°C	30 seconds	40 cycles
72°C	1 minute	
72°C	5 minutes	1 cycle
4°C	Hold	

Recovery of the aqueous phase

Following the PCR, the emulsion was added to a clean 0.5ml tube and 25µl of Phusion 1xGC buffer was used to recover the remaining emulsion from the original tube, leading to a dilution of the fusion product. The aqueous phase was recovered using hexane extraction. 200µl of hexane was added to each emulsion and vortexed to mix thoroughly, before centrifugation at 13000g for 3 minutes. The hexane phase was removed and the process repeated twice for a total of three extractions. The samples were left to air dry for 10 minutes at room temperature.

Reamplification of fusion products

As the fusion products are present at a low concentration, reamplification is necessary to allow further downstream processing. The genotype information for the telomeric SNPs was obtained from the HapMap project [116], whilst genotypes for rs4512398 and rs17382102 were obtained during sequencing of the centromeric flanking region of *DEFA1A3* , as described in section 2.5. For the Telomeric Gene, Telomeric Indel5, Centromeric Gene and Centromeric Indel5 systems, allele-specific reamplifi-

SNP	Primers 5′ to 3′	PCR cycling conditions
rs2978951	CGTTCGATCACATGAGTGGA or GTTCGATCACATGAGTGGG with GGTTCCAGAGTTGGGTCTCA (Telomeric Gene) or TCTGGGTGCTGCATGTAAAGC (Telomeric Indel5)	95°C for 30 seconds 67°C for 30 seconds 70°C for 1 minute for 35 cycles
rs2738046	CTCATCTGCCCCCTTCCA or CTCATCTGCCCCCTTCCC with GGTTCCAGAGTTGGGTCTCA (Telomeric Gene) or TCTGGGTGCTGCATGTAAAGC (Telomeric Indel5)	95°C for 30 seconds 68.1°C for 30 seconds 70°C for 1 minute for 35 cycles
rs4512398 Centromeric Indel5	AGGCCAGCTCATGAGA or AGGCCAGCTCATGAGG with GAAGAGGGAGAGCGGGTG	95°C for 30 seconds 65.3°C for 30 seconds 70°C for 15 seconds for 35 cycles
rs4512398 Centromeric Gene	AGGCCAGCTCATGAGA or AGGCCAGCTCATGAGG with TGAAGCCCAAACTCCTGCTTG	95°C for 30 seconds 65.3°C for 30 seconds 70°C for 1 minute for 35 cycles
rs17382102 Centromeric Indel5	TTAACCACAGAAATTTACAAACAT or TTAACCACAGAAATTTACAAACAC with GAAGAGGGAGAGCGGGTG	95°C for 30 seconds 69°C for 30 seconds 70°C for 15 seconds for 35 cycles
rs17382102 Centromeric Gene	TTAACCACAGAAATTTACAAACAT or TTAACCACAGAAATTTACAAACAC with TGAAGCCCAAACTCCTGCTTG	95°C for 30 seconds 69°C for 30 seconds 70°C for 1 minute for 35 cycles

Table 2.7: Primers and PCR cycling conditions used for the allele-specific reamplification of the Gene and Indel5 EHF-PCR products.

cation was performed, using allele-specific primers for either rs2978951 or rs2738046 for telomeric fusions and rs4512398 or rs17382102 for centromeric fusions. All reamplifications used BIOTAQ, with the exception of the Centromeric Indel5 which used 10x LD buffer, and 1µl of fusion product in place of the genomic DNA input in a 20µl reaction. The primers and PCR conditions are shown in table 2.7. The reamplified fusion products were sequenced as described in section 2.1.

The Telomeric Microsatellite, Telomeric Microsatellite rs2738046 and Centromeric Microsatellite EHF-PCR products were reamplified using 10x LD buffer in a 20µl reaction, containing 1µl of the EHF-PCR product, in place of genomic DNA, and 1µM each primer. The forward primer for all reactions was the HEX-labelled primer used for the *DEFA1A3* microsatellite assay (HEX-5′-GGATCCAGGTGGAGTCTCA-3′; section 2.3). For the Telomeric Microsatellite system, this was used with either a non-allele specific reverse primer (5′-CCACTCATGTGATCGAACG-3′) at an annealing temperature of 60°C or allele-specific primers for the SNP rs2978951 (5′-CCTACTAGGCATTTTAAAGAAC-3′ or 5′-CCTACTAGGCATTTTAAAGAACT-3′), also at 60°C. For the Telomeric Microsatellite rs2738046 system, it was used with allele-specific primers for the SNP rs2738046 (5′-TGAGAAGTATCCATCAATCCAAT-3′ or 5′-TGAGAAGTATCCATCAATCCAAG-3′) at an annealing temperature of 64°C. For the Centromeric Microsatellite system, it was used with either a non-allele specific reverse primer (5′-TCCTAGAAAGCTGGTTGCTCA-3′) with an annealing temperature of 60°C, or allele-specific primers for the SNP rs4512398 (5′-AGGCCAGCTCATGAGA-3′ or 5′-AGGCCAGCTCATGAGG-3′) at 64°C. For all reamplifications, the PCR conditions were:

95°C	30 seconds	
60 or 64°C	1 minute	24 cycles
70°C	1 minute	
70°C	10 minutes	1 cycle

0.5µl of the PCR product was analysed using capillary electrophoresis at 1.5kV for 30 seconds, as described in section 2.1. The peak areas were used to determine the relative location of the microsatellite alleles; i.e. the centromeric-most allele should give the largest peak area for the Centromeric Microsatellite assay.

2.12 Quantifying *DEFA1A3* expression in neutrophils

The total expression of HNP-1, HNP-2 and HNP-3 in neutrophil cells was measured for the 120 BBSRC Project BB/I006370/1 volunteers by Danielle Carpenter and Laura Mitchell. A sandwich ELISA, using a polyclonal capture antibody (Abnova) and a biotinylated monoclonal detection antibody (Hycult biotech), was used quantify HNP1-3 expression per 10^6 neutrophils. A dilution series of purified HNP1-3 of known concentration (Hycult biotech) was used to generate a standard curve, from which the amount of HNP1-3 in each of the BBSRC Project BB/I006370/1 volunteer samples was quantified.

2.13 Statistical Analysis

Comparing features of the *DEFA1A3* locus with *DEFA1A3* haplotype class

Chi square or Cochran-Armitage tests were used to compare features of the *DEFA1A3* locus with *DEFA1A3* haplotype class. The copy number and frequency categories were designated such that each category was comparably populated. For Class 1 and Exchange 2, homozygous and heterozygous were grouped, due to their low frequencies. Individuals were counted multiple times; for example, an individual homozygous positive for the Reference Sequence would have also been counted as homozygous negative for Class 1, Class 2, Exchange 1 and Exchange 2. To account for this, p-values were adjusted using Bonferroni correction, using the formula:

$$1 - ((1 - p)^n$$

where n= number of tests performed. All tests were performed in Excel using programs written by John Armour.

Comparing HNP1-3 expression with *DEFA1A3* haplotype class

A Kolmogorov-Smirnov test was used to show that the HNP1-3 expression data did not conform to a normal distribution (p=0.001). Therefore, non-parametric tests were used to compare HNP1-3 expression and *DEFA1A3* haplotype class. For Reference Sequence and Class 2, a Kruskal-Wallis test was used, whereas for Class 1, Exchange 1 and

Exchange 2, a Mann-Whitney U test was used. All analysis was done using SPSS [149].

3 *DEFA1A3* copy number and haplotypes

At multiallelic copy number variable loci, the combination of the two haplotypes, rather than the diploid total, is likely to explain the effect of copy number variation on the expression and function of the genes involved [111]. In addition, in order to study the evolutionary processes shaping CNVRs, it is necessary to have haplotype information, as a diploid total could be composed of many different haplotype combinations. This is particularly true of the *DEFA1A3* locus, at which there is variation not only in the total number of copies, but the copy number of each of the *DEFA1* and *DEFA3* genes. Therefore, previous studies at the *DEFA1A3* locus have focused on obtaining haploid, and not just diploid, copy number information. Khan *et al.* [50] used segregation in three-generation CEPH pedigrees to infer the *DEFA1A3* haplotype copy numbers for 151 independent haplotypes. In cases where a three-generation pedigree resource is unavailable, PFGE can be used to determine *DEFA1A3* haplotype copy numbers, as shown by Aldred *et al.* [55].

In addition, information about the internal variation within a CNV locus can be useful, in terms of comparing the structures of related haplotypes, to infer mechanisms of change in copy number, and in the estimation of a mutation rate for the locus. Therefore, previous work at the *DEFA1A3* locus has not only included looking at the ratio of the *DEFA1* to *DEFA3* genes, but also the ratio of two additional internal variants- a

5bp insertion located upstream of each copy of the gene (Indel5) and a 7bp duplication located in intron 1 of the gene (7bp dup) [50]. These ratios also provide a useful measure of copy number. The aim was to add to the *DEFA1A3* haplotype information available, both in terms of determining the haplotype copy numbers of additional samples and looking at additional variation within the repeat units. This would provide a more detailed dataset from which the *DEFA1A3* mutation rate, as well as mechanisms of copy number change, could be investigated. In addition, as the inference of haplotype information relies on the ability to first determine the diploid copy number of a sample accurately, the aim was to investigate alternative methods of determining diploid *DEFA1A3* copy number.

3.1 Pulsed-field Gel Electrophoresis

As described by Aldred *et al.* [55], the digestion of genomic DNA with the enzyme *Hpa*I allows the entire *DEFA1A3* locus to be cut out on a single restriction fragment (figure 3.1A). The separation of the digested genomic DNA sample using PFGE, followed by Southern blotting with a probe targeting the *DEFA1A3* locus, allows the *DEFA1A3* haplotype copy numbers of the sample to be determined. However, Southern blotting is a process which lacks sensitivity; whilst the *DEFA1A3* haplotype copy numbers could be determined using this method, information about the internal content of each haplotype could not be deduced. The aim was to use PFGE to separate the two *DEFA1A3* haplotypes, but to analyse the separated gel fragments using PCR-based techniques, in order to gain information about both the haplotype copy number and internal variation of each *DEFA1A3* haplotype.

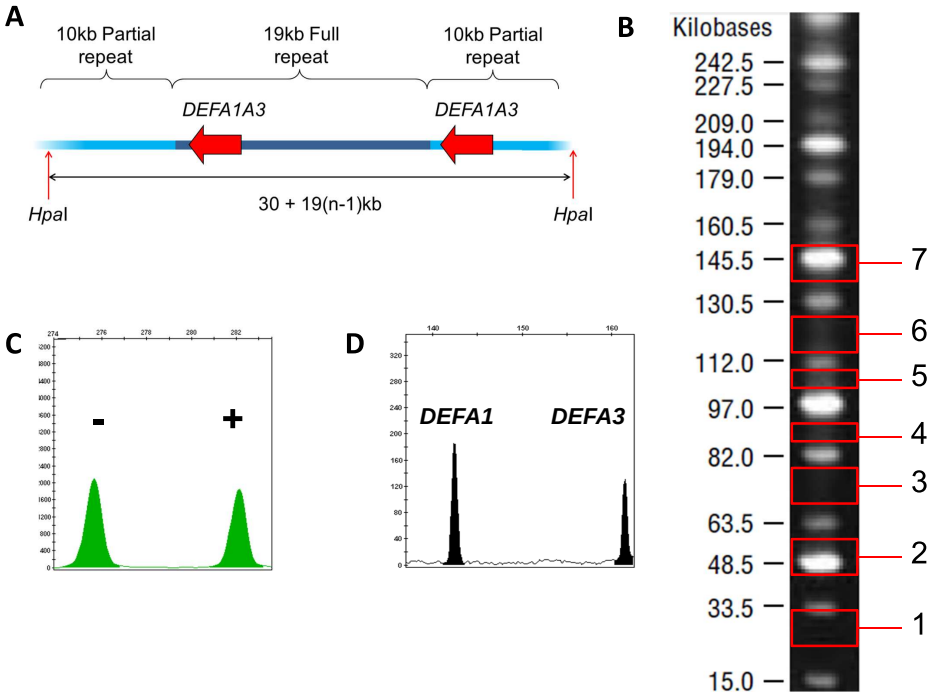


Figure 3.1: A: A *Hpa*I digest cuts out the entire *DEFA1A3* repeat array on a single restriction fragment, with length as shown, where n= the number of copies of *DEFA1A3* . The enzyme does not cut within the *DEFA1A3* array. B: Following PFGE, gel blocks were sampled at the positions shown, using the MidRange I PFG marker as a guide (image adapted from New England Biolabs). Each block represents the point expected to be occupied by a *DEFA1A3* haplotype with between 1 and 7 copies. Blocks were sampled between these regions to act as controls, as they would not be expected to contain a *DEFA1A3* haplotype. C and D: Ratios from gel blocks do not always fit the expected copy number or ratio. C: The 7bp duplication ratio for gel block 5B from sample NA07008 has a ratio of 1.2, where a 4:1 ratio is expected. D: The Indel5 ratio for gel block 1A from sample NA06990 has a ratio of 1.6, where the haplotype is expected to be *DEFA1* absent.

In order to assess the viability of this method, two samples, for which the *DEFA1A3* haplotype copy numbers and haplotype ratios for three internal variants were known from segregation, were analysed. In theory, the *Hpa*I-digested DNA should occupy the regions of the gel corresponding to the size of their two haplotypes, whilst other regions of the lane should

not contain DNA from the *DEFA1A3* region (figure 3.1B). However, a PCR assay designed to detect DNA from the *DEFA1A3* region showed that most regions of the gel, and not just those expected to be occupied by the *DEFA1A3* haplotypes, contained DNA from the *DEFA1A3* region, although blank lanes were *DEFA1A3* DNA absent. This suggests that, not only is some DNA being retarded in the gel, but some of the *DEFA1A3* haplotypes are on restriction fragments shorter than would be expected, probably due to the short length of the input genomic DNA fragments. Therefore, a simple positive/negative PCR assay is unsuitable for determining the *DEFA1A3* haplotype copy numbers of a sample following PFGE.

However, it would be assumed that the ratio of internal *DEFA1A3* features would match the copy number for regions of the gel expected to contain a *DEFA1A3* haplotype, whereas other regions of the gel would have ratios that did not correspond to the haplotype copy number expected in that region. The ratios from three allelic ratio systems were obtained for the same two samples, as shown in tables 3.1 and 3.2. As observed previously, most regions of the DNA lane contain DNA from the *DEFA1A3* region. For sample NA07008 (table 3.1), the absence of *DEFA3* from many regions of the gel may suggest there is a *DEFA3* absent haplotype, although this is not the case. Whilst the DefHae3 and Indel5 ratios may allow the deduction of a 5-copy haplotype, with both ratios close to 4:1, the 7bp dup ratio of 1.2 does not fit with the 4:1 ratio predicted by segregation (figure 3.1C). It is not possible to identify that the other haplotype has 2 copies, but if this is assumed, the Indel5 and 7bp duplication ratios fit what is expected, whilst the DefHae3 ratio differs considerably from the expected 1:1 ratio.

NA07008	DefHae3 Ratio	Indel5 Ratio	7bp dup Ratio
Diploid ratios	5:2	6:1	5:2
Haplotype ratios from segregation			
Haplotype A	4:1	4:1	4:1
Haplotype B	1:1	2:0	1:1
Ratios for each gel block			
7 copies	*DEFA1* only	Deletion only	Unduplicated only
6-7	-	-	3:1
6	-	Deletion only	1.9
5-6	*DEFA1* only	Deletion only	3
5B	3.4	3.3	1.2
5A	-	2.7	-
4-5	*DEFA1* only	Deletion only	1.6
4	*DEFA1* only	Deletion only	3.2
3-4	*DEFA1* only	Deletion only	2.3
3	*DEFA1* only	Deletion only	3.4
2-3	-	Deletion only	3.0
2C	2.4	Deletion only	1.6
2B	-	Deletion only	1.1
2A	1.5	Deletion only	1.2
1-2	*DEFA1* only	Deletion only	0.9
1	-	3.3	1.9
<1	-	Deletion only	2.6

Table 3.1: Ratios from three allelic ratio assays on gel samples from HapMap individual NA07008 following PFGE. Gel blocks were sampled at and between the regions expected to be represented by a *DEFA1A3* haplotype with between 1-7 copies, with the region expected to be occupied by a haplotype separated into 2-3 smaller blocks to allow a more detailed analysis (labelled A-C). - indicates the absence of a peak or peaks at too low a level to call. If only one allele is observed or can be called, the allele present is stated.

NA06990	DefHae3 Ratio	Indel5 Ratio	7bp dup Ratio
Diploid ratios	2:1	2:1	2:1
Haplotype ratios from segregation			
Haplotype A	2:0	1:1	1:1
Haplotype B	0:1	1:0	1:0
Ratios for each gel block			
7	1.7	-	Unduplicated only
6-7	*DEFA1* only	-	Duplicated only
6	*DEFA1* only	-	Unduplicated only
5-6	*DEFA1* only	-	-
5	*DEFA1* only	-	Unduplicated only
4-5	*DEFA1* only	0.8	-
4	1.1	-	Unduplicated only
3-4	1.2	2.3	1.0
3	-	0.9	1.3
2-3	1.0	1.0	1.2
2C	-	0.8	2.3
2B	8.1	1.1	1.1
2A	5.4	1.2	1.7
1-2	2.8	2.2	1.0
1C	0.6	3.3	3.8
1B	0.4	4.6	3.8
1A	1.6	4.3	2.1
$<$1	0.9	4.9	6.2

Table 3.2: Ratios from three allelic ratio assays on gel samples from HapMap individual NA06990 following PFGE. Gel blocks were sampled at and between the regions expected to be represented by a *DEFA1A3* haplotype with between 1-7 copies, with the region expected to be occupied by a haplotype separated into 2-3 smaller blocks to allow a more detailed analysis (labelled A-C). - indicates the absence of a peak or peaks at too low a level to call. If only one allele is observed or can be called, the allele present is stated.

For sample NA06990 (table 3.2), given that it has only 3 copies of *DEFA1A3*, it could be assumed that the haplotype copy numbers are 1 and 2. In the 1-copy region, the Indel5 ratios and 7bp dup ratios suggest a 1:0 ratio, whereas it is unclear that the haplotype is *DEFA1* absent, given the presence of *DEFA1* DNA in the 1-copy region (figure 3.1D). It is more obvious that the 2-copy haplotype contains only *DEFA1*, and the 1:1 ratios for the Indel5 and 7bp dup assays can be identified. However, a certain amount of prior knowledge about the haplotype composition is required in the interpretation of the results; the haplotype copy numbers and the ratios of variants on each haplotype cannot both be deduced correctly. Therefore, PFGE followed by PCR-based analysis failed to identify both the haplotype copy numbers and haplotype allelic ratios for a sample.

3.2 Segregation in three-generation families

An alternative method for obtaining *DEFA1A3* haplotype copy numbers involves the segregation of diploid copy number information in three-generation families; this has previously been applied to 21 CEPH families [50]. 3rd-generation individuals from an additional 3 CEPH families were typed for *DEFA1A3* copy number (table 3.3), allowing the identification of an additional 24 haplotype *DEFA1A3* copy numbers and haplotype *DEFA3* and Indel5 insertion frequencies (figure 3.2 and table 3.4). For some samples, the minimum ratio value is below 10, indicating there is less confidence in the assigned copy number. However, in the case of sample NA07053, for which the Indel5 ratio is clearly 2:1, the diploid copy number must be a multiple of three, giving confidence that the copy number is 12, and not 11 or 13, for example. In addition, the use of segregation, in which at least three different haplotype

combinations are observed in the 3rd generation, constrains the possible haplotype copy numbers, adding confidence to both the haploid and diploid copy numbers assigned to the 2nd and 3rd generation individuals.

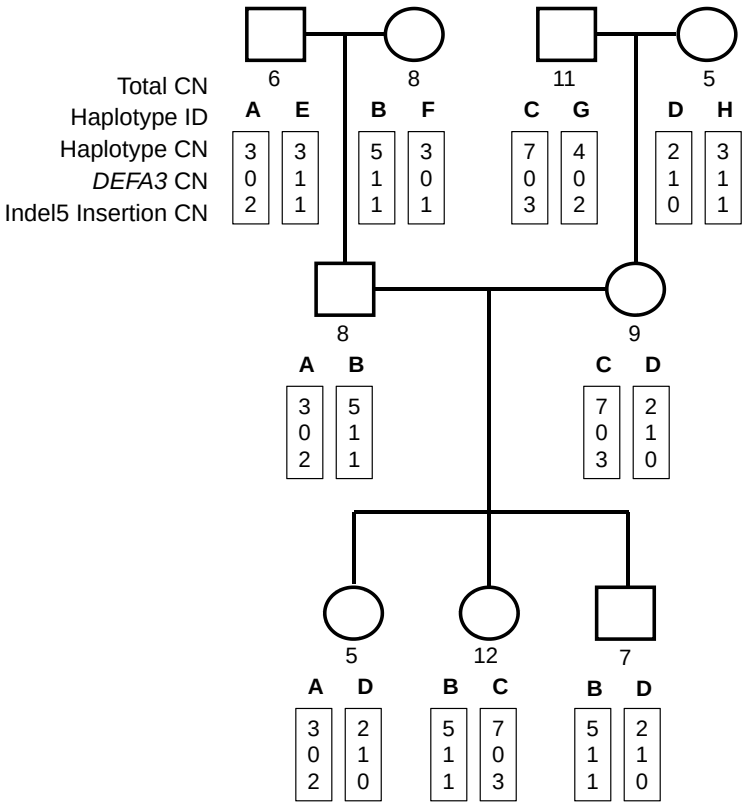


Figure 3.2: The segregation of *DEFA1A3* diploid copy number into the haplotype components shown for the CEPH family 1340. Each of the eight first-generation haplotypes is assigned a letter A-H and linkage data was used to show the transmission pattern of haplotypes A-D to the third generation. This allows the copy number of haplotypes A-D to be determined, with the copy number of haplotypes E-H inferred from the diploid copy number of the first-generation individuals. *DEFA3* and Indel5 insertion copy number can also be split into the haplotype components, as shown.

Sample	MLT1A0	DEFA4	DefHae3	Indel5	MLCN	MinRatio
NA07062	5.09	4.81	3.89	1.58	5	349.94
NA07053	12.67	11.63	9.91	2.03	12	3.50
NA07008	7.39	6.69	2.34	6.01	7	24.83
NA11998	6.28	6.01	5.15	1.06	6	41.16
NA11999	6.17	5.77	*DEFA1*	2.09	6	41.99
NA12000	6.05	6.43	5.33	5.03	6	9.87
NA12806	9.24	9.16	7.81	1.36	9	4.19
NA12807	7.38	7.14	*DEFA1*	1.45	7	7.72
NA12810	4.66	5.49	3.91	Deletion	5	16.40

Table 3.3: The unrounded copy number values for the two PRTs (MLT1A0 and DEFA4) and the ratios from two allelic ratio assays (DefHae3 and Indel5), used to determine the diploid *DEFA1A3* copy number for nine 3rd-generation individuals from the CEPH families 1340, 1420 and 1454. The maximum likelihood copy number (MLCN) and Minimum ratio values, from the output of the DEFAML program, are also shown.

Sample	Haplotype A			Haplotype B		
	CN	*DEFA3*	Indel5 ins	CN	*DEFA3*	Indel5 ins
NA07062	3	0	2	2	1	0
NA07053	5	1	1	7	0	3
NA07008	5	1	1	2	1	0
NA11998	3	1	1	3	0	2
NA11999	3	0	0	3	0	2
NA12000	3	0	0	3	1	1
NA12806	4	0	3	5	1	1
NA12807	4	0	3	3	0	0
NA12810	2	1	0	3	0	0

Table 3.4: Haplotype copy numbers, *DEFA3* copy number and Indel5 insertion copy number for the nine 3rd generation individuals from the CEPH families 1340, 1420 and 1454, obtained through segregation.

3.3 Using Read Depth to estimate *DEFA1A3* copy number

The deduction of haplotype *DEFA1A3* copy numbers relies on the ability to first obtain accurate diploid copy number calls. Previous measurements of diploid copy number at the *DEFA1A3* locus have used a combination of 2 PRTs and 3 allelic ratio assays to accurately call *DEFA1A3* copy number (section 2.2 [50]). Next-generation sequencing data provides an additional measure of copy number at multiallelic loci, as the number of reads mapping to a copy number variable locus should be

proportional to the copy number of the locus. This can be compared to the number of reads mapping to a two-copy reference locus. For 98 of the HapMap CEU samples, whole genome sequencing data is available via the 1000 Genomes project [20]. The diploid *DEFA1A3* copy number estimated using read depth was compared to the copy numbers assigned by Khan *et al.* [50] (figure 3.3). An r^2 value of 0.8715 shows a strong correlation between the two measures, with 70% of integer copy number values in agreement. Of the 29 samples for which read depth predicts a different copy number to PRT-based methods, 24 differ by only a single integer value. This both corroborates the accuracy of the PRT-based methods and suggests that read depth is a useful tool for measuring *DEFA1A3* copy number. The *DEFA1A3* copy number distributions for each population in the 1000 Genome dataset are shown in figure 3.4. These distributions are similar to those observed previously [50, 55, 85, 108], with a *DEFA1A3* copy number range from 3-17 copies.

It would be assumed that reads covering variants across the *DEFA1A3* region, such as the variant distinguishing *DEFA1* from *DEFA3*, would also be proportional to the frequency of each allele. For 45 independent CEU1 individuals, the reads displaying either the G (*DEFA1*) or T (*DEFA3*) alleles were counted. The ratio was compared to that obtained by Khan *et al.* (figure 3.5) [50]. This shows a weak correlation between the two ratios, with an r^2 of only 0.3794. In addition, there is an example of an 8:1 ratio which appears as *DEFA3* absent in read depth analysis, due to a lack of reads displaying the *DEFA3* allele. For this sample, there were only 15 reads covering the variant. This suggests that the low coverage sequencing prevents detailed analysis of any one variant, as there are too few reads covering the position.

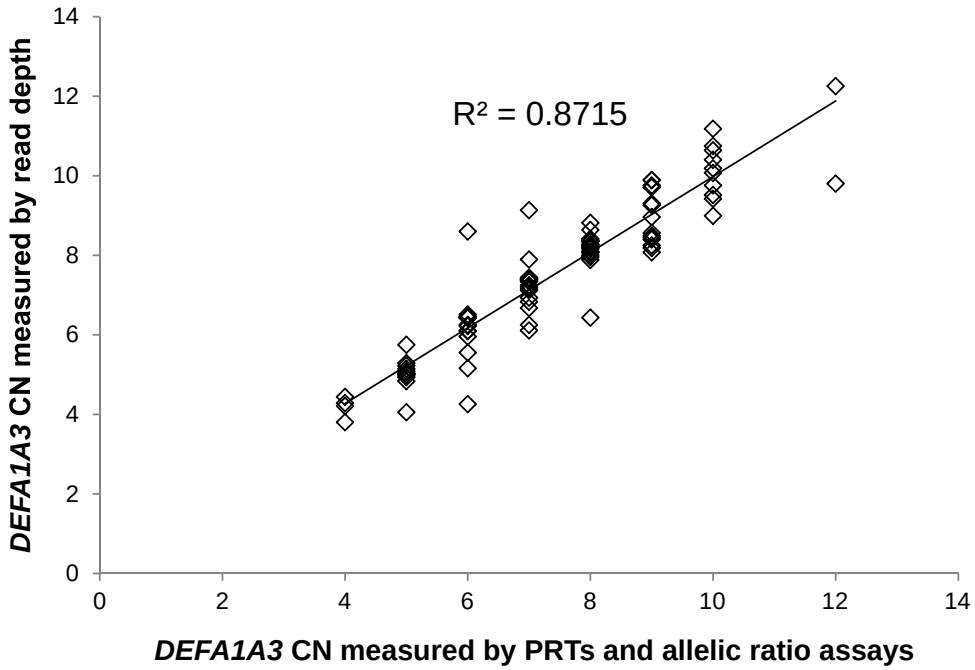


Figure 3.3: The *DEFA1A3* diploid copy numbers assigned by Khan *et al* [50] plotted against the unrounded diploid *DEFA1A3* copy number estimated using read depth for the 98 CEU samples included in the 1000 Genomes project [20].

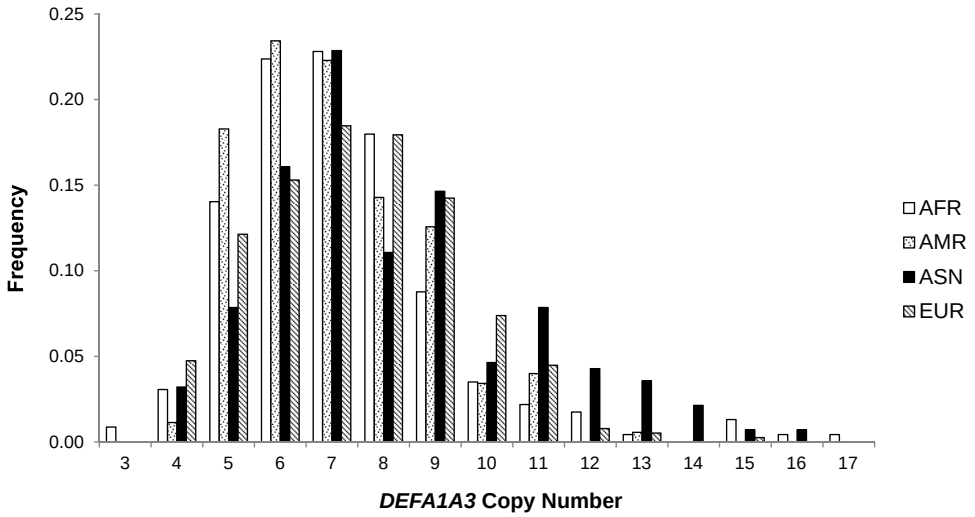


Figure 3.4: The *DEFA1A3* copy number distributions for the 1062 samples in the 1000 Genome project, separated by population [20]. AFR= samples with African ancestry, AMR= samples with American ancestry, ASN= samples with Asian ancestry, EUR= samples with European ancestry.

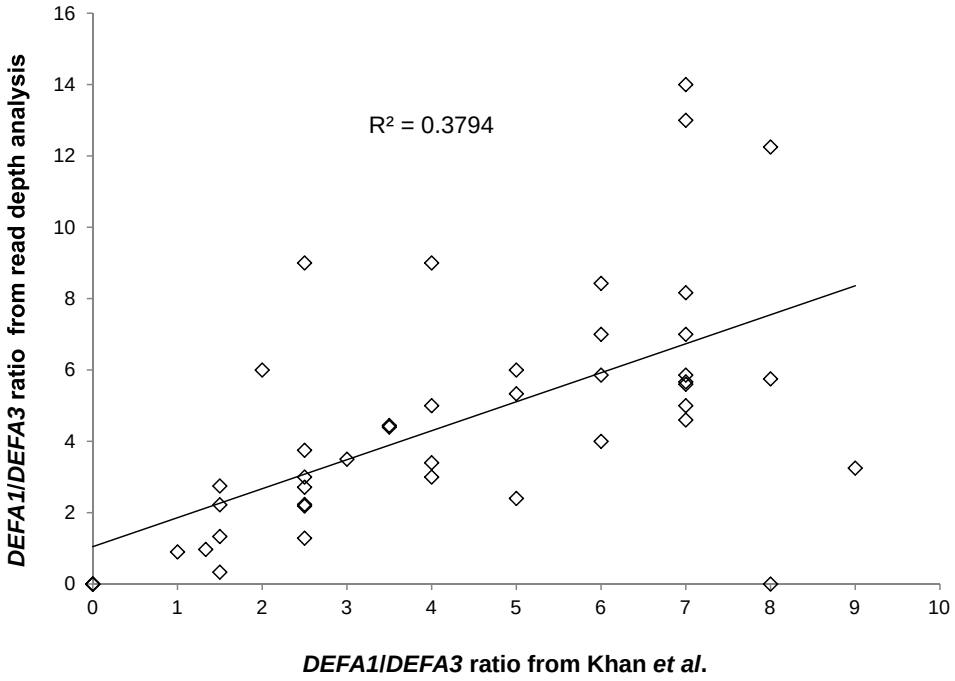


Figure 3.5: *DEFA1* vs. *DEFA3* ratio for 45 HapMap CEU1 individuals estimated using read depth analysis [20] compared to the values assigned by Khan *et al.* [50].

3.4 *DEFA1A3* microsatellite

Allelic ratio assays provide an accurate method for measuring multiallelic copy number, because the amplification differences between the different alleles should be minimised, as they are found in the same sequence context and, as shown for the Indel5 assay, the difference in amplification efficiency between the different products occurs at a consistent rate. In addition, different allelic variants within the *DEFA1A3* repeat array could act as tags of variants affecting the expression and function of the *DEFA1A3* genes. Internal variants are also useful for comparing the structures of related haplotypes, from which the mechanisms responsible for changes in copy number can be inferred.

Therefore, an allelic ratio assay was designed to measure the allele sizes

of a microsatellite located downstream of each copy of the *DEFA1A3* gene, to provide additional information about the internal sequence variation within the *DEFA1A3* repeat array and to potentially provide an additional measure of *DEFA1A3* copy number (figure 3.6). This assay was used to measure the microsatellite allele sizes for individuals from 24 CEPH families and additional trios from the HapMap CEU1 population (total 331 individuals). This identified 13 different allele sizes for the microsatellite, with product sizes ranging from 351-428bp, corresponding to microsatellite sizes of 43-120bp. The *DEFA1A3* microsatellite is highly variable, with each individual having between 2 and 6 different length microsatellite alleles.

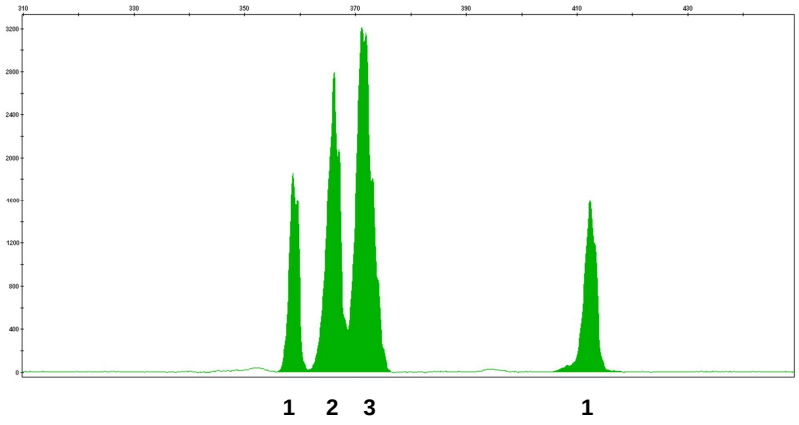


Figure 3.6: *DEFA1A3* microsatellite capillary electrophoresis trace for the CEPH sample 1420-10. This sample has 7 copies of *DEFA1A3* and the microsatellite trace shows four peaks, the areas of which are in the ratio 1:2:3:1, with allele sizes of 51, 57, 66 and 106bp.

Segregation in three-generation families

For all 24 CEPH families, the diploid microsatellite profiles were segregated into the two component haplotypes. For the majority of families, due to the high variability of the microsatellite, the segregation pattern

could be determined using the microsatellite data alone, with no prior inference of the haplotype combination inherited by the children. This allowed a confirmation of the expected inheritance patterns and, in some cases, allowed the identification of incorrectly assigned haplotype combinations. In addition, the haplotype copy numbers could be deduced for an additional trio, from family 1444, for which a three-generation resource is unavailable, providing copy numbers for an additional four haplotypes (figure 3.7).

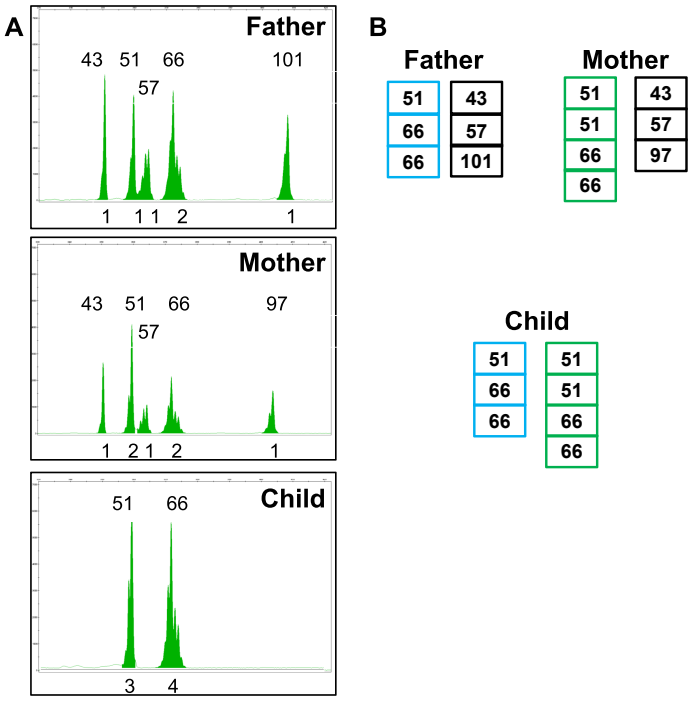


Figure 3.7: *DEFA1A3* microsatellite profiles (A) and haplotypes (B) for a trio from the CEPH family 1444. A: Only the 51bp and 66bp alleles are transmitted to the child. The peak ratios show all copies of the 51bp and 66bp alleles from the first generation must be transmitted and the other alleles, 43, 57 and 101bp for the Father and 43, 57 and 97bp for the Mother, must reside on the untransmitted haplotype. B: The diploid microsatellite profiles split into the two haplotypes.

Using the *DEFA1A3* microsatellite to estimate copy number

The accuracy of the ratios obtained using the *DEFA1A3* microsatellite assay was assessed by calculating the normalised ratios of each pair of peaks; the ratio obtained between the two peaks was divided by the expected ratio (figure 3.8), in which the expected ratios were based on the total diploid copy number determined independently of the microsatellite. For ratios of 1:1, all minimum ratio values are 1.0 or greater, as the observed ratio between two peaks was always calculated by dividing the larger peak area by the smaller peak area. For all other ratios, the mean is 1.03, which is close to the expected mean of 1. The standard deviation was 0.135, which suggests that the assay is able to accurately measure the ratio of different microsatellite alleles.

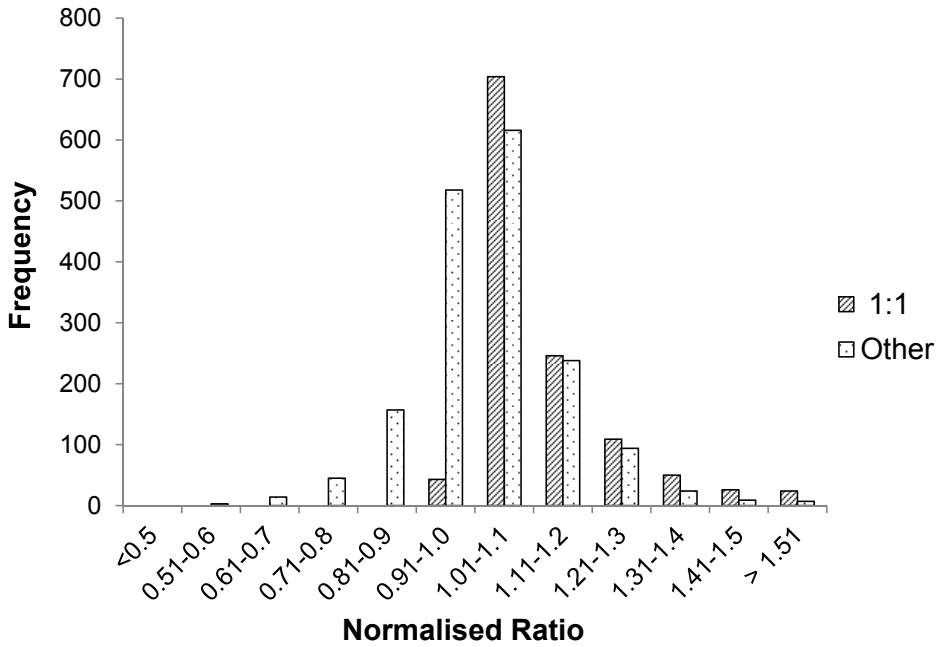


Figure 3.8: The normalised ratio values obtained by comparing the observed and expected ratios between pairs of peaks obtained for the *DEFA1A3* microsatellite expected to have either a 1:1 ratio or any other ratio.

As there is one copy of the *DEFA1A3* microsatellite per copy of *DEFA1A3*, the microsatellite could potentially provide a copy number typing assay. In order to assess the ability of the *DEFA1A3* microsatellite to determine *DEFA1A3* copy number, the peak areas for each sample were assessed using the microsatellite peak analysis program, which assigns the most likely diploid copy number represented by the peak area ratios. This was compared to the maximum likelihood copy number estimated by Khan *et al.* (figure 3.9) [50]. This shows that for the majority of samples, the microsatellite peak analysis program assigns the same copy number as predicted using the PRT and allelic ratio assays. At higher copy number values, the microsatellite peak analysis program has a tendency to assign a MLCN one integer higher or lower than expected. This is likely due to the fact that high ratios between two peaks, for example a 7:1 ratio, can easily appear as 6:1 or 8:1, even when the standard deviation is low. As these higher ratios are more likely to appear in samples with a higher copy number, it leads to some samples being assigned a copy number one higher or lower than expected. 88% of individuals have been assigned either the same copy number or a copy number differing by a single integer. In addition, many samples are assigned a MLCN of 12, the maximum allowed by the program, as the deviations of the ratios from what would be expected leads to the program estimating a copy number much higher than its actual value. Despite this, the two samples with a copy number of 12 are assigned correctly; this gives the microsatellite a better ability to accurately assign high copy numbers in cases where a sample has many peaks, because each of the ratios being measured is low. This is in comparison to PRTs or a two-allele allelic ratio system, in which an increase in copy number often leads to an increase in the observed ratio. Therefore, the *DEFA1A3* microsatellite alone is

not as accurate as the two PRT and three allelic ratio assays combined at determining diploid *DEFA1A3* copy number. However, in combination, it provides an independent measure of copy number which can help to clarify the predicted copy number and provide additional information about the internal repeat array variation.

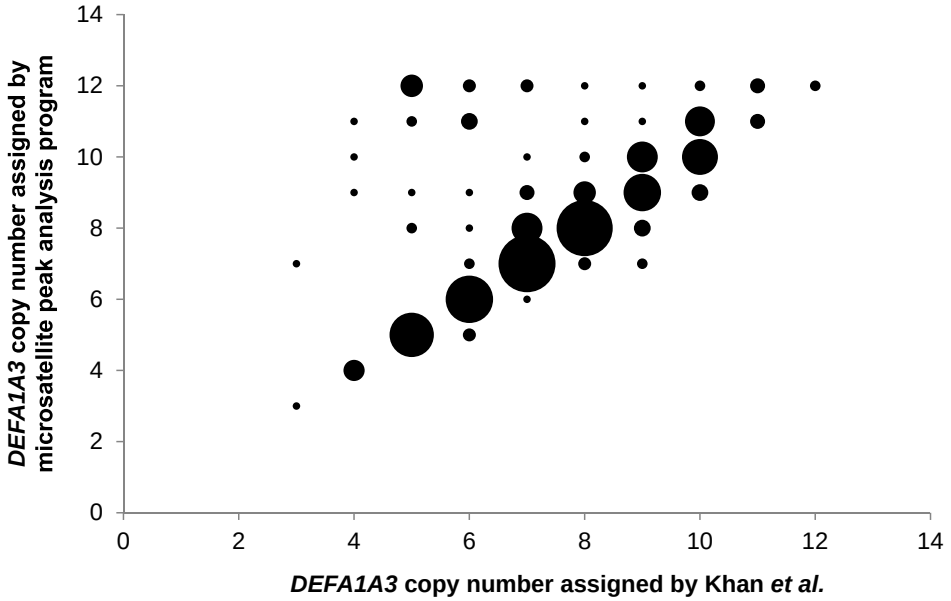


Figure 3.9: The diploid *DEFA1A3* copy number assigned by Khan *et al.* [50] compared to the copy number assigned by the *DEFA1A3* microsatellite peak analysis program. The area of points are scaled to represent the frequency of individuals.

3.5 Estimating the *DEFA1A3* mutation rate

There is extensive variation in *DEFA1A3* copy number; however, in the 302 transmissions observed across the CEPH families, no examples of *de novo* copy number changes have been observed at *DEFA1A3*. Therefore, two alternative indirect method were used to estimate the *DEFA1A3* mutation rate.

Ewens sampling formula follows the infinite alleles model, in which each mutation is expected to create a new allele. The formula is:

$$E(k) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + ... + \frac{\theta}{\theta+(k-1)}$$

where k= number of alleles sampled and E(k)= expected number of different alleles [150]. *DEFA1A3* haplotypes were considered to be the same allele if they shared the same *DEFA1A3* copy number, *DEFA3* copy number, Indel5 insertion copy number and 7bp duplication copy number. Within the CEPH families, 78 independent individuals were sampled (k= 156 haplotypes), across which 44 different *DEFA1A3* alleles (E(k)=44) were observed. For these values, Ewens sampling formula estimates the population diversity parameter, θ, as 20.05. Using the formula $\theta=4N_e\mu$, with an effective population size (N_e) of 10,000, the per generation mutation rate at *DEFA1A3* can be estimated as μ= 5x10^{-4}, or one mutation at *DEFA1A3* per 2000 transmissions.

However, Ewens formula is likely to underestimate the mutation rate at the *DEFA1A3* locus, because it is possible that the same allele could arise more than once independently. Therefore, the *DEFA1A3* mutation rate was also calculated using the stepwise mutation model, originally used to calculate the mutation rate for microsatellites, in which alleles differ by a fixed size repeat unit and alleles of the same size can appear more than once and not be identical by descent. The equation used is:

$$p_0 = \sqrt{\frac{1}{1+2\theta}}$$

in which p_0= frequency of homozygotes [151]. Out of 78 individuals,

6 homozygotes were observed (p_0=0.08). This gives θ= 81.85, which equates to a per generation mutation rate of μ= 2x10^{-3}, or 1 mutation per 500 transmissions.

3.6 Conclusions

The use of PCR-based analysis following PFGE appears to be unsuitable for determining *DEFA1A3* haplotype copy numbers and the haplotype ratios of internal sequence variants. The major limitation to this technique appears to be the length of DNA in solution. This makes it difficult to accurately assign haplotype *DEFA1A3* copy numbers and, even if this is possible, the ratios of internal variants are distorted, such that it is hard to identify the true ratio. Contamination from endemic PCR products from the *DEFA1A3* locus will also contribute to this, especially when using what is, post-PFGE, small amounts of genomic DNA and large cycle numbers for the allelic ratio assays. In addition, PFGE uses a large amount of DNA (2µg), in comparison to the 10ng required for each of the five assays (2 PRTs and 3 allelic ratio assays) used to assign a diploid *DEFA1A3* copy number. It is clear that these five assays provide an accurate and consistent method for determining *DEFA1A3* copy number and they have been successfully applied to an additional three CEPH families, with segregation providing an additional 24 haplotype copy numbers. However, a three-generation resource is not always available; the CEPH pedigrees are, in many ways, unique in terms of the number of families and the number of individuals per family available. Secondly, the transmission of haplotypes was inferred using publicly available linkage data; the generation of this information can be time-consuming and adds an additional cost to the process. Therefore, whilst

the three-generation CEPH pedigrees were very useful in allowing the identification of haplotype *DEFA1A3* copy numbers, this is not always a realistic option.

The *DEFA1A3* microsatellite can, in some cases, allow the identification of haplotype copy numbers without any additional information, due to its high variability. However, this is not a reliable method, as it depends on the variability of individuals and the inheritance pattern of the different alleles. Whilst the *DEFA1A3* microsatellite alone is not the most accurate method for determining diploid *DEFA1A3* copy number, in combination with the other assays, it is very useful in confirming a copy number, particularly in cases of high copy numbers where individuals have several different microsatellite alleles. The poor ability of the microsatellite to deal with large ratios appears to be due in part to the unequal amplification of different alleles, due to both their length and possibly their sequence. Whilst it may be possible to correct for this and in turn make the ratios more reflective of the true copy number, it would be a difficult process, because the context in which the peak was observed would differ from case to case. In addition, PCR leads to an underrepresentation of the largest products, which would also need taking into account. Read depth provides an additional measure of *DEFA1A3* copy number and there is strong correlation between the copy number estimated using this method and those assigned by Khan *et al.* [50]. However, the low coverage prevents ratio analysis of internal variants, something which will be useful for further analysis at the *DEFA1A3* locus.

The *DEFA1A3* microsatellite is particularly valuable as an internal sequence variant, due to its high variability, which will aid the structural analysis of *DEFA1A3* haplotypes. In total, using segregation and the

DEFA1A3 microsatellite, 179 *DEFA1A3* haplotype copy numbers have now been deduced across the CEPH families, for which the ratios of the DefHae3, Indel5 and 7bp duplication variants are also known. This provides a comprehensive dataset from which further analysis can be completed.

One use of the dataset came in the estimation of the *DEFA1A3* mutation rate. This process did not take into account the microsatellite alleles when assigning haplotypes, as it would be expected that the microsatellite alleles are prone to *in situ* changes in length at a higher rate than changes in copy number, and, as such, would not provide an accurate representation of the *DEFA1A3* mutation rate. By classifying alleles using three internal variants, it is possible for mutational processes not changing the *DEFA1A3* copy number to change an allele; this may occur via gene conversion, for example, changing a *DEFA1* gene to a *DEFA3*. However, it was important to classify alleles using more than just the haplotype copy number, which would not represent the wide variation observed and would lead to an underestimation of the mutation rate, as it would assume all haplotypes with the same copy number were more closely related to each other than to haplotypes with a different copy number, which is not necessarily the case. The true mutational landscape at *DEFA1A3* is likely to fall somewhere between the two models used, which gives a mutation rate of 1 mutation per 500-2000 transmissions, higher than that for surrounding sequence variants, such as SNPs. However, the rate is sufficiently low that an example amongst the 302 transmission events studied would be unlikely.

4 *DEFA1A3* haplotype classes

The *DEFA1A3* locus has been well characterised in terms of diploid and, in the case of the European population, haploid copy numbers [50]. In addition, there is extensive information about the ratios of different internal variants, including the ratios of *DEFA1* to *DEFA3*. However, it has not been determined how *DEFA1A3* haplotypes are related to each other. For example, it could be that all two-copy haplotypes are more closely related to each other than they are to haplotypes with a different number of *DEFA1A3* repeats. Alternatively, haplotypes with different *DEFA1A3* copy numbers, but other similar features (e.g. *DEFA3* content) could share a more recent common ancestor than haplotypes sharing the same copy number. Each of these scenarios would suggest that different mutational processes lead to a change in copy number at the *DEFA1A3* locus. In addition, the association of flanking sequence variants with copy number is rare at multiallelic CNV loci, due to the limited ability of a biallelic SNP in tagging multiple copy number states [29]. However, the SNP rs4300027 has previously been identified as a tag of *DEFA1A3* haplotype copy number, distinguishing 2-3 copy haplotypes from those with 4-5 copies [50]. This suggests there may be other variants that are able to tag additional features of the *DEFA1A3* locus, such as *DEFA3* content. The aim was to identify sequence variants flanking the *DEFA1A3* locus, in order to allow the identification of haplotypes that were identical by descent. This would then allow investigations of whether there are tags of additional features of the *DEFA1A3* locus, as

well as the identification of the mutational processes occurring at the *DEFA1A3* locus. As shown in chapter 3, the *DEFA1A3* copy number is mutating at a faster rate than the surrounding sequence; therefore, flanking sequence variation should tag *DEFA1A3* haplotypes sharing a common ancestor.

4.1 *DEFA1A3* flanking sequence variants

In order to identify sequence variants flanking the *DEFA1A3* locus, a 4.1kb region centromeric to the *DEFA1A3* locus (chr8: 6864188-6868287) was sequenced across 30 HapMap CEU1 trios; this allowed phased sequence haplotypes to be obtained for 120 independent haplotypes. The region was selected as it is sufficiently diverged from the *DEFA1A3* full repeats to allow specific amplification from a single locus per haplotype, whilst being close enough to the *DEFA1A3* locus to be associated and act as a tag of it, given that it occurs within the region of high LD including the *DEFA1A3* locus. In addition, the region contains the SNP rs4300027, which has previously been associated with *DEFA1A3* copy number in the European population; hence, it is possible that SNPs further sub-categorising the *DEFA1A3* haplotypes may be contained within this region.

Across the 4.1kb region sequenced, 48 different sequence variants were identified; 46 single base substitutions and 2 Indels (table 4.1). The centromeric-most site for the Indel5 variant is located within the region sequenced; however, within the 120 haplotypes sequenced, only the deleted form was observed in this location. The ancestral and derived alleles for the observed variants were assigned based on the ancestral sta-

rs number	Coordinate chr8	Ancestral allele	Derived allele	Variant status
-	6864195	G	A	PSV
-	6864197	C	T	PSV
-	6864198	G	A	Exchange
-	6864200	G	T	PSV
-	6864222	G	A	PSV
rs71511142	6864337	C	T	PSV
-	6864338	A	G	Exchange
-	6864348	C	G	PSV
-	6864363	T	C	PSV
-	6864365	G	C	PSV
-	6864370	G	A	PSV
-	6864379	C	G	Exchange
rs62487509	6864445	C	T	PSV
rs62487510	6864484	A	G	PSV
-	6864507	T	A	PSV
-	6864652	G	A	PSV
-	6864726	G	A	PSV
-	6864773	T	C	Exchange
-	6864790	A	G	PSV
rs11137085	6864878	C	G	PSV
rs2739218	6864926	C	T	Exchange
rs59380237	6864943	A	G	PSV
rs2615770	6864962	G	A	PSV
rs4012963	6865114	G	C	Exchange
-	6865254	G	A	PSV
-	6865276	G	A	PSV
rs4841815	6865288	T	C	PSV
-	6865330	G	A	PSV
rs10112401	6865500	A	G	PSV
rs4284061	6865667	A	T	PSV
-	6865783	G	A	PSV
rs34894091	6866188	G	A	Exchange
rs17466573	6866247	G	A	PSV
rs71509228	6866454	T	G	Exchange
-	6866462	C	G	PSV
-	6866545	C	T	PSV
-	6866729	- -	CA	PSV; Indel
-	6866748	C	T	PSV
-	6866846	G	T	PSV
rs7825750	6867213	T	C	PSV
rs62487513	6867289	C	G	PSV
-	6867294	-	T	PSV; Indel
rs7814783	6867295	C	T	PSV
rs62487514	6867504	C	A	PSV
-	6867854	C	A	PSV
rs17382102	6867931	A	G	PSV
rs4300027	6867985	T	C	PSV
rs4512398	6868004	T	C	PSV

Table 4.1: Variants identified in the *DEFA1A3* flanking region (chr8: 6864188-6868287). All variants SNVs unless stated. PSV= partial repeat-specific variant; Exchange= replacement of partial repeat sequence with a contiguous block of sequence from equivalent region of full repeat.

tus of the partial repeat, defined using Chimpanzee, Gorilla and Orangutan sequence [93]. Previous work at the *DEFA1A3* locus identified an exchange of genetic material in which a 2.2kb region of the *DEFA1A3* telomeric partial repeat (chr8: 6813274-6815492) was replaced with the sequence from the equivalent region of the *DEFA1A3* full repeats in a gene conversion event ([38], Fayeza Khan, personal communication). Therefore, it is possible for similar events to have occurred within the *DEFA1A3* centromeric partial repeat. The first 3kb of the 4.1kb region sequenced (chr8: 66864188-6867141) shares high sequence identity to the *DEFA1A3* full repeats and could facilitate a sequence replacement event. However, the final 1kb of sequence (chr8: 6867142-6868287) differs considerably from the *DEFA1A3* full repeats; therefore, variants within this region were assumed to be partial repeat-specific variants (PSVs). The Human, Chimpanzee and Gorilla sequences were obtained for the full repeat sequence [93] and used to assess whether similar sequence replacement events had occurred at the centromeric end of the *DEFA1A3* locus. Ideally, an exchange event could be identified when the ancestral status of the full and partial repeats differed, with the full repeat variant appearing as a derived allele in the partial repeat location. However, there are no examples of this.

There are eight variants which potentially display evidence of a sequence exchange. In all cases, the ancestral status of the full and partial repeats is the same, with the derived full repeat base appearing in the partial repeat location, suggesting either a possible exchange event, in which the direction is unclear, or a recurrent mutation. However, when multiple exchange variants are observed across the same haplotype, it increases the likelihood of a sequence exchange having occurred across that region. For example, there are six examples of a haplotype which con-

tains seven consecutive exchange positions, likely representing a single sequence replacement event. Therefore, the *DEFA1A3* flanking region is highly variable, containing possible sequence exchanges with the *DEFA1A3* full repeats.

4.2 *DEFA1A3* haplotype classes

The advantage of sequencing individuals who are part of two-generation trios is that it allowed the 48 sequence variants to be phased across the 4.1kb region sequenced, providing complete haplotype sequences. In total, 21 different sequences were observed across the region- i.e. combinations of variants found together on the same haplotype (table 4.2). 20 of the 21 sequences can be grouped into five classes, based on sequence similarity in which each class is distinct at several variant positions compared to all other classes. Haplotype ”RefSeqA” matches the human reference assembly sequence and haplotypes RefSeqB-G all differ at only a single variant position from haplotype RefSeqA. These are therefore grouped into the ”Reference Sequence” class. Similarly, haplotypes Exchange 1A-D can be grouped as ”Exchange 1”, haplotypes Class 2A-E as ”Class 2” and haplotypes Class 1 A-C as ”Class 1”. Haplotype ”Exchange 2” represents the fifth class, with all six examples sharing the same sequence. The remaining haplotype, ”Class 3”, was only observed on a single occasion and is distinct from all other classes in terms of its sequence.

29 of the 48 variants are found only within a single class, whilst there are 19 variants shared between different combinations of classes. There are five variants unique to the Reference Sequence class, which distinguish

Table 4.2: The sequence of each of the 21 observed *DEFA1A3* haplotypes for each variant position identified across the region chr8: 6864188-6868287. Derived alleles are shaded in grey.

Haplotype	Freq.	6864195	6864197	6864198	6864200	6864222	6864337	6864338	6864348	6864363	6864365	6864370	6864379	6864445	6864484	6864507	6864652	6864726	6864773	6864790	6864878	6864926	6864943	6864962	6865114
RefSeq A	28	A	T	G	T	A	C	A	C	T	G	A	C	T	A	T	G	G	T	G	G	T	A	G	G
RefSeq B	6	A	T	G	T	A	C	A	C	T	C	A	C	T	A	T	G	G	T	G	G	T	A	G	G
RefSeq C	7	A	T	G	T	A	C	A	C	T	G	A	C	T	A	A	G	G	T	G	G	T	A	G	G
RefSeq D	1	A	T	G	T	A	C	A	C	T	G	A	C	T	A	T	A	G	T	G	G	T	A	G	G
RefSeq E	1	A	T	G	T	A	C	A	C	T	G	A	C	T	A	T	G	G	T	G	G	T	A	A	G
RefSeq F	1	A	T	G	T	A	C	A	C	T	G	A	C	T	A	T	G	G	T	G	G	T	A	G	G
RefSeq G	1	A	T	G	T	A	C	A	C	T	G	A	C	T	A	T	G	G	T	G	G	T	A	G	G
Class 1 A	4	A	T	G	T	A	C	A	G	T	G	A	C	T	A	T	G	G	T	G	G	T	G	G	G
Class 1 B	9	A	T	G	T	A	C	A	G	T	G	A	C	T	A	T	G	G	T	G	G	T	G	G	G
Class 1 C	1	A	T	G	T	A	C	A	G	T	G	A	C	T	A	T	G	G	T	G	G	T	G	G	G
Class 2 A	1	A	T	G	T	A	C	A	C	T	G	A	C	T	A	T	G	G	T	G	G	C	A	A	G
Class 2 B	24	A	T	G	T	A	C	A	C	T	G	A	C	T	A	T	G	G	T	G	C	C	A	A	G
Class 2 C	2	A	T	G	T	A	C	A	C	T	G	A	C	T	A	T	G	A	T	G	C	C	A	A	G
Class 2 D	1	A	T	G	T	A	C	A	C	T	G	A	C	T	A	T	G	G	T	G	C	C	A	A	G
Class 2 E	1	A	T	G	T	A	C	A	C	T	G	A	C	T	A	T	G	G	T	G	C	C	A	A	G
Exchange 1 A	18	A	C	A	G	G	T	G	C	T	G	A	C	C	G	T	G	G	T	G	G	T	A	G	G
Exchange 1 B	1	A	C	A	G	G	T	G	C	T	G	A	C	C	G	T	G	A	T	G	G	T	A	G	G
Exchange 1 C	1	A	C	A	G	G	T	G	C	T	G	A	C	C	G	T	G	G	T	G	G	T	A	G	C
Exchange 1 D	5	A	C	A	G	G	T	G	C	T	G	A	C	C	G	T	G	G	T	G	G	T	A	G	G
Exchange 2	6	G	C	A	G	G	C	G	C	C	G	G	G	C	A	T	G	G	C	A	C	T	A	A	C
Class 3	1	G	C	A	G	G	C	A	C	T	G	A	C	T	A	T	G	G	T	G	G	T	A	G	G

Haplotype	Freq.	6865254	6865276	6865288	6865330	6865500	6865667	6865783	6866188	6866247	6866454	6866462	6866545	6866729	6866748	6866846	6867213	6867289	6867294	6867295	6867504	6867854	6867931	6867985	6868004
RefSeq A	28	G	A	T	G	A	T	G	G	G	T	G	C	- -	C	G	T	C	-	C	C	C	A	T	C
RefSeq B	6	G	A	T	G	A	T	G	G	G	T	G	C	- -	C	G	T	C	-	C	C	C	A	T	C
RefSeq C	7	G	A	T	G	A	T	G	G	G	T	G	C	- -	C	G	T	C	-	C	C	C	A	T	C
RefSeq D	1	G	A	T	G	A	T	G	G	G	T	G	C	- -	C	G	T	C	-	C	C	C	A	T	C
RefSeq E	1	G	A	T	G	A	T	G	G	G	T	G	C	- -	C	G	T	C	-	C	C	C	A	T	C
RefSeq F	1	G	A	T	G	A	T	A	G	G	T	G	C	- -	C	G	T	C	-	C	C	C	A	T	C
RefSeq G	1	G	A	T	G	A	T	G	G	G	T	G	C	CA	C	G	T	C	-	C	C	C	A	T	C
Class 1 A	4	G	A	T	G	A	T	G	G	G	G	C	C	- -	C	G	T	C	-	C	C	C	A	C	T
Class 1 B	9	G	A	T	A	A	T	G	G	G	G	C	C	- -	C	G	T	C	-	C	C	C	A	C	T
Class 1 C	1	G	A	T	G	A	T	G	G	G	G	C	C	- -	C	T	T	C	-	C	C	C	A	C	T
Class 2 A	1	G	G	T	G	A	A	G	G	G	T	G	C	- -	C	G	C	C	-	T	C	C	A	C	T
Class 2 B	24	G	G	T	G	A	A	G	G	G	T	G	C	- -	C	G	C	C	-	T	C	C	A	C	T
Class 2 C	2	G	G	T	G	A	A	G	G	G	T	G	C	- -	C	G	C	C	-	T	C	C	A	C	T
Class 2 D	1	G	G	T	G	A	A	G	G	G	T	G	C	- -	C	G	C	C	-	T	C	A	A	C	T
Class 2 E	1	G	G	T	G	A	A	G	G	G	T	G	C	- -	C	G	C	C	T	T	C	C	A	C	T
Exchange 1 A	18	G	A	T	G	A	T	G	G	A	T	G	C	- -	C	G	T	G	-	C	A	C	G	T	C
Exchange 1 B	1	G	A	T	G	A	T	G	G	A	T	G	C	- -	C	G	T	G	-	C	A	C	G	T	C
Exchange 1 C	1	G	A	T	G	A	T	G	G	A	T	G	C	- -	C	G	T	G	-	C	A	C	G	T	C
Exchange 1 D	5	A	A	T	G	A	T	G	G	A	T	G	C	- -	C	G	T	G	-	C	A	C	G	T	C
Exchange 2	6	G	A	C	G	G	A	G	A	G	T	G	T	- -	C	G	T	C	-	C	C	C	A	C	T
Class 3	1	G	A	T	G	A	T	G	G	G	T	G	C	- -	T	G	T	C	-	C	C	C	A	T	C

RefSeq B, C, D, F and G from RefSeq A, whilst the variant distinguishing RefSeq E from RefSeq A is shared with two other classes. Class 1 haplotypes contain three unique variants, one of which may represent a sequence exchange, although given that it is a single variant position, this is unclear. There are two additional variants distinguishing Class 1 B and C from Class 1 A. Class 2 haplotypes have two variants unique to the class, as well as four positions distinguishing Class 2 A-E, three of which are shared with other classes. Exchange 2 haplotypes contain seven unique positions, whilst Exchange 1 haplotypes contain six unique positions and three distinguishing Exchange 1A-D, two of which are shared with other classes. The Class 3 haplotype has one unique position and a potential exchange event at position chr8: 6864198, although given it is a single position, it is unclear whether it is a true exchange event. The SNP rs4300027, which has previously been associated with copy number, clearly separates Class 1, Class 2 and Exchange 2 haplotypes from the Reference Sequence, Exchange 1 and Class 3 haplotypes, although this does not necessarily suggest that all rs4300027 C haplotypes are more closely related to each other than they are to rs4300027 T haplotypes, and vice versa.

A 2.2kb region of the *DEFA1A3* telomeric partial repeat has been shown to be exchanged with sequence from the equivalent region of the *DEFA1A3* full repeats, in a gene conversion event referred to as the ”Telomeric replacement polymorphism” ([38], Fayeza Khan, personal communication). Sequencing data from the centromeric partial repeats suggests that there are two haplotype classes which display evidence of a full repeat sequence replacement event- Exchange 1 and Exchange 2 (figure 4.1).

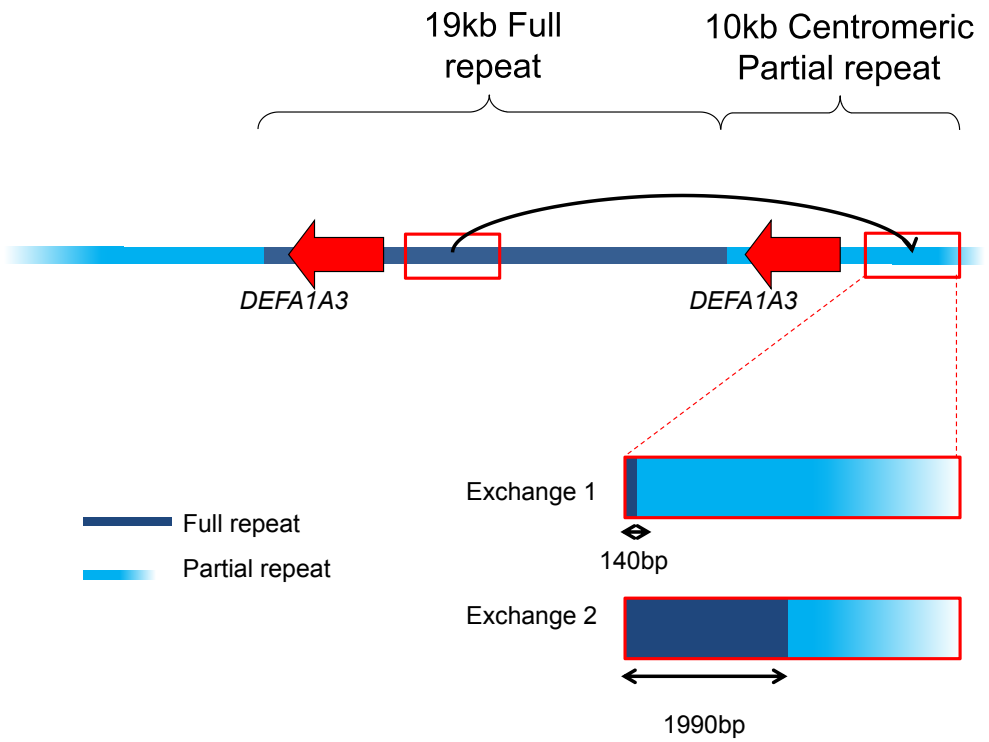


Figure 4.1: The *DEFA1A3* centromeric partial repeat shares a high sequence similarity to the full repeats, such that the 4.1kb region of the centromeric partial repeat that has been sequenced has an equivalent location in the full repeats (red boxes). There are sequence differences between the two locations and it appears there is evidence of the partial repeat sequence being replaced by that from the full repeats (arrow). This has occurred over a 140bp region on Exchange 1 haplotypes and a 1990bp region on Exchange 2 haplotypes.

Exchange 1 haplotypes contain two consecutive variants, at positions chr8:6864198 and chr8:6864338, which represent a possible exchange event (i.e. there are no other variants in between these positions which are considered exchanges). These variants stretch over a 140bp region, with a centromeric boundary between chr8: 6876928-6876969. The telomeric boundary extends beyond chr8: 6876788, although it may be continuous with the block of full repeats. Exchange 2 haplotypes contain seven consecutive exchange positions over a region of 1990bp, with a centromeric boundary between chr8: 6878778-6879044. The telomeric

boundary extends beyond chr8: 6876788, although it may be continuous with the block of full repeats. Therefore, Exchange 2 is a convincing example of a full repeat sequence replacement event.

DEFA1A3 haplotype class lineage

Defining *DEFA1A3* haplotype classes has allowed the identification of haplotypes sharing a common ancestor; however, it is not known how the classes are related to each other. The reconstruction of an evolutionary tree to determine the relationship of the *DEFA1A3* haplotype classes could be done using the sequence information from the centromeric flanking region (chr8: 6864188-6868287), but the observation of both sequence replacement and recombination events across this region makes the task more complex. Therefore, two independent phylogenetic trees were constructed using different parts of the flanking region. The first tree was constructed using the final 1kb of sequence (chr8: 6867142-6868264), which appears to be exchange and recombination-free, across all haplotype classes, whilst the second tree used the first 3kb of sequence (chr8: 6864195-6867141) for the non-exchange haplotypes (Reference Sequence, Class 1, Class 2 and Class 3). Both trees are unrooted, due to the lack of a reliable outgroup sequence. Whilst the reference assemblies for the Chimpanzee and Orang-utan provide complete sequence for this region, it is not clear if it is representative of the true partial sequence in these species, due to the poor assembly of the *DEFA1A3* locus, and whether these would provide an accurate representation of the ancestral status of the region in humans, which could have undergone many changes since the divergence of Chimpanzee and Human.

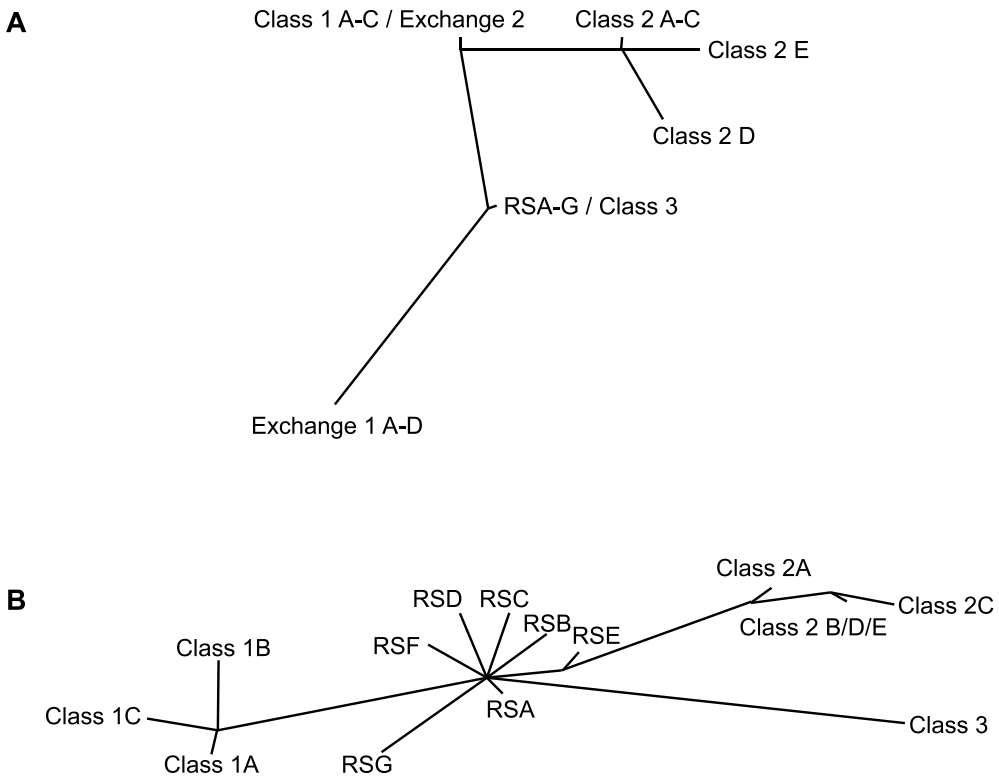


Figure 4.2: Phylogenetic trees showing the evolutionary relationship between each haplotype class at *DEFA1A3* . A: Tree constructed using final 1kb of sequence for all haplotypes. B: Tree constructed using first 3kb of sequence for non-Exchange haplotypes. RS= Reference Sequence

The two phylogenetic trees are shown in figure 4.2. As expected, the haplotypes within each major class cluster together in both trees. Tree A places Class 1 and Exchange 2 close to both Class 2 and the Reference Sequence, whereas in Tree B, the Reference Sequence is more closely related to both Class 1 and Class 2 than they are to each other. This is due to the shared status at both rs4300027 and rs4512398 of Class 1 and Class 2 haplotypes, which differs from the Reference Sequence, and the absence of Class 1-specific variants in the final 1kb of sequence, making them appear more closely related in Tree A than in Tree B. Tree A suggests that the Exchange 1 class is evolutionary more distant from all other classes, with the Reference Sequence its closest

class. Tree A is unable to distinguish Class 3 from the Reference Sequence and Tree B places Class 3 as being most closely related to the Reference Sequence, although diverged at many positions, as shown by the long branch length. Given that these two classes also share the same rs4300027 status, Class 3 will be grouped with the Reference Sequence for further analysis.

As well as differences in the evolutionary distance between Class 1 and Class 2, tree B suggests that RefSeq E is more closely related to Class 2 than the other Reference Sequence haplotypes, due to a shared sequence variant between them, which may have occurred due to a possible recombination event. In addition, it implies that RefSeq G is more distant from RefSeq A-F haplotypes than they are to each other, due to the inability of the algorithm to deal with the 2bp insertion observed in RefSeq G, which appears as two independent events. Overall, both trees are representative of the sequence used to generate them, but suggest different phylogenies between the *DEFA1A3* haplotype classes.

Worldwide *DEFA1A3* haplotype class frequency

The five *DEFA1A3* haplotype classes were identified in the 60 HapMap CEU1 European individuals, representing only 120 haplotypes. Hence, it was unclear whether the observed frequencies of each class were representative of the wider European or worldwide populations.

In order to investigate this in the European population, haplotype classes were assigned to an additional 539 individuals with European ancestry (HapMap CEU2 and ECACC HRC 1-5), using a combination of four SNP genotypes which act as tags of the five haplotype classes observed in

the CEU1 population (figure 4.3). Three of the four SNPs used were identified from sequencing. However, the SNP rs7826487, used as a tag of the Class 1 lineage, was identified in the HapMap dataset [116], as the high sequence similarity between the *DEFA1A3* full and partial repeats prevented the development of an assay to genotype a Class 1 tag within the sequenced region. For the 539 samples typed, 538 conformed to the expected pattern of LD; the remaining sample was excluded from further analysis. The genotype frequencies observed for each of the four SNPs conformed to Hardy-Weinberg equilibrium (rs7826487: p=0.65; rs7825750: p=0.70; rs62487514: p=0.04; rs4300027: p=0.63).

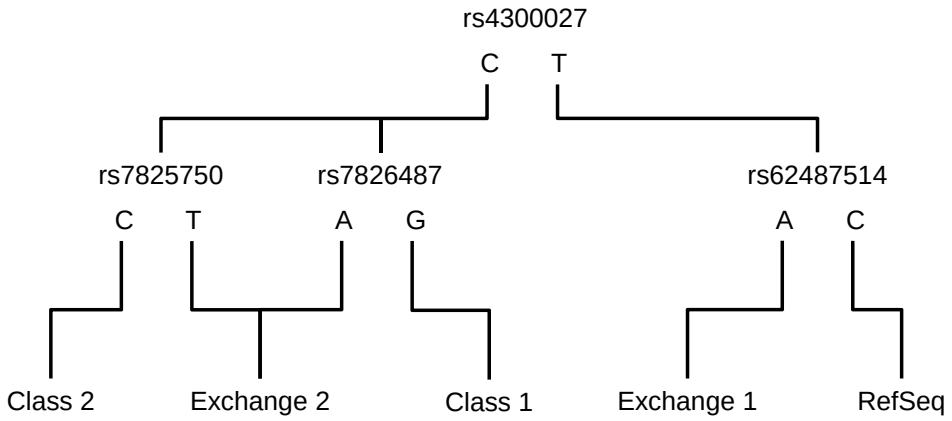


Figure 4.3: The four SNPs used to tag the five haplotype classes at *DEFA1A3* exhibit LD, such that a diploid genotype profile can be used to indicate the classes of the haplotypes for the sample.

The frequencies of the five *DEFA1A3* haplotype classes in the CEU1 samples were compared to the larger European dataset (figure 4.4). There is a significant difference in the class frequency distribution between the two populations (χ^2= 12; p= 0.016), which appears to be due to a higher frequency of Class 2 and Exchange 2 haplotypes and a lower frequency of Exchange 1 haplotypes in the CEU2 and HRC 1-5 samples, compared to CEU1. However, given that the CEU1 population repre-

sents a small sample of only 120 haplotypes and, as such, is likely to be distorted due to sampling effect, the difference in class frequencies is minimal.

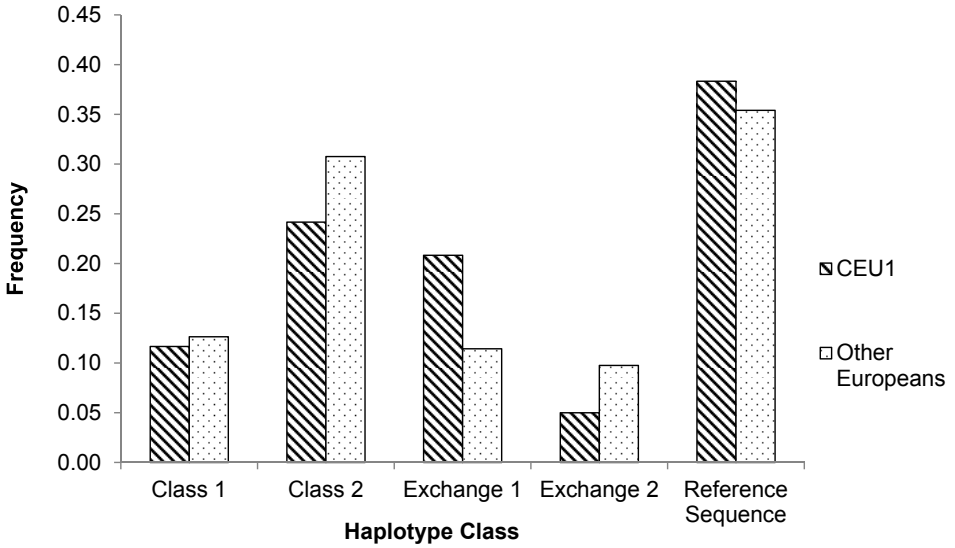


Figure 4.4: The frequencies of each of the five *DEFA1A3* haplotype classes in the CEU1 population compared to an additional 538 individuals of European ancestry (CEU2 and HRC 1-5.)

The same four SNPs were genotyped as part of the 1000 Genomes project [20], providing haplotype class status for 2184 haplotypes from 14 worldwide populations, grouped as African, European, Asian and American. This confirms that the expected pattern of LD between the four SNPs is conserved worldwide, with just two exceptions, both of which are in the Asian population. This data also shows that the SNPs rs4300027 and rs4512398 are not in complete LD, as previously assumed, and there are exceptions in all worldwide populations. Some of the HapMap CEU samples are included within the 1000 Genomes dataset and, in all cases, the SNP genotypes agree.

Figure 4.5 shows the frequencies of the five *DEFA1A3* haplotype classes in different worldwide populations. The most notable difference is the

near absence of Exchange 1 from the Asian and African populations, with only two examples observed in the African population and no examples in the Asian population. This observation has been confirmed experimentally for the HapMap CHB, JPT and YRI samples, in which the Exchange 1 sequence replacement typing assay was negative for all samples tested (n=150)(section 2.6).

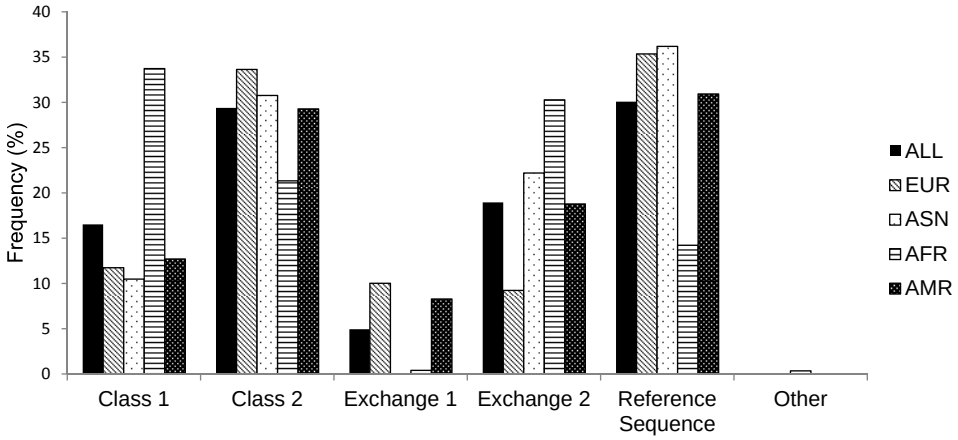


Figure 4.5: Frequencies of the five *DEFA1A3* haplotype classes across the European (EUR), Asian (ASN), African (AFR) and American (AMR) populations as observed in the 1000 Genomes samples [20].

Whilst four of the five *DEFA1A3* haplotype classes are found outside of Europe, the frequencies of each vary greatly between populations (figure 4.5); χ^2= 362; p= 6.08 x 10^{-68}. Whilst the Reference Sequence class is the most frequent in the European, Asian and American populations, with the Class 2 haplotype the second-most frequent, the Reference Sequence haplotype is much less common in African populations, where the Exchange 2 and Class 1 haplotypes are more prominent. In addition, the European population has an unusually low frequency of the Exchange 2 class, compared to other worldwide populations.

Data from the 1000 Genomes project confirms the LD observed across the *DEFA1A3* locus in HapMap CEU1 individuals (figure 4.6). This sug-

gests that, although haplotype classes have been defined on the basis of sequence variants identified in the centromeric flanking region of the *DEFA1A3* locus, they are tags of haplotypes across the entire *DEFA1A3* region.

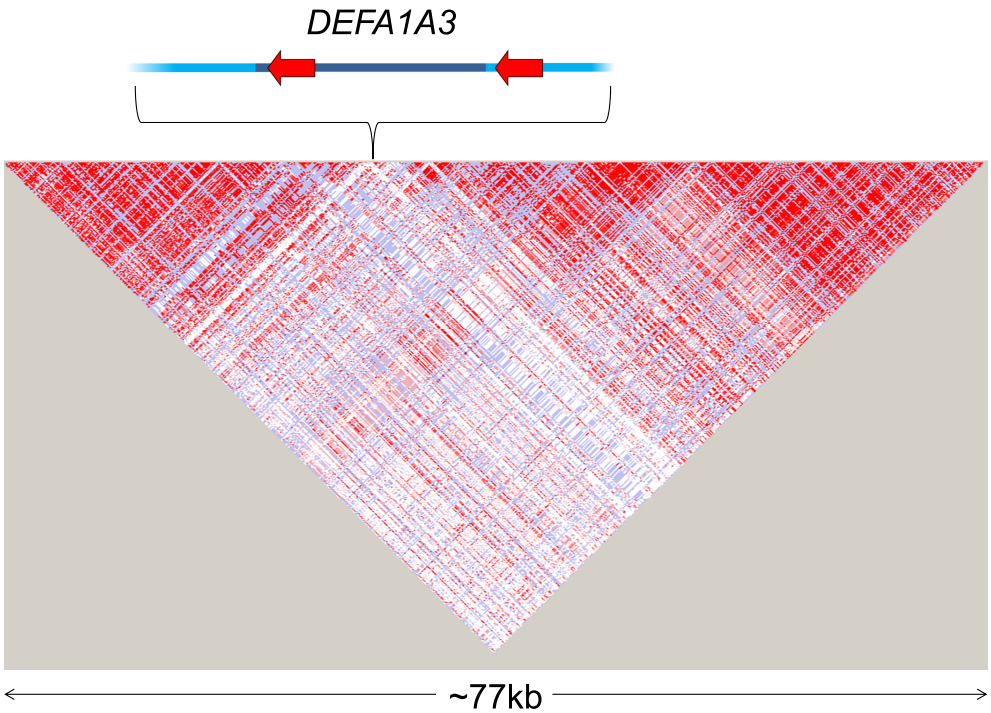


Figure 4.6: Data for the 1000 Genomes European samples confirms that the *DEFA1A3* locus falls within a region of high LD, as SNPs either side of the locus have a D′=1 [20]

.

4.3 Evidence of selection at *DEFA1A3*

In order to look for evidence of selection at *DEFA1A3*, the EHH was assessed for the ancestral and derived alleles of SNPs tagging the five *DEFA1A3* haplotype classes and the SNP rs4300027. EHH can act as an indicator of selection across a genomic region, as selective sweeps resulting from selection at a particular locus will lead to extended regions of homozygosity on haplotypes carrying the advantageous allele [143].

Therefore, if a particular *DEFA1A3* haplotype class has been under selection, the allele tagging that class (i.e. the derived allele) will display evidence of a selective sweep.

The EHH plots obtained for the five *DEFA1A3* haplotype classes and the SNP rs4300027 using phased haplotype data for the HapMap CEU1 individuals are shown in figure 4.7. These show that, for rs4300027 and the Reference Sequence, there is little difference in the EHH plots for the derived and ancestral alleles, suggesting no evidence for selection on backgrounds carrying the derived alleles for either of these variants. However, for the four other SNPs, there is a difference in the curves for the derived and ancestral alleles, suggesting selection is occurring on the derived allele background. However, in order to assess the significance of these differences, a standardised integrated haplotype score (iHS) was generated, as described in section 2.8 (table 4.3). This is necessary, as variants with a low minor allele frequency are expected to be younger and, as such, will be accompanied by increased EHH, as there will have been less time for new mutations to arise on that background. Relative to the genome-wide distribution of iHS values, the iHS for Exchange 1 is weakly significant (p=0.027), indicating evidence of selection on the Exchange 1 class background.

However, the HapMap CEU1 population represents only 120 independent haplotypes. Therefore, the same analysis was repeated for the 1000 Genomes European samples, excluding CEU1 individuals, to see if the same signature of selection on Exchange 1 haplotypes could be observed (figure 4.8 and table 4.3). The iHS obtained using data for the 1000 Genomes haplotypes does not reach genome-wide significance, suggesting the Exchange 1 haplotype class is not under selection.

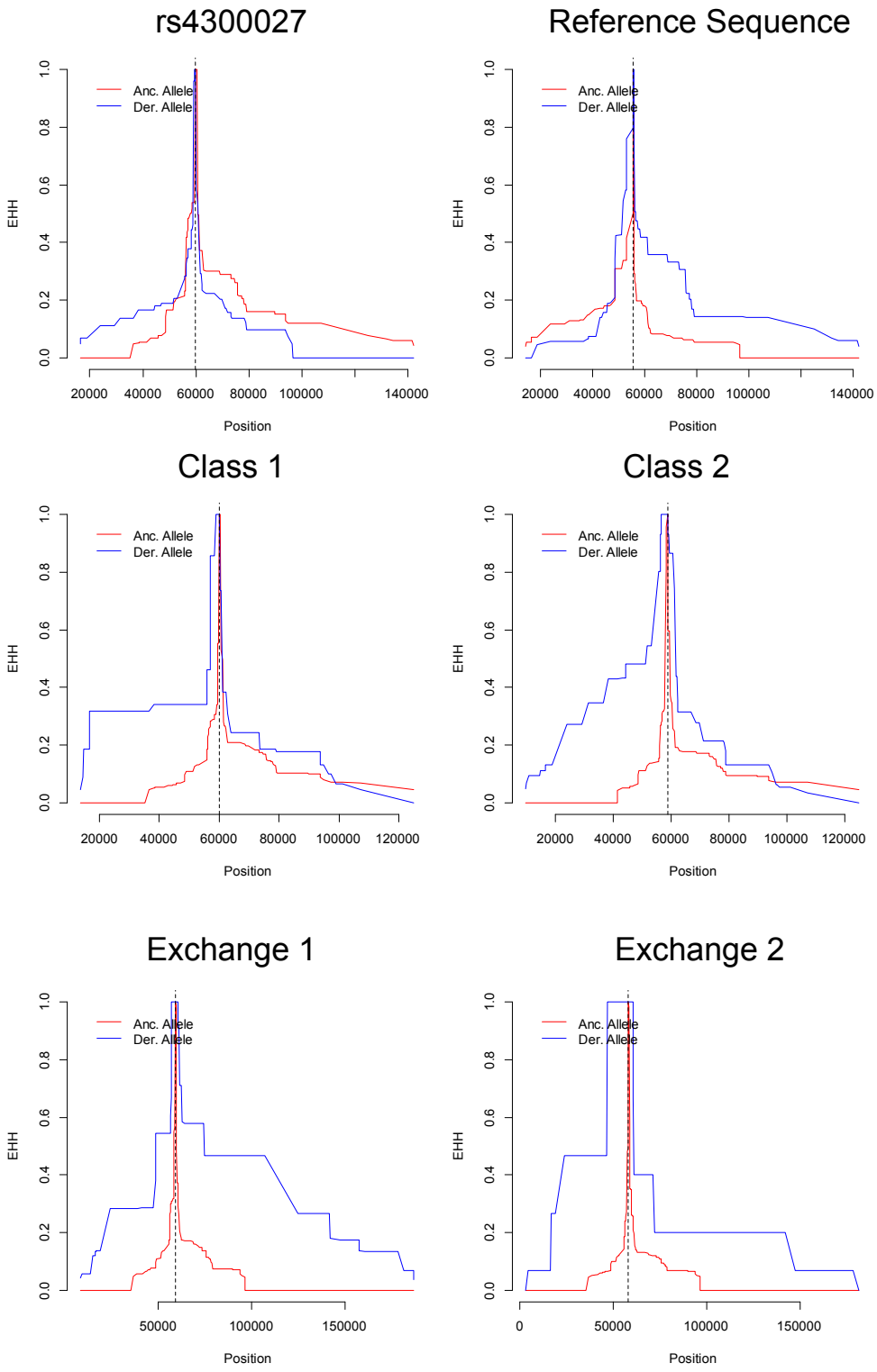


Figure 4.7: EHH plots for SNPs tagging the five *DEFA1A3* haplotype classes and the SNP rs4300027 for haplotypes in the HapMap CEU1 population [18]. EHH measured from SNP of interest to point where EHH <0.05.

Population/Variant	MAF	iHHa	iHHd	ln(iHH)	mean iHH	SD iHH	iHS	Significance
HapMap CEU1								
rs4300027	0.408	11689.839	9426.924	0.215	-0.499	0.789	0.905	NS
Reference Sequence	0.375	6297.867	15951.547	-0.929	-0.592	0.692	-0.488	NS
Class 1	0.117	6313.356	20636.625	-1.184	-1.143	0.619	-0.067	NS
Class 2	0.242	5730.922	22900.039	-1.385	-0.808	0.627	-0.921	NS
Exchange 1	0.208	4890.363	49384.86	-2.312	-0.873	0.652	-2.206	p=0.027
Exchange 2	0.05	3833.618	40399.271	-2.355	-1.361	0.585	-1.699	NS
1000 Genomes Europeans								
rs4300027	0.561	16637.619	9831.867	0.526	-0.516	0.669	1.556	NS
Reference Sequence	0.347	8903.699	25386.108	-1.048	-0.495	0.866	-0.638	NS
Class 1	0.121	7276.543	19595.377	-0.991	-1.016	0.589	0.043	NS
Class 2	0.346	7072.106	23972.457	-1.221	-0.549	0.699	-0.961	NS
Exchange 1	0.092	6001.556	51833.912	-2.156	-1.005	0.664	-1.733	NS
Exchange 2	0.094	7139.312	17027.167	-0.869	-1.005	0.664	0.205	NS
1000 Genomes Asians								
rs4300027	0.638	20261.39	12305.57	0.499	-0.212	0.781	0.910	NS
Reference Sequence	0.362	14976.61	20329.49	-0.306	-0.421	0.734	0.158	NS
Class 1	0.108	8160.662	43698.794	-1.678	-1.196	0.696	-0.693	NS
Class 2	0.311	8165.982	39566.691	-1.578	-0.679	0.623	-1.444	NS
Exchange 2	0.222	10696.88	22222.44	-0.731	-0.75	0.664	0.029	NS
1000 Genomes Africans								
rs4300027	0.854	8787.538	3023.547	1.067	0.722	0.887	0.388	NS
Reference Sequence	0.142	3697.431	8735.508	-0.86	-1.059	0.578	0.344	NS
Class 1	0.337	3492.153	13532.957	-1.355	-0.632	0.633	-1.142	NS
Class 2	0.213	3359.435	14895.914	-1.489	-0.707	0.541	-1.445	NS
Exchange 2	0.303	4423.179	5787.68	-0.269	-0.621	0.672	0.524	NS

Table 4.3: The iHS associated with each *DEFA1A3* haplotype class and the SNP rs4300027 for the HapMap CEU1 and 1000 Genomes European, Asian and African populations, calculated as described in section 2.8. The relative position of the iHS in the genome-wide distribution of iHS values was used to indicate significance at a genome-wide level. iHS values between -2 and 2 are labelled not significant (NS), as this range includes 95.5% of iHS values.

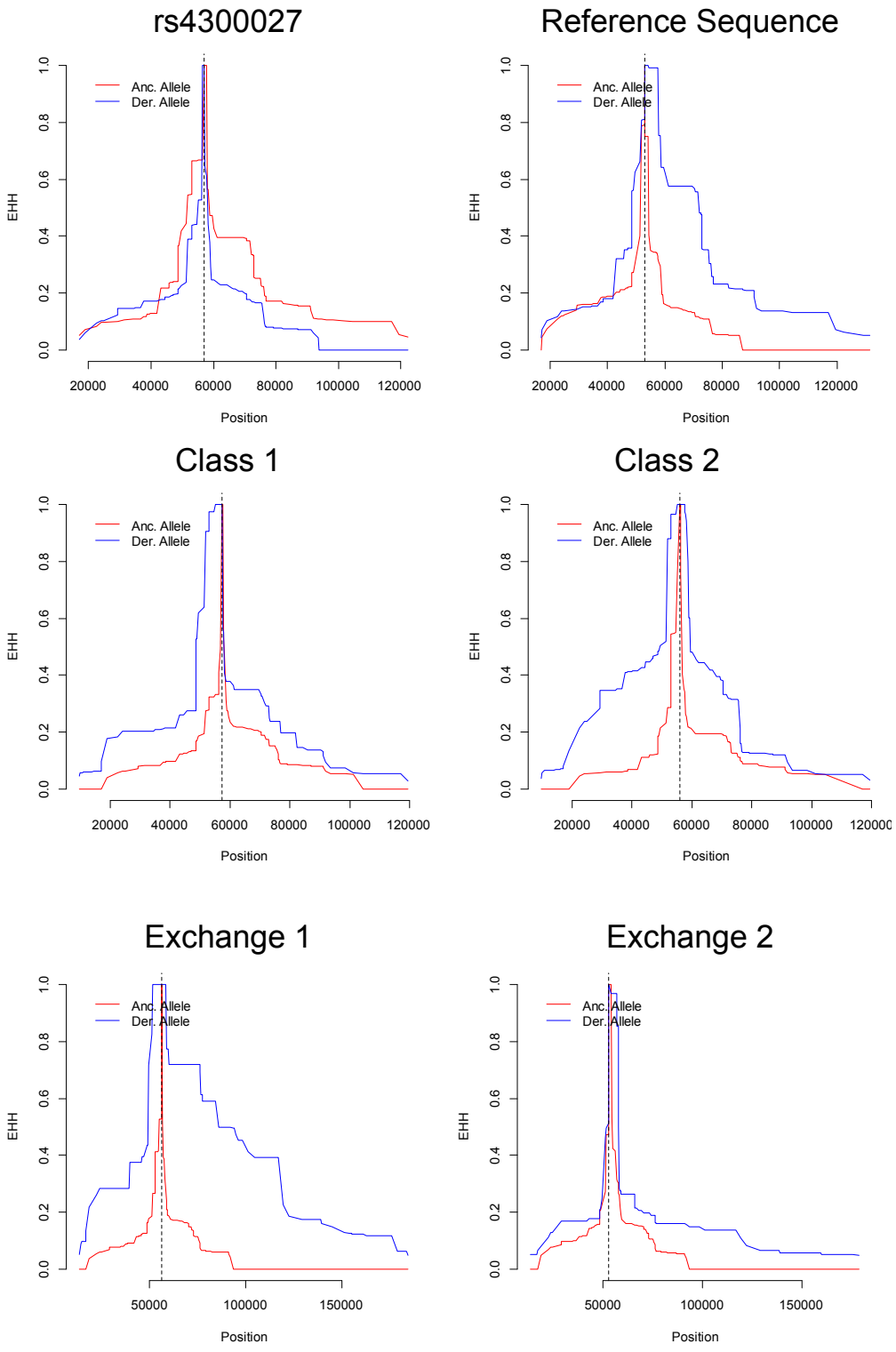


Figure 4.8: EHH plots for SNPs tagging the five *DEFA1A3* haplotype classes and the SNP rs4300027 for the 1000 Genomes European samples [20]. EHH measured from SNP of interest to point where EHH <0.05.

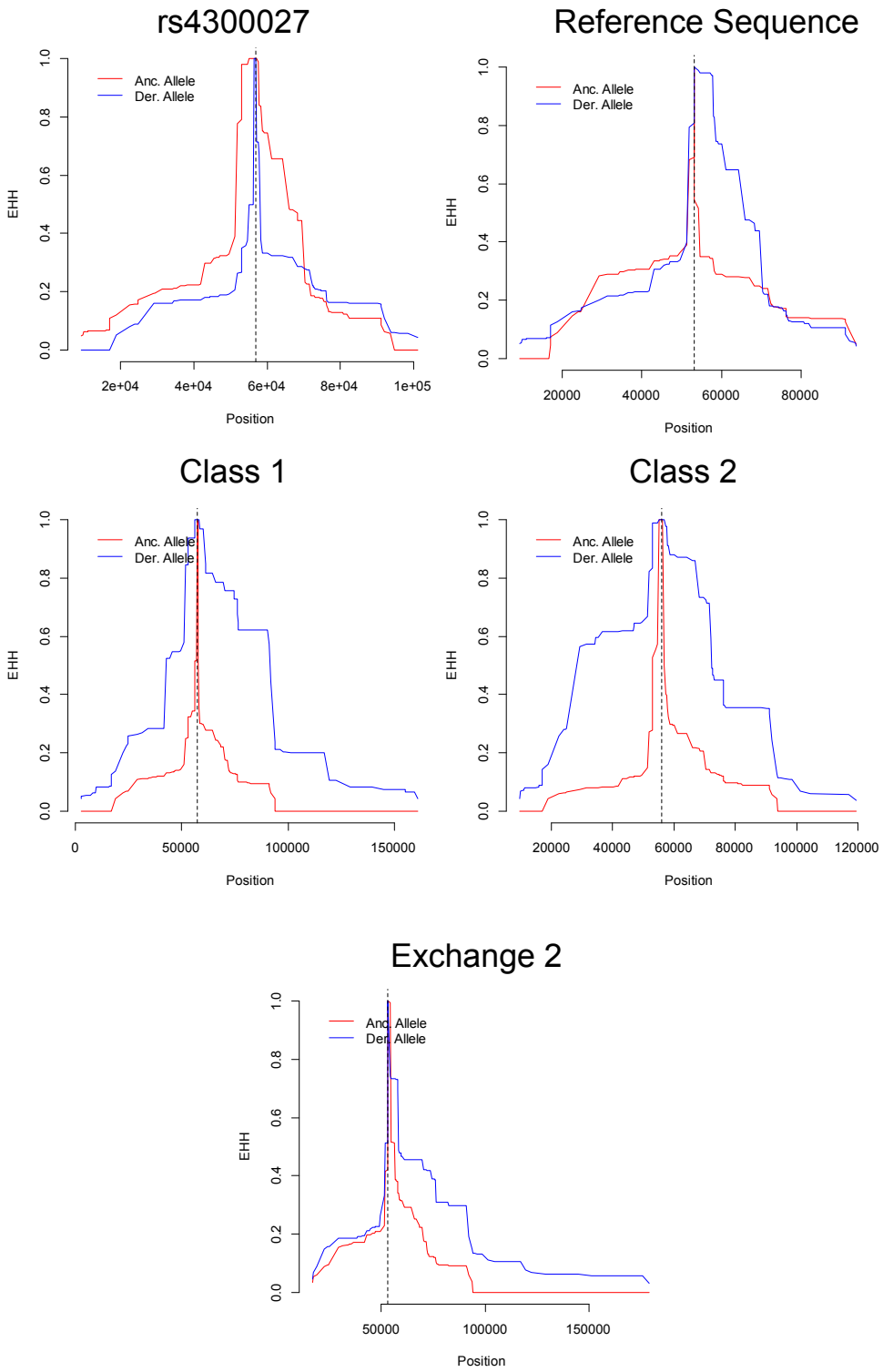


Figure 4.9: EHH plots for SNPs tagging the five *DEFA1A3* haplotype classes and the SNP rs4300027 for the 1000 Genomes Asian samples [20]. EHH measured from SNP of interest to point where EHH <0.05.

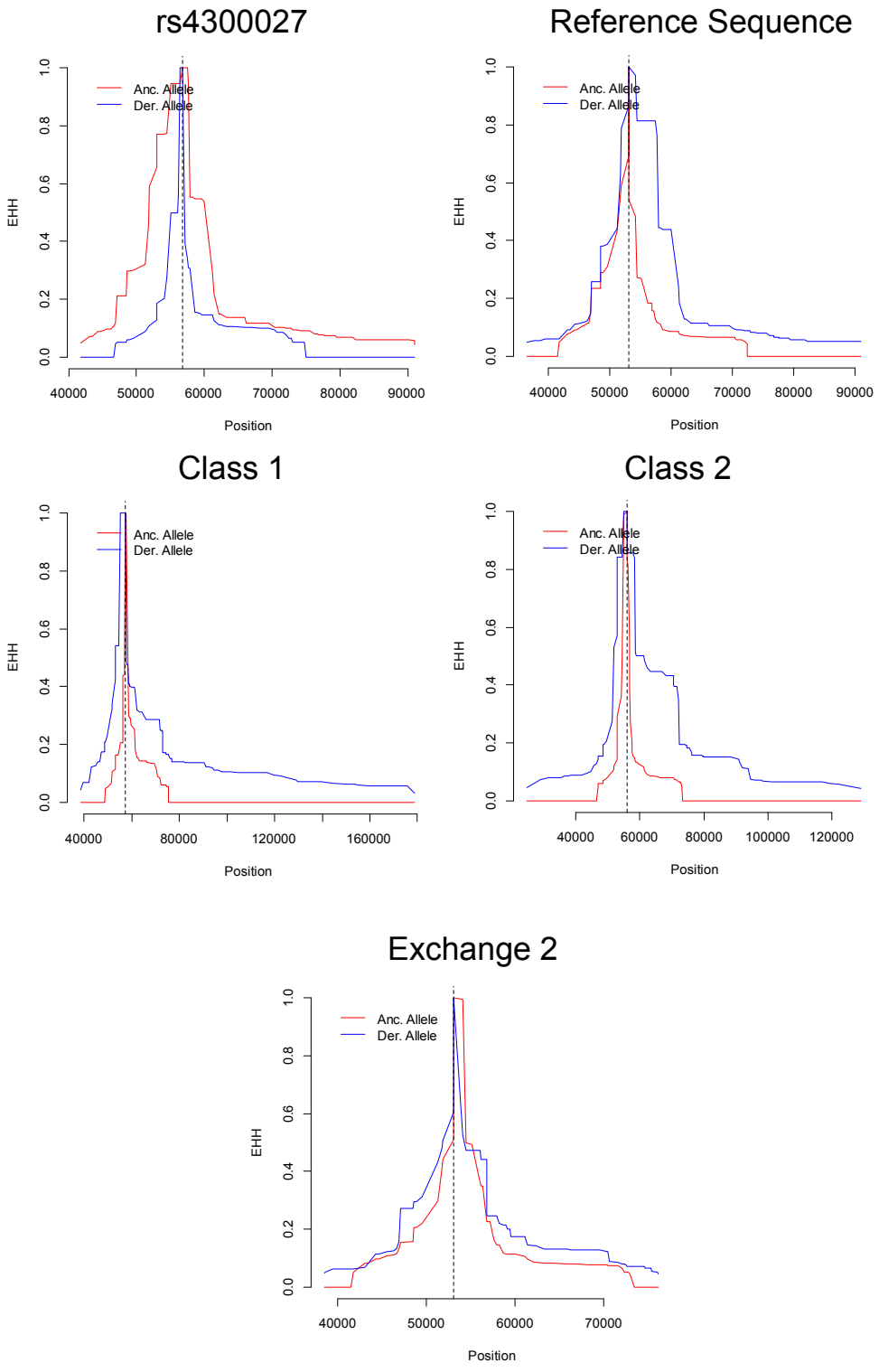


Figure 4.10: EHH plots for SNPs tagging the five *DEFA1A3* haplotype classes and the SNP rs4300027 for the 1000 Genomes African samples [20]. EHH measured from SNP of interest to point where EHH <0.05.

In order to compare these results to other populations, the iHS was calculated for four of the five *DEFA1A3* haplotype classes and the SNP rs4300027 for the 1000 Genomes Asian and African samples (figures 4.9 and 4.10 and table 4.3); Exchange 1 was excluded due to its low frequency/absence in these populations. In the Asian population, the EHH plots show differences between the derived and ancestral alleles for the Class 1 and Class 2 variants, but neither show genome-wide significance. In the African population, the EHH plots for the derived and ancestral alleles are similar for all five variants and this is reflected in the corresponding iHS, which were not significant. Therefore, in both the Asian and African ancestry populations, no *DEFA1A3* haplotype class displays evidence of selection.

4.4 Conclusions

Flanking sequence information provides further evidence that the *DEFA1A3* locus is highly variable. Within this sequence information, we see not only five clear haplotype classes at *DEFA1A3*, but evidence of possible localised gene conversion events between *DEFA1A3* full and partial repeat sequences. There are a number of limitations in trying to identify sequence exchange (i.e. gene conversion) events, not least in assigning the ancestral and derived states for each variant position in both the full and partial repeats. This is due to the dynamic nature of the locus, in which full and partial repeat sequences can be exchanged, as well as the poor reference assemblies for the non-human primate species. In addition, the method of assigning exchange events is conservative, as there are no sequence variants annotated within the *DEFA1A3* full repeats in the human reference assembly, due to the region's CNV

status. The full repeat sequence information is based on the two copies of the reference assembly and, if the allele observed matches the ancestral status, it will be assumed that the position is invariable in the full repeats. Therefore, if an unobserved derived allele is transferred, the exchange event will be missed. Further sequencing in which full repeat variants are captured would be required to clarify the status of possible exchange variants. In summary, there is evidence of both sequence exchanges and recombination across the region flanking the *DEFA1A3* locus, which is interesting given that it falls within a region of high LD, but small-scale, unidentified gene conversion events could be responsible for the apparent recombination. The high variability and potential for exchanges makes the *DEFA1A3* locus appear as more of a continuum, in which repeat boundaries are not fixed, but the sequence identity gradually declines at the boundaries of the locus, within the partial repeats.

The complex nature of the locus creates issues when trying to reconstruct the evolutionary history between the *DEFA1A3* haplotype classes. Due to apparent recombination and sequence exchange events, the flanking sequence, from which the phylogeny was inferred, had to be split into two sequences. This means neither tree is able to capture the full extent of flanking variation, but it would be unsuitable to include the entire sequence for all classes, as the tree drawing algorithms are unable to deal with complex events like these accurately, as shown by the inability to determine that the 2bp insertion on RefSeq G represents a single event. This does mean that information is lost; for example, Exchange 2 and Class 1 are grouped together in Tree A and Exchange 2 cannot be included in Tree B, due to the complex nature of its variation. As the trees consist of European variation alone, yet the same classes can be identified in all worldwide populations, it is likely that haplotypes rep-

resenting the intermediate steps between different classes are missing and inclusion of additional worldwide populations would help to clarify the relationship between each class.

Despite the complexity observed at the *DEFA1A3* locus, the four SNPs used to tag sequence class show a remarkable consistency of LD worldwide, with only three observed exceptions amongst the 3098 independent haplotypes observed. It is possible that other haplotypes which do not fit this rule have been missed, due to the assumptions about the LD between the genotyped SNPs imposed upon diploid genotype calls. However, this is likely to affect a very small proportion of the dataset. It is clear that the frequency of the five haplotype classes varies markedly worldwide, with a clear absence of the Exchange 1 haplotype from Asia and Africa. The two examples in the African population most likely represent an admixture event, suggesting the Exchange 1 sequence replacement is a recent event. In addition, it appears that the class frequencies observed in the CEU1 sample are close enough to the European distribution to provide a representative sample.

There is no evidence of selection at *DEFA1A3*, which would have suggested different functional properties of different haplotype classes at the locus. However, the locus must perform an important function, given its association with IgA Nephropathy in the Han Chinese population [110]. It is unsurprising that haplotypes do not display EHH, given that all classes, except Exchange 1, are found in all worldwide populations. It is possible that only a subset of haplotypes within a *DEFA1A3* class contain an allele under selection, or that multiple haplotypes contain advantageous variants, events which could not be detected using the iHS method. This shows that, even when the EHH plots for the ancestral and derived alle-

les appear to differ, it is the combination of the EHH values, the distance involved and the minor allele frequency of the variant that affect whether the difference is significant.

There are limitations with the methods used to assess EHH. As the *DEFA1A3* locus is CNV, there are no SNPs annotated across an approximately 47kb region. In order to avoid false positive results due to apparently high EHH across this region, the *DEFA1A3* CNVR was collapsed to account for this lack of variants. Inevitably, it does not accurately reflect the true and variable distance between the flanking markers. Also, Voight *et al.* [144] used recombination rate to generate genetic distances, in cMs, which was used instead of chromosome coordinate position to generate EHH plots. This will have affected the iHS scores obtained. However, it would be hard to assess accurately the recombination rate across the *DEFA1A3* locus, as it is CNV and displays evidence of many complex mutational processes. In summary, the iHS values obtained suggest no evidence of selection on each *DEFA1A3* haplotype classes.

5 Association of flanking sequence variation with features of the *DEFA1A3* locus

It is rare to find a SNP associated with copy number at a multiallelic CNV locus, due to the limited ability of a biallelic SNP to predict multiple different copy number states [29]. However, the SNP rs4300027 has previously been identified as a tag of *DEFA1A3* haplotype copy number in the European population, where the C allele is associated with 2 and 3-copy haplotypes and the T allele is associated with 4 and 5-copy haplotypes ($p=1.3x10^{-45}$) [50]. Despite the convincing association, this SNP alone demonstrates a poor ability to predict diploid *DEFA1A3* copy numbers, with only a weak correlation between the observed diploid copy number and that estimated using rs4300027 ($r^2=0.35$), as a biallelic SNP can only predict three different copy number states.

It is possible that other SNPs may be able to further partition the association of rs4300027 with copy number; for example, an additional SNP could distinguish 2-copy from 3-copy haplotypes. In addition, flanking sequence variants may tag other features of the *DEFA1A3* locus, such as *DEFA3* frequency. It is also possible that the use of multiple SNPs may improve on rs4300027 genotype alone in estimating diploid *DEFA1A3* copy number. The aim was to identify whether each *DEFA1A3* haplotype class, as described in chapter 4, is associated with particular fea-

tures of the *DEFA1A3* locus, for example, *DEFA1A3* copy number. In addition, the aim was to assess whether the use of multiple SNPs, which tag *DEFA1A3* haplotype class, may be able to estimate *DEFA1A3* copy number better than rs4300027 alone.

5.1 Comparing haplotype class with *DEFA1A3* features

As discussed in chapter 4, there are five different common classes of *DEFA1A3* haplotype: Reference Sequence, Class 1, Class 2, Exchange 1 and Exchange 2. As 29 of the 48 sequence variants identified in the region flanking the *DEFA1A3* locus (chr8:6864188-6868287) are found within a single class, it was decided that haplotype class would be compared to features of the *DEFA1A3* locus, rather than testing each SNP individually. The class genotype (e.g. Reference Sequence homozygous/ heterozygous/ negative) was compared to the *DEFA1A3* copy number or frequency of *DEFA3*, the Indel5 insertion and the 7bp duplication, using either a Chi Square or Cochran Armitage test (section 2.13). This analysis was initially performed in the HapMap CEU1 population and replicated in additional populations, as described below.

Preliminary analysis using HapMap CEU1 individuals

The *DEFA1A3* copy number and ratios of three allelic variants are known for 84 haplotypes in the HapMap CEU1 population (table 5.1). This shows that there appear to be single copy, stepwise differences between haplotypes within each class. For example, there are 2, 3, 4, 5 and 7-copy Exchange 1 haplotypes, all lacking *DEFA3* and having only the 7bp duplication allele, each with only one copy of the Indel5 deletion

Sequence Class	Frequency in CEU1	Copy Number	*DEFA1*	*DEFA3*	Indel5 Del	Indel5 Ins	7bp Undup	7bp Dup
Reference Sequence	1	2	2	0	2	0	2	0
	3	4	3	1	4	0	4	0
	1	4	3	1	4	0	3	1
	3	4	3	1	3	1	3	1
	3	5	5	0	4	1	4	1
	16	5	4	1	4	1	4	1
	1	5	4	1	4	1	3	2
	1	5	4	1	3	2	3	2
	1	5	3	2	5	0	4	1
	1	5	3	2	4	1	4	1
Class 1	1	2	2	0	1	1	1	1
	6	3	3	0	1	2	1	2
	2	3	2	1	2	1	1	2
	1	4	4	0	1	3	1	3
Class 2	10	2	1	1	2	0	1	1
	1	3	3	0	2	1	2	1
	11	3	2	1	2	1	1	2
	1	6	6	0	4	2	4	2
Exchange 1	2	2	2	0	1	1	0	2
	3	3	3	0	1	2	0	3
	1	4	4	0	2	2	0	4
	8	4	4	0	1	3	0	4
	1	5	5	0	1	4	0	5
	1	5	4	1	2	3	1	4
	1	7	7	0	4	3	0	7
Exchange 2	3	3	3	0	3	0	3	0

Table 5.1: *DEFA1A3* haplotype copy numbers and ratios of three internal allelic variants for 84 independent haplotypes from the HapMap CEU1 population, split by *DEFA1A3* haplotype class. The frequency of each haplotype is also shown.

and a variable number of copies of the Indel5 insertion. Therefore, it is clear that multiple deletion and/or duplication events, involving repeat units highly similar in sequence, could have led to the presence of these five different alleles. In addition, related haplotypes appear to have similar properties. For example, most Exchange 1, Exchange 2 and Class 1 haplotypes lack *DEFA3*, whereas Class 2 and Reference Sequence haplotypes tend to have at least one copy of *DEFA3*.

However, without statistical analysis, it is unclear whether any of the observed associations are significant. This was addressed using a series of Chi Square and Cochran-Armitage tests comparing *DEFA1A3* haplotype class with features of the *DEFA1A3* locus for the HapMap CEU1 individuals (table 5.2). This shows that Class 2 haplotypes are associated with a low *DEFA1A3* copy number and Reference Sequence haplotypes with a high copy number. Class 2 haplotypes also display a low frequency of the Indel5 insertion, a high *DEFA3* frequency and a low frequency of the 7bp duplication. Exchange 1 haplotypes, whilst not associated with a high or low copy number, have a significant association with a high frequency of the Indel5 insertion, a near-absence of *DEFA3* and a high frequency of the 7bp duplication.

Whilst Exchange 2 haplotypes appear to share a high similarity across the *DEFA1A3* region, with all three examples having three copies and only *DEFA1* and the Indel5 deletion and 7bp unduplicated alleles, this class is present at such a low frequency in the CEU1 population that no significant associations can be identified. It would also be expected that Class 1 would be associated with a low *DEFA1A3* copy number, a high frequency of the Indel5 insertion and a low frequency of *DEFA3*; again, the low frequency of the Class 1 haplotype prevents identification

Class	**Copy number**			***DEFA3***		**Indel5 ins**			**7bp dup**	
Ref Seq	3-7		8-12	0-1	2-7	0-2		3-6	0-3	4-11
++	1		5	2	4	5		1	6	0
+-	15		19	19	15	19		15	21	13
- -	15		5	14	6	8		12	8	12
p-value	0.044; high			0.652		0.459			0.084	
Class 1	3-6	7-8	9-12	0-1	2-7	0-1	2-3	4-6	0-3	4-11
++/+-	5	8	1	11	3	0	11	3	6	8
- -	8	25	13	24	22	20	13	13	29	17
p-value	0.422			0.809		0.709			0.981	
Class 2	3-7		8-12	0-1	2-7	0-2		3-6	0-3	4-11
++	5		0	0	5	5		0	5	0
+-	15		4	8	11	14		5	14	5
- -	12		24	27	9	13		23	16	20
p-value	0.001; low			0.003; high		0.007; low			0.040; low	
Exchange 1	3-7		8-12	0-1	2-7	0-2		3-6	0-3	4-11
++	2		2	4	0	0		4	0	4
+-	5		12	14	3	2		15	2	15
- -	25		14	17	22	30		9	33	6
p-value	0.497			0.015; low		$2x10^{-5}$; high			$8x10^{-7}$; high	
Exchange 2	3-6	7-8	9-12	0-1	2-7	0-1	2-3	4-6	0-3	4-11
++/+-	1	5	0	6	0	5	1	0	5	1
- -	12	28	14	29	25	15	23	16	30	24
p-value	0.996			0.446		0.098			0.985	

Table 5.2: Outcome of Chi-Square and Cochran-Armitage tests comparing the *DEFA1A3* copy number or frequency of other variants with haplotype class at *DEFA1A3* in the HapMap CEU1 population. ++ = homozygous; +- = heterozygous; - - = negative. The p-value and the direction of significant associations are shown. P-values were adjusted for multiple testing using Bonferroni correction.

of these associations.

Wider European population

Whilst there are significant associations between haplotype class and *DEFA1A3* features in the HapMap CEU1 dataset, it is not clear if these 60 samples are representative of a wider European population. Therefore, the analysis was repeated using an additional 538 European individuals (CEU2 and HRC 1-5; table 5.3), the haplotype classes for which were identified using SNP genotyping, as described in section 2.9.

The results show that all of the association observed in the CEU1 population are replicated in the wider European dataset, with the exception of an association between Class 2 haplotypes and 7bp duplication frequency. However, this was a weak association and could have been a result of sampling effect. In addition, some novel associations have been identified. Class 1 haplotypes are associated with a low *DEFA1A3* copy number, a high frequency of the Indel5 insertion and a low frequency of the *DEFA3* gene; these associations were observed in the CEU1 population, but were not statistically significant. This is also true of the associations between Exchange 2 and a low frequency of both *DEFA3* and the Indel5 insertion. In addition, Exchange 1 haplotypes show a European-wide association with a high copy number, whilst Reference Sequence haplotypes are associated with a high frequency of *DEFA3* and a low frequency of both the Indel5 insertion and 7bp duplication, in addition to a high copy number. Overall, the results suggest that the *DEFA1A3* haplotypes found within the CEU1 population are representative of the *DEFA1A3* haplotypes found within the wider European population.

Class	**Copy number**			***DEFA3***		**Indel5 ins**			**7bp dup**	
Ref Seq	3-7		8-12	0-1	2-7	0-2		3-6	0-3	4-11
++	8		56	16	43	51		12	47	17
+-	122		114	107	111	175		57	182	53
- -	184		33	126	64	120		89	94	122
p-value	$2x10^{-27}$; high			$3x10^{-7}$; high		$9x10^{-5}$; low			$7x10^{-10}$; low	
Class 1	3-6	7-8	9-12	0-1	2-7	0-1	2-3	4-6	0-3	4-11
++/+-	80	39	6	88	24	15	90	18	76	50
- -	121	184	87	161	194	164	162	55	247	142
p-value	$1x10^{-10}$; low			$2x10^{-8}$; low		$9x10^{-5}$; high			1	
Class 2	3-7		8-12	0-1	2-7	0-2		3-6	0-3	4-11
++	41		2	2	37	40		2	25	18
+-	175		58	106	106	171		55	156	74
- -	98		143	141	75	135		101	142	100
p-value	$3x10^{-18}$; low			$7x10^{-10}$; high		$8x10^{-8}$; low			0.940	
Exchange 1	3-7		8-12	0-1	2-7	0-2		3-6	0-3	4-11
++	1		11	11	0	0		12	0	12
+-	54		40	69	11	20		74	11	83
- -	259		152	169	207	326		72	312	97
p-value	0.027; high			$3x10^{-12}$; low		$8x10^{-34}$; high			$1x10^{-39}$; high	
Exchange 2	3-6	7-8	9-12	0-1	2-7	0-1	2-3	4-6	0-3	4-11
++/+-	43	38	16	64	21	57	30	4	72	25
- -	158	185	77	185	197	122	222	69	251	167
p-value	0.959			$1x10^{-4}$; low		$6x10^{-8}$; low			0.170	

Table 5.3: Outcome of Chi-Square and Cochran-Armitage tests comparing the *DEFA1A3* copy number or frequency of other variants with haplotype class at *DEFA1A3* in the HapMap CEU2 and HRC 1-5 populations. ++ = homozygous; +- = heterozygous; - - = negative. The p-value and the direction of significant associations are shown. P-values were adjusted for multiple testing using Bonferroni correction.

The SNP rs4300027, which has previously been associated with *DEFA1A3* haplotype copy number in the European population, separates different classes of *DEFA1A3* haplotype. Class 1 and Class 2 haplotypes, both associated with a low copy number, have the C allele, whilst Exchange 1 and Reference Sequence haplotypes, both associated with a high copy number, have the T allele. In addition, the most common 2 and 3-copy haplotypes are both within Class 2, whilst the most common 4-copy haplotype is an Exchange 1 haplotype and the most common 5-copy a Reference Sequence haplotype (table 5.1). This suggests that the basis of the association between rs4300027 and *DEFA1A3* copy number is due to the combined effects of these four classes.

Other Worldwide Populations

Whilst there are some highly significant associations between haplotype class and features of the *DEFA1A3* locus in the European population, it is not clear if these associations are applicable to other worldwide populations. The same haplotype classes can be observed in all worldwide populations, but at varying frequencies (section 4.2). For samples used in the 1000 Genomes project [20], we have information on both the haplotype classes (derived using 1000 Genomes SNP genotype information) and *DEFA1A3* copy number (estimated using read depth). This allows tests comparing diploid *DEFA1A3* copy number with haplotype class genotype to be replicated for all worldwide populations (table 5.4). Exchange 1 is absent, or virtually absent, from Africa and Asia, so was not assessed in these populations.

The results from the European 1000 Genomes samples show a difference in the pattern of associations compared to the HapMap CEU and

Class	Africa	America	Asia	Europe
Ref Seq	0.768	$3x10^{-4}$; high	0.009; high	$2x10^{-26}$; high
Class 1	0.994	0.999	$9x10^{-16}$; high	0.346
Class 2	1	$2x10^{-6}$; low	$1x10^{-8}$; low	$1x10^{-18}$; low
Exchange 1	-	0.003; high	-	0.103
Exchange 2	1	0.879	0.045; low	$3x10^{-4}$; low

Table 5.4: Outcome of Cochran-Armitage tests comparing the diploid *DEFA1A3* copy number with haplotype class at *DEFA1A3* in the 1000 Genomes individuals. The p-values and the direction of significant associations are shown. P-values were adjusted for multiple testing using Bonferroni correction.

HRC individuals, despite the inclusion of 98 CEU samples within the 1000 Genomes dataset. Although the association of Class 2 haplotypes with a low copy number and Reference Sequence with a high copy number are observed, the associations between Class 1 and Exchange 1 with *DEFA1A3* copy number are not replicated. This suggests that there is additional variation within these classes in the European population that has not previously been observed. For the American population, we see similarities to the observed associations in the European population- a low copy number with Class 2 and a high copy number with the Reference Sequence and Exchange 1 classes. In the Asian population, we see an association of Class 1 with a high *DEFA1A3* copy number, in contrast to the previous observation of an association with a low copy number in the European population. In addition, we see associations of Class 2 and Exchange 2 with a low copy number and the Reference Sequence with a high copy number in the Asian population. There are no associations between haplotype class and *DEFA1A3* copy number in the African population, suggesting the within-class variation is much higher in the African population, compared with other populations. Whilst this comparison only looks at *DEFA1A3* copy number, it highlights the differences between worldwide populations in terms of the *DEFA1A3* compo-

sition of each haplotype class.

In the European population, the basis of the rs4300027 association with *DEFA1A3* copy number can be attributed to Class 1 and Class 2 haplotypes having a low *DEFA1A3* copy number and Reference Sequence and Exchange 1 haplotypes having a high copy number. This association is not replicated in all worldwide populations. Data from the 1000 Genomes project suggests that rs4300027 shows an association with *DEFA1A3* copy number in the American ($p=7x10^{-10}$) population and less so in the Asian population ($p=5x10^{-4}$), but no association in the African population (p=0.114). The basis of the association in the European population is the SNP's ability to split Class 1 and Class 2 haplotypes from Exchange 1 and Reference Sequence haplotypes; this effect is mirrored in the American population, in which the association of Class 2 with a low copy number and the Reference Sequence and Exchange 1 with a high copy number is observed. This is similar to the Asian population, in which Class 2 and Reference Sequence haplotypes, responsible for over 70% of haplotypes in the population, are associated with a low or high *DEFA1A3* copy number respectively. This cancels out the association of Class 1 with a high *DEFA1A3* copy number, which is at a low frequency in the Asian population. However, this does make the association weaker than that observed in the European population. Therefore, any association of rs4300027 with *DEFA1A3* copy number can be attributed to the properties of each haplotype class, but the association is not applicable to all worldwide populations, due to differences in the associations between haplotype class and *DEFA1A3* copy number between different populations and because of the varying frequencies of the classes between populations.

5.2 Ability of haplotype class to predict features of the *DEFA1A3* locus

Despite the strong association of the SNP rs4300027 with *DEFA1A3* haplotype copy number in the European population, it shows a poor ability to estimate diploid *DEFA1A3* copy number, with the correlation between the observed diploid *DEFA1A3* copy number and that estimated using rs4300027 genotype having an r^2 of just 0.35. However, using four SNP genotypes, which tag the five *DEFA1A3* haplotype classes, could improve on this ability to predict diploid *DEFA1A3* copy number, as well as predict the frequency of additional variants, including *DEFA3*. This is possible given the observed associations between haplotype class and features of the *DEFA1A3* locus in the European population (section 5.1). Using the haplotype information from the HapMap CEU1 samples, the average *DEFA1A3* copy numbers and frequency of variants were calculated for haplotypes within each of the five *DEFA1A3* classes. These were used to provide the expected values for each diploid class combination, with values rounded to the nearest integer value (table 5.5).

Using these expected values, it was possible to determine the ability of *DEFA1A3* haplotype class (i.e. 4 SNP genotypes) to predict features of the *DEFA1A3* locus. By comparing the expected values to those observed for an independent European dataset, the HapMap CEU2 and HRC 1-5 populations (table 5.6). This shows that using a combination of four SNP genotypes to predict features of the *DEFA1A3* locus is always better than rs4300027 alone; this is seen most dramatically for the ability to predict Indel5 insertion frequency (figure 5.1), as the addition of more SNPs allows for more possible expected values. However, this improvement in correlation is variable. For example, only a small improvement

Haplotype Classes	*DEFA1A3* CN	*DEFA3* freq	Indel5 ins freq	7bp dup freq
RS / RS	9	2	2	2
RS / C1	8	1	3	3
RS / C2	7	2	1	2
RS / E1	8	1	3	5
RS / E2	8	1	1	1
C1 / C1	6	0	4	4
C1 / C2	6	1	2	4
C1 / E1	7	0	4	6
C1 / E2	6	0	2	2
C2 / C2	5	2	1	3
C2 / E1	7	1	3	5
C2 / E2	6	1	1	2
E1 / E1	8	0	5	8
E1 / E2	7	0	3	4
E2 / E2	6	0	0	0

Table 5.5: The most frequent diploid *DEFA1A3* copy number and ratios for the *DEFA1*/*DEFA3*, Indel5 and 7bp duplication variants for each diploid haplotype class combination, based on the haplotypes observed in the HapMap CEU1 population. RS= Reference Sequence; C1= Class 1; C2= Class 2; E1= Exchange 1; E2= Exchange 2.

Feature	rs4300027	Haplotype Class
Copy Number	0.401; 2.31×10^{-59}	0.418; 1.61×10^{-62}
DEFA3	0; 1	0.266; 5.01×10^{-33}
Indel5 Insertion	0.068; 2.68×10^{-9}	0.501; 6.8×10^{-78}
7bp Duplication	0.031; 5.4×10^{-5}	0.311; 1.7×10^{-43}

Table 5.6: The r^2 and p-values for the correlations between the observed features of the *DEFA1A3* locus and those estimated using either rs4300027 genotype or Haplotype class (table 5.5) for the HapMap CEU2 and HRC 1-5 samples. Both values were calculated using the regression function of Excel.

in the ability to predict *DEFA1A3* copy number is observed, despite the fact that four of the five haplotype classes are significantly associated with either a high or low *DEFA1A3* copy number. One interesting observation is the poor ability to predict *DEFA3* copy number, despite the fact that four classes show significant associations with *DEFA3* frequency. In contrast, the four SNPs can better estimate 7bp duplication frequency, despite this only showing an association with two classes. Overall, this shows that, whilst increasing the number of variants improves the ability to predict features of the *DEFA1A3* CNV locus, even the use of four

SNPs provides a poor estimation.

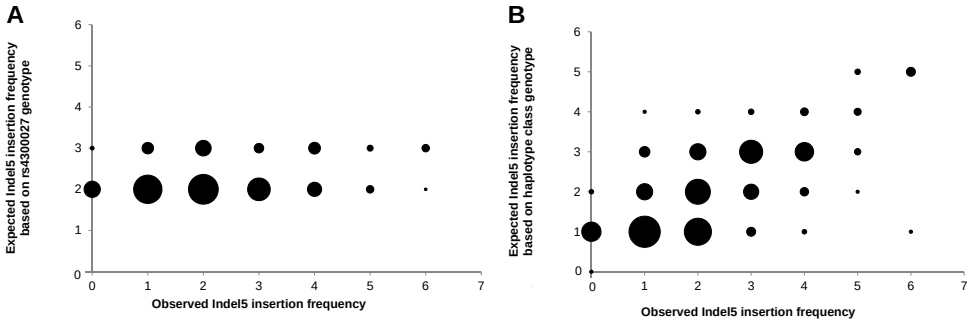


Figure 5.1: Plots comparing the observed and expected Indel5 insertion frequencies estimated using either the SNP rs4300027 or *DEFA1A3* haplotype class. Whilst rs4300027 estimates only two different frequencies, using haplotype class gives six different expected values, allowing a more accurate estimation of Indel5 insertion frequency and increasing the r^2 value from 0.068 to 0.501. The area of points are scaled to represent the frequency of individuals.

5.3 Testing for associations between haplotype class and *DEFA1A3* microsatellite allele length

As each *DEFA1A3* haplotype class has been associated with features of the *DEFA1A3* locus, it was proposed that each haplotype class may also show associations with particular length alleles of the *DEFA1A3* microsatellite. 108 individuals homozygous for sequence class from the ECACC HRC population were typed for the *DEFA1A3* microsatellite, in order to identify potential associations between sequence class and *DEFA1A3* microsatellite allele sizes. These hypotheses were then tested using 76 independent CEPH individuals.

The ECACC HRC samples indicated that 9 of the 13 microsatellite alleles were rare or did not appear to show any association with a particular sequence class; these were the alleles in the range 89bp-120bp. However,

the alleles 43bp, 51bp, 57bp and 66bp appeared to show associations with *DEFA1A3* haplotype class (table 5.7).

Haplotype Class	43bp	51bp	57bp	66bp
Reference Sequence	-	-	High	-
Class 1	High	-	-	-
Class 2	-	High	Low	-
Exchange 1	-	-	High	Low
Exchange 2	-	-	-	-

Table 5.7: Hypothesised associations between *DEFA1A3* haplotype class and *DEFA1A3* microsatellite allele frequencies based on observations in the HRC population. High= appears to contain a high frequency of microsatellite allele of given size; Low= appears to contain a low frequency of microsatellite allele of given size.

These associations were then tested for significance in the CEPH population using either Chi Square or Cochran-Armitage tests, in which categories were based on each containing an equal number of individuals (table 5.8). As for previous tests for association, homozygotes and heterozygotes for Class 1 and Exchange 2 were grouped, due to their low frequency in the study population. This shows that all observed associations in the HRC population are replicated in the CEPH population and are statistically significant. In addition, some weaker associations of the Reference Sequence and Exchange 1 with a low frequency of the 51bp allele and Exchange 2 with a high frequency of the 51bp allele, are identified. Class 2 is also significantly associated with a high frequency of the 66bp allele. This suggests that, although highly variable, the *DEFA1A3* microsatellite shows associations with haplotype class at *DEFA1A3*, in which particular alleles are at a significantly high or low frequency, compared with the rest of the population.

Class	**43bp**		**51bp**			**57bp**			**66bp**	
Ref Seq	0	1+	0	1+		0-3	4+		0-1	2+
++	6	1	6	1		1	6		4	3
+-	30	8	16	22		18	20		16	22
- -	22	9	5	26		24	7		13	8
p-value	0.971		0.003; low			0.006; high			0.999	
Class 1	0	1+	0	1	2+	0-1	2-3	4+	0-1	2+
++/+-	4	12	8	7	1	5	3	8	12	4
- -	54	6	19	24	17	18	17	25	21	39
p-value	1x10^{-6}; high		0.450			1			0.079	
Class2	0	1+	0	1+		0-3	4+		0-1	2+
++	10	0	0	10		9	1		0	10
+-	22	6	1	27		26	2		6	22
- -	26	12	26	12		8	30		27	11
p-value	0.330		4x10^{-7}; high			3x10^{-7}; low			1x10^{-5}; high	
Exchange 1	0	1+	0	1+		0-3	4+		0-1	2+
++	3	1	4	0		0	4		4	0
+-	15	3	9	9		6	12		12	6
- -	40	14	14	39		37	17		17	37
p-value	0.999		0.012; low			0.005; high			0.005; low	
Exchange 2	0	1+	0	1	2+	0-1	2-3	4+	0-1	2+
++/+-	7	3	0	6	4	5	3	2	4	6
- -	51	15	27	25	14	18	17	31	29	37
p-value	1		0.011; high			0.577			1	

Table 5.8: Outcome of Chi-Square and Cochran-Armitage tests comparing the frequency of *DEFA1A3* microsatellite allele sizes with *DEFA1A3* haplotype class in 76 CEPH individuals. ++ = homozygous; +- = heterozygous; - - = negative. The p-value and the direction of significant associations are shown. High= associated with a high frequency of the microsatellite allele with given size; Low= associated with a low frequency of the microsatellite allele with given size. P-values were adjusted for multiple testing using Bonferroni correction.

5.4 Conclusions

In conclusion, each of the five *DEFA1A3* haplotype classes is associated with features of the *DEFA1A3* locus in the European population, with each class displaying a unique profile of associations. These indicate the likely properties of each haplotype class, for example, Class 1 haplotypes having a low *DEFA1A3* copy number, a low frequency of the *DEFA3* gene and a high frequency of the Indel5 insertion. However, haplotype class is a poor predictor of the absolute values of these features. This is likely due to within-class variation, which means each class may be associated with an overall trend toward a low or high frequency of a particular feature, but the absolute value can differ. For example, haplotypes within the Exchange 1 class have been observed with 2-7 copies, yet the class still has a significant association with a high *DEFA1A3* copy number. Although these features (e.g. Indel5, 7bp duplication) may not themselves be functionally important, they may be associated with variants which do influence the expression of *DEFA1A3* and, as such, knowing their composition within each *DEFA1A3* haplotype class is important.

The maximum possible r^2 value for a SNP tagging multiallelic copy number variation is 0.73, in which the diploid copy number range was 3-12, with the same haploid and diploid copy number frequencies as observed for *DEFA1A3*. The four SNPs tagging *DEFA1A3* haplotype class are only able to achieve an r^2 value of 0.413. This is representative of the fact that *DEFA1A3* copy number is changing faster than the surrounding sequence, such that even a combination of SNPs tagging haplotypes with distinct properties cannot perfectly tag *DEFA1A3* copy number. Therefore, this suggests that SNPs will act as poor tags of exact copy number states, but can be very informative in indicating the general features of a

haplotype.

The use of data from the 1000 Genomes project allowed a comparison of *DEFA1A3* haplotype class with copy number in other worldwide populations. This indicated that the associations varied between worldwide populations, especially in the African population, which appears to be highly variable, with no haplotype class showing an association with *DEFA1A3* copy number. Whilst the associations with *DEFA1A3* copy number have been investigated in other non-European populations, it has not yet been possible to compare other features of the *DEFA1A3* locus with haplotype class. This will be necessary to gain a detailed understanding about the within-class variation between different worldwide populations. Using read depth to infer the ratio of variants at *DEFA1A3* is unsuitable, due to the relatively low coverage of the sequencing achieved by the 1000 Genomes project.

One outcome of the tests for association has been determining the basis of the previously observed association of the SNP rs4300027 in the European population with *DEFA1A3* copy number. The SNP's ability to separate different *DEFA1A3* haplotype classes provides the basis of this association. Therefore, the Asian population shows a weaker association than the European population, because Class 1 haplotypes are associated with a high copy number in Asia, as opposed to a low copy number in Europe. Therefore, this SNP is acting as a tag of *DEFA1A3* lineages, rather than *DEFA1A3* copy number itself.

The observation that certain *DEFA1A3* microsatellite alleles are associated with *DEFA1A3* haplotype class is interesting, given that the mutation rate for the microsatellite will exceed that for *DEFA1A3* copy number. This, along with the smaller sample size, is likely to contribute to the less

significant associations than observed for some of the other features. However, the associations suggest additional features shared between related haplotypes and provide another internal variant from which structural information could be derived. In summary, there appears to be high similarity in the features of related haplotypes at *DEFA1A3*.

6 *DEFA1A3* structural haplotypes

The structural organisation and phasing of variants in the human genome can affect gene expression and phenotype [111]. At CNV loci, this issue is made more complex, as it requires not only the identification of variants on each haplotype, but their positions relative to each other across a haplotype. As demonstrated in previous chapters, much is known about *DEFA1A3* haplotypes in the European population, in which we know the total haplotype composition for multiple variants. However, very little is known about the positions of the alleles of these variants in relation to each other.

Aldred *et al.* used long PCR to show that *DEFA3*, the less common of the two genes at *DEFA1A3*, can occupy any position across the repeat array, but was biased towards the centromeric-most repeat [55]. Emulsion haplotype fusion-PCR (EHF-PCR) has previously been applied to the *DEFA1A3* locus to obtain spatial information about the relative positions of the *DEFA1* and *DEFA3* genes across a haplotype [131]. This technique could be applied to additional variants (e.g. Indel5) to provide additional structural information for *DEFA1A3* haplotypes. Sequencing has allowed us to identify related haplotypes and the comparisons of the structures of related haplotypes would allow the inference of possible mechanisms resulting in the deletion and/or duplication of copies of *DEFA1A3*. In addition, in comparison with gene expression data, it would be possible to infer how the spatial arrangement of the *DEFA1A3* locus is related to the expression of the *DEFA1A3* genes.

The aim was to obtain spatial information for *DEFA1A3* haplotypes of European ancestry from each of the five major *DEFA1A3* haplotype classes and use the structures obtained to investigate the mechanisms responsible for changes in *DEFA1A3* copy number, as well as investigate how haplotype structure may impact on *DEFA1A3* expression. It is not necessarily thought that these non-coding variants (i.e. Indel5, 7bp duplication, *DEFA1A3* microsatellite) are themselves responsible for changes in gene expression, but it is possible that they could act as tags of functional variants and, indirectly, of sequence-independent position effects, given their associations with particular haplotype classes.

6.1 Separation of the *DEFA1A3* full and partial repeats

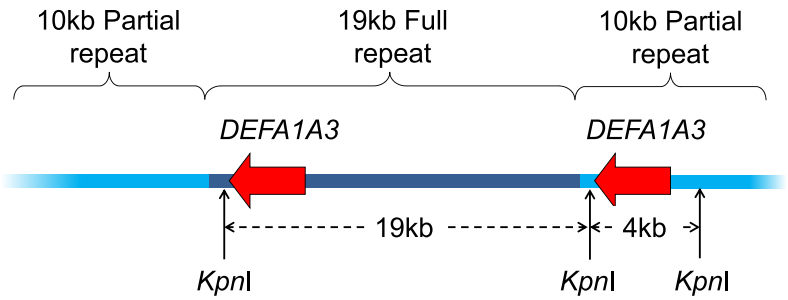


Figure 6.1: The *Kpn*I restriction sites found across the *DEFA1A3* locus are highlighted. There is a single site per full repeat and two sites within the partial repeat, leaving the partial-repeat copy of *DEFA1A3* on a 4kb restriction fragment and all full-repeat copies on a 19kb fragment, which can be separated using gel electrophoresis.

Due to sequence differences between the *DEFA1A3* full and partial repeats, it is possible to digest genomic DNA with the enzyme *Kpn*I and separate the full repeat copies of the *DEFA1A3* gene from the partial repeat copies (figure 6.1). This allows the contents of the *DEFA1A3* full and partial repeats to be analysed separately which, in combina-

tion with phased haplotype information, provides structural information across *DEFA1A3* haplotypes.

The *DEFA1A3* full and partial repeats were separated for nine CEPH individuals. For each sample, the DefHae3 and 7bp duplication ratios were obtained for the *DEFA1A3* full and partial repeats separately (table 6.1 and figure 6.2); the Indel5 assay was not used, because the centromeric-most Indel5 site falls outside of the 4kb *Kpn*I partial repeat restriction fragment. In most cases, the full and partial DefHae3 ratios can be rounded to integer values which sum to give the expected diploid total for that sample. For example, NA07008 has a full repeat ratio of 3.55, which is close to 4:1, and a partial repeat ratio of 1.17, which is close to 1:1, giving a diploid total of 5:2, as expected. The exception is sample NA07053, for which only *DEFA1* is observed in the full repeats; therefore, the partial repeat DefHae3 ratio should be 1:1, but the observed ratio is 0.33.

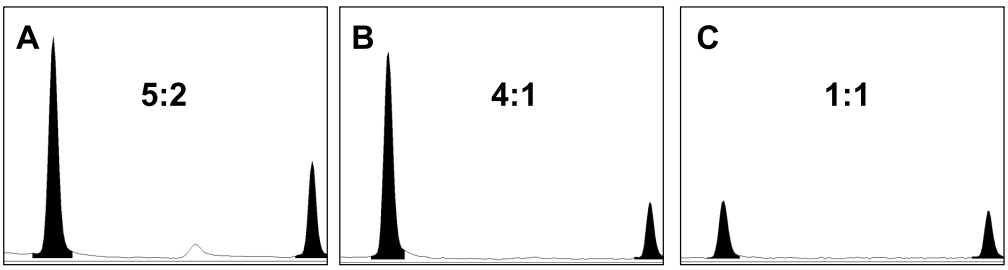


Figure 6.2: DefHae3 ratios for the CEPH sample 1340-05. A: total diploid ratio 5 *DEFA1* : 2 *DEFA3*; B: Full repeat ratio 4:1, taken from $\geq$10kb gel block; C: Partial repeat ratio 1:1, taken from 4kb gel block.

The 7bp duplication ratios are harder to interpret. For samples NA07008, NA12000, NA12806, NA12807 and NA12810, it is possible to round the observed 7bp duplication ratios to values which sum to give the expected diploid total. However, for the other samples, the results are less clear. For example, the partial repeat ratio for sample NA07062 should be 1:1,

Sample	Diploid ratio		Unrounded Defhae3 ratio		DefHae3 ratio		Unrounded 7bp dup ratio		7bp dup ratio	
	DefHae3	7bp dup	Full	Partial	Full	Partial	Full	Partial	Full	Partial
NA07062	4:1	1:4	2.78	3.1	2:1	2:0	Dup only	12.75	0:3	1:1
NA07053	11:1	4:8	*DEFA1* only	0.33	10:0	1:1	0.48	4.61	3:7	1:1
NA07008	5:2	5:2	3.32	1.17	4:1	1:1	1.73	13.75	3:2	2:0
NA11998	5:1	2:4	3.55	*DEFA1* only	3:1	2:0	0.24	4.53	-	-
NA11999	6:0	4:2	*DEFA1* only	*DEFA1* only	4:0	2:0	2.02	2.01	-	-
NA12000	5:1	4:2	3.91	*DEFA1* only	3:1	2:0	1.64	8.75	2:2	2:0
NA12806	8:1	5:4	*DEFA1* only	1.37	7:0	1:1	0.71	10.85	3:4	2:0
NA12807	7:0	4:3	*DEFA1* only	*DEFA1* only	5:0	2:0	1.03	5.74	2:3	2:0
NA12810	4:1	4:1	3.04	6.92	2:1	2:0	1.93	Undup only	2:1	2:0

Table 6.1: Ratios obtained from the DefHae3 and 7bp duplication assays for the *DEFA1A3* full and partial repeats, compared to the expected diploid total, for nine CEPH individuals. The ratios are rounded to show the most likely integer ratio.

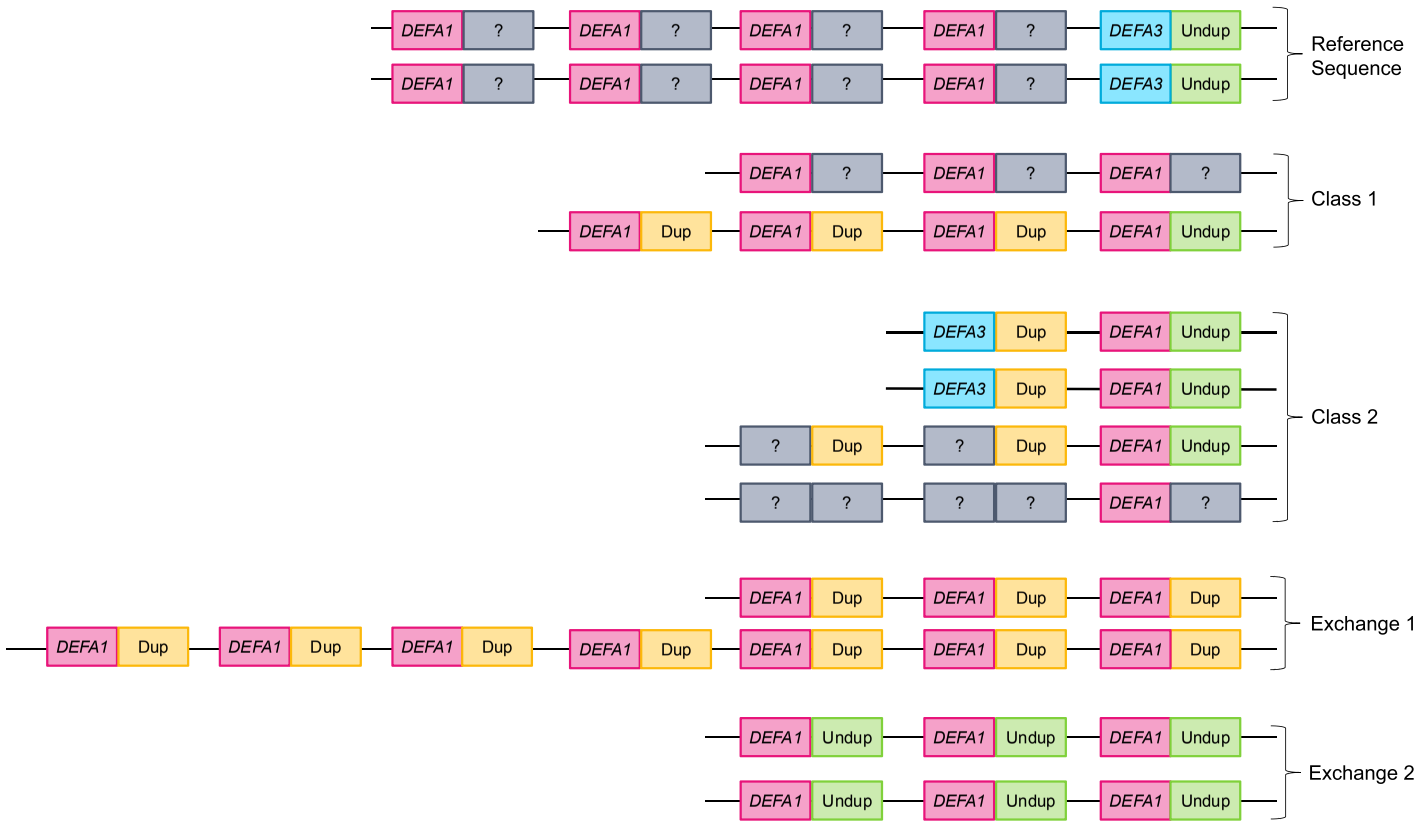


Figure 6.3: Structural information for 12 independent haplotypes, based on the *DEFA1*/*DEFA3* and 7bp duplication variants, separated by *DEFA1A3* haplotype class. Although some variant positions remain unclear (indicated by ?), there still appears to be some within-class coherence in structure.

given the presence of only the duplication allele in the full repeats; however, a ratio of 12.75 was observed. For sample NA07053, the most likely ratios are 3:7 for the full repeats and 1:1 for the partial repeats, but the ratios could be interpreted to give 2:8 and 2:0. For samples NA11998 and NA11999, it is unclear whether the partial repeat ratio is 1:1 or 2:0 and the full repeat ratios do not clarify this. Therefore, no ratio values have been assigned for these samples.

In combination with the DefHae3 and 7bp duplication ratios for each haplotype, this technique has provided complete structural information for 3 independent haplotypes and partial structural information for 5 independent haplotypes, in relation to the *DEFA1*/*DEFA3* and 7bp duplication variants (figure 6.3). For example, NA07053 has a partial DefHae3 ratio of 1:1, but the 7-copy haplotype is known to be *DEFA3* absent, so the *DEFA3* must be the centromeric-most gene on the 5-copy haplotype. The 7bp duplication ratio for the partial repeat is also 1:1 and again, the 7-copy haplotype contains only the duplication, so the centromeric-most copy of the 5-copy haplotype must be the unduplicated allele. The structure of four haplotypes, in relation to these two variants, was already known, as the haplotypes contained only one allele of these variants, but this information helped to confirm the positions of alleles on the other haplotype for that sample. Whilst the information gained from this technique is limited and is incomplete for some haplotypes, we still see evidence of shared and similar structures of haplotypes within the same *DEFA1A3* haplotype class.

6.2 Emulsion haplotype fusion-PCR

EHF-PCR can provide positional information at CNV loci [131]. Due to the relatively small size of the *DEFA1A3* repeat unit (19kb), fusions to the second repeat have been observed, identified as the minor product following Sanger sequencing. However, fusions to the third copy on a haplotype cannot be observed using Sanger sequencing (Jess Tyson, personal communication). This allows up to four copies of a variant to be observed per haplotype.

The order of the *DEFA1* and *DEFA3* genes, as well as the Inserted and Deleted alleles of a 5bp Indel located upstream of each copy of *DEFA1A3*, were determined for 61 independent haplotypes of European ancestry (HapMap CEU1) (figure 6.4). This was achieved using a combination of four EHF-PCR systems, in addition to inferring the structure of haplotypes containing only one allele for a particular variant (e.g. *DEFA1* only) and sequencing data (chapter 4), which captured the centromeric-most copy of the 5bp Indel. Across the 120 haplotypes sequenced, only the Deleted Indel5 allele was observed in the centromeric-most repeat.

The structures of 17 Reference Sequence haplotypes were determined, 12 of which display a structure shown in figure 6.4. There are three different structures shown, which are highly similar to each other and there is a single copy difference between the structures that could be accounted for by a single deletion or duplication event. The exceptions to this are two haplotypes with only two copies of *DEFA1A3*, which is uncommon for Reference Sequence haplotypes. This is in addition to three haplotypes in which the telomeric-most Indel5 site is the Deleted allele and the second site from the telomeric end contains the Inserted allele, showing an

apparent rearrangement from the more frequently observed structure.

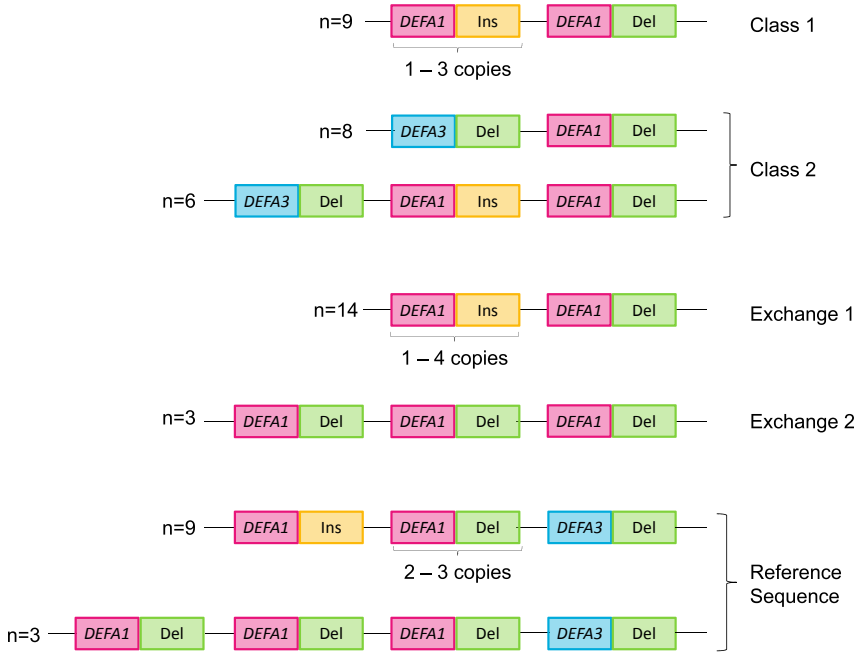


Figure 6.4: The common structures of *DEFA1A3* haplotypes for each class as observed in the HapMap CEU1 population, with regards to the *DEFA1*/*DEFA3* and Indel5 variants. Ins/Del= status of Indel5 variant.

For Class 1 haplotypes, 9 of the 11 structures observed display a structure shown in figure 6.4. These contain a variable number of a repeat unit containing *DEFA1* and the Indel5 insertion. Therefore, not only are the structures similar, but they share a variable number of copies of a highly similar repeat unit. The two exceptions to this are haplotypes containing a copy of *DEFA3*, which is rare for Class 1 haplotypes.

14 of the 15 Class 2 haplotypes observed display a structure shown in figure 6.4. Again, the two different structures observed are highly similar and a single deletion or duplication event could account for the difference between the two. The exception is a *DEFA3* -absent haplotype, which is rare for Class 2 haplotypes.

For Exchange 1 haplotypes, 14 of the 15 structures observed display

a structure shown in figure 6.4. They contain a variable number of copies of a repeat unit containing *DEFA1* and the Indel5 Inserted allele. Therefore, they are structurally similar to Class 1 haplotypes, although this does not necessarily make them more closely related than to other classes. Again, Exchange 1 haplotypes share both high structural and repeat unit similarity. The exception is a haplotype containing *DEFA3*, which is rare for haplotypes within the Exchange 1 class. Although an EHF-PCR has not been used to determine the relative positions of the 7bp duplication variant across a haplotype, for the majority of Exchange 1 haplotypes, all repeats contain the duplicated allele, further showing a high sequence similarity between repeat units across a haplotype.

There are only three Exchange 2 haplotype for which haplotype-specific copy number information is available. All three haplotypes have the same structure; three copies of a repeat unit containing *DEFA1* and the Indel5 Deletion allele. In addition, all repeat units contain the unduplicated allele of the 7bp duplication.

The structures obtained suggest high within-class and low between-class structural similarity at *DEFA1A3*. However, a more detailed analysis would come from looking at additional variants across the *DEFA1A3* locus. Therefore, the positions of the *DEFA1A3* microsatellite alleles were determined for 25 independent haplotypes of European ancestry (HapMap CEU1)(figure 6.5). Although the *DEFA1A3* microsatellite will be mutating at a faster rate than *DEFA1A3* copy number, there is still high within-class coherence in the structures observed.

For the Reference Sequence haplotypes, the centromeric-most *DEFA1A3* microsatellite allele is, in most cases, the 66bp allele, and all internal repeats carry the 57bp allele, whilst the telomeric-most allele is variable.

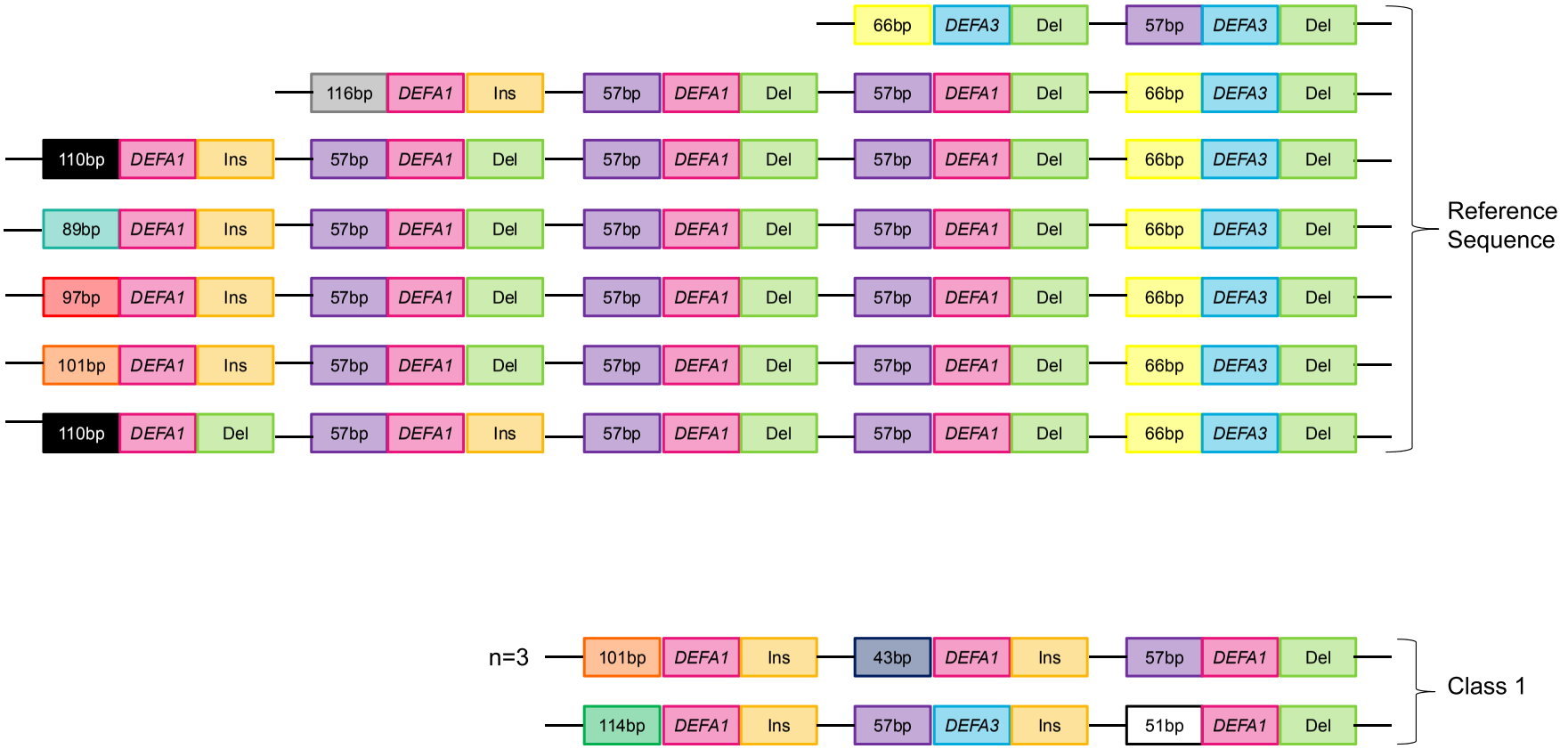
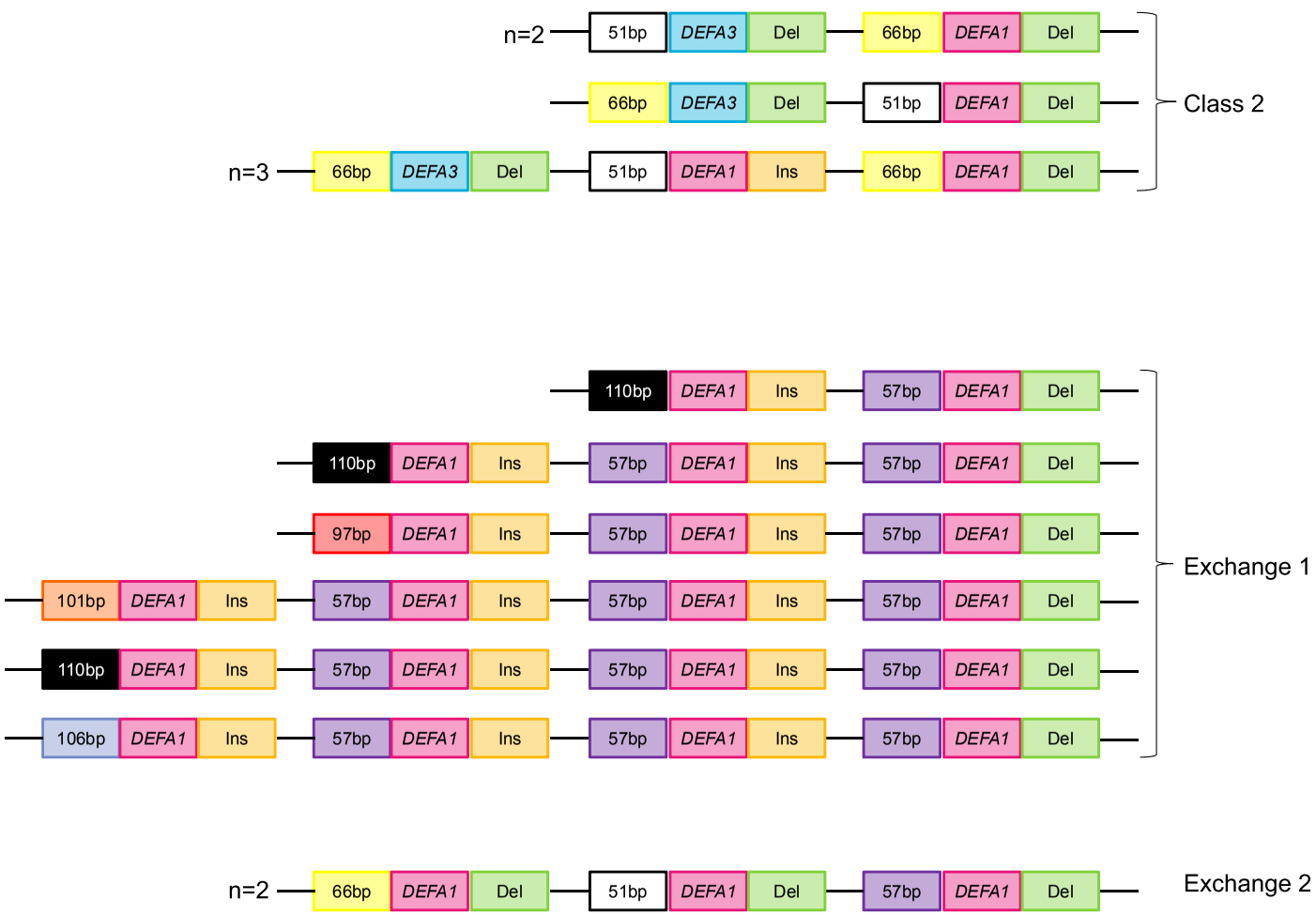


Figure 6.5: Structures of 25 HapMap CEU1 haplotypes with regards to the *DEFA1A3* microsatellite, *DEFA1*/*DEFA3* and Indel5 variants. 43-116bp= size of microsatellite allele; Ins or Del= status of Indel5 variant. For all structures, n=1 unless stated. Continues overleaf.


n=2
51bp DEFA3 Del
66bp DEFA1 Del
66bp DEFA3 Del
51bp DEFA1 Del
n=3
66bp DEFA3 Del
51bp DEFA1 Ins
66bp DEFA1 Del
Class 2
110bp DEFA1 Ins
57bp DEFA1 Del
110bp DEFA1 Ins
57bp DEFA1 Ins
57bp DEFA1 Del
97bp DEFA1 Ins
57bp DEFA1 Ins
57bp DEFA1 Del
101bp DEFA1 Ins
57bp DEFA1 Ins
57bp DEFA1 Ins
57bp DEFA1 Del
110bp DEFA1 Ins
57bp DEFA1 Ins
57bp DEFA1 Ins
57bp DEFA1 Del
106bp DEFA1 Ins
57bp DEFA1 Ins
57bp DEFA1 Ins
57bp DEFA1 Del
Exchange 1
n=2
66bp DEFA1 Del
51bp DEFA1 Del
57bp DEFA1 Del
Exchange 2

The exception is the two copy haplotype, which, as stated previously, is unusual for a Reference Sequence haplotype. Again, this suggests a high similarity in the sequence of individual repeat units across a haplotype, in addition to within-class structural similarity.

Three of the four Class 1 haplotypes have the same microsatellite allele structure, whilst the fourth has a different structure, but also contains the *DEFA3* gene, which is uncommon for Class 1 haplotypes. For the Class 2 haplotypes, the three 3-copy haplotypes observed have the same *DEFA1A3* microsatellite structure and one 2-copy haplotype shows a simple single-copy deletion compared to the 3-copy haplotypes. However, more than one rearrangement event would be required to explain the differences between the other type of 2-copy haplotype and the 3-copy haplotypes observed.

The Exchange 1 haplotypes are very coherent, in which all repeat units contain the 57bp allele, with the exception of the telomeric-most site, which is variable, again showing a high similarity between repeat units across a haplotype. Only two Exchange 2 haplotypes were observed, both of which have the same microsatellite profile. Overall, the *DEFA1A3* microsatellite structural information supports the within-class structural similarity observed at the *DEFA1A3* locus and, in some cases, a high similarity between repeat units across a haplotype, in addition to between-class differences in structure. The higher variability for this feature is likely due to the higher mutation rate of the *DEFA1A3* microsatellite, relative to the *DEFA1A3* mutation rate.

One advantage of capillary electrophoresis for the analysis of EHF-PCR products over Sanger sequencing appears to be the increased sensitivity. Figure 6.6 shows the profiles obtained for the sample NA12282

following EHF-PCR, clearly showing a fusion product is obtained from the third repeat. However, the ability to fuse to the third repeat is only useful for a locus like the *DEFA1A3* microsatellite, for which there are many different allele sizes.

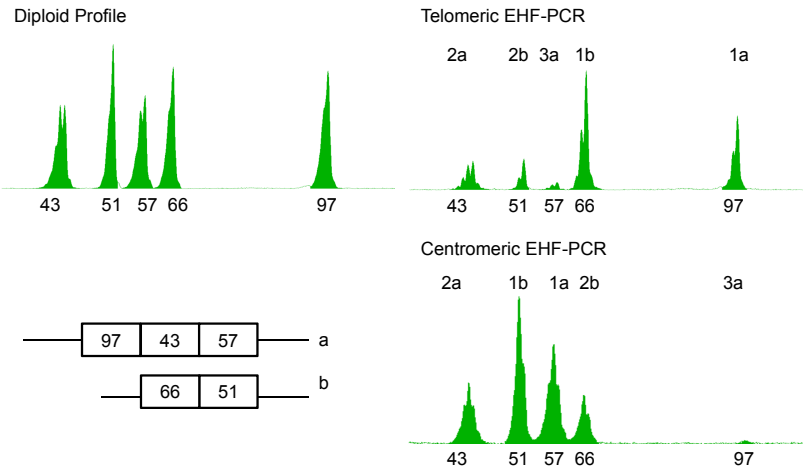


Figure 6.6: The sample NA12282 (HapMap CEU2) has five different *DEFA1A3* microsatellite alleles and a diploid *DEFA1A3* copy number of 5. Non-allele specific reamplification of telomeric and centromeric microsatellite EHF-PCRs identifies the same five peaks, but the areas of the peak are relative to their position along the array. For example, the 97bp allele has one of the largest peak areas from the Telomeric EHF-PCR, yet has the smallest peak area from the Centromeric EHF-PCR. This indicates its position to be the telomeric-most allele on the three-copy haplotype. Combined with phased haplotype information, the position of all five alleles can be determined, showing capillary electrophoresis is sensitive enough to detect fusion products from the third repeat across a *DEFA1A3* haplotype.

6.3 Comparing *DEFA1A3* haplotype class with HNP1-3 expression

As shown in sections 6.1 and 6.2, each *DEFA1A3* haplotype class appears to display within-class coherence in structure, whilst being distinct from the structures observed within the other *DEFA1A3* haplotype classes. Therefore, it appears that, within the European population,

DEFA1A3 haplotype class alone can be used to indicate *DEFA1A3* haplotype structure and thus, by comparing *DEFA1A3* haplotype class with HNP1-3 expression, it would allow the inference of how *DEFA1A3* haplotype structure influences the expression of the *DEFA1A3* genes. The *DEFA1A3* locus expresses three peptides, HNP1 from *DEFA1*, HNP3 from *DEFA3* and HNP2, which is a proteolytic product of HNP1 and/or HNP3 [55, 84]. The relationship between *DEFA1A3* copy number and the amount of HNP1-3 expressed is unclear; whilst one small study identified a positive correlation [85], a larger study found no relationship (Danielle Carpenter, personal communication). Another study found no association between *DEFA1A3* copy number and mRNA expression, although the ratio of *DEFA1* and *DEFA3* appeared to be maintained during mRNA expression [55]. However, mRNA expression was measured in neutrophils and not promyelocytes, where the mRNA was expressed; therefore, this may not be a true reflection of the real expression levels. Despite the lack of associations between *DEFA1A3* copy number and HNP1-3 expression, it may be that other features, such as the relative positions of the *DEFA1* and *DEFA3* genes, as well as other internal variants, may influence HNP1-3 expression.

Total HNP1-3 expression has been measured for 118 of the 120 UK individuals as part of BBSRC Project BB/I006370/ (Danielle Carpenter and Laura Mitchell; section 2.12). DNA was also collected for these samples, allowing the diploid *DEFA1A3* haplotype class profile to be obtained via the genotyping of four SNPs, as described for the HapMap CEU and HRC samples in section 4.2. Complete SNP genotype data was obtained for 119 of the 120 samples; the remaining sample was excluded from further analysis. Total HNP1-3 expression was compared with *DEFA1A3* haplotype class genotype for 117 samples (figure 6.7).

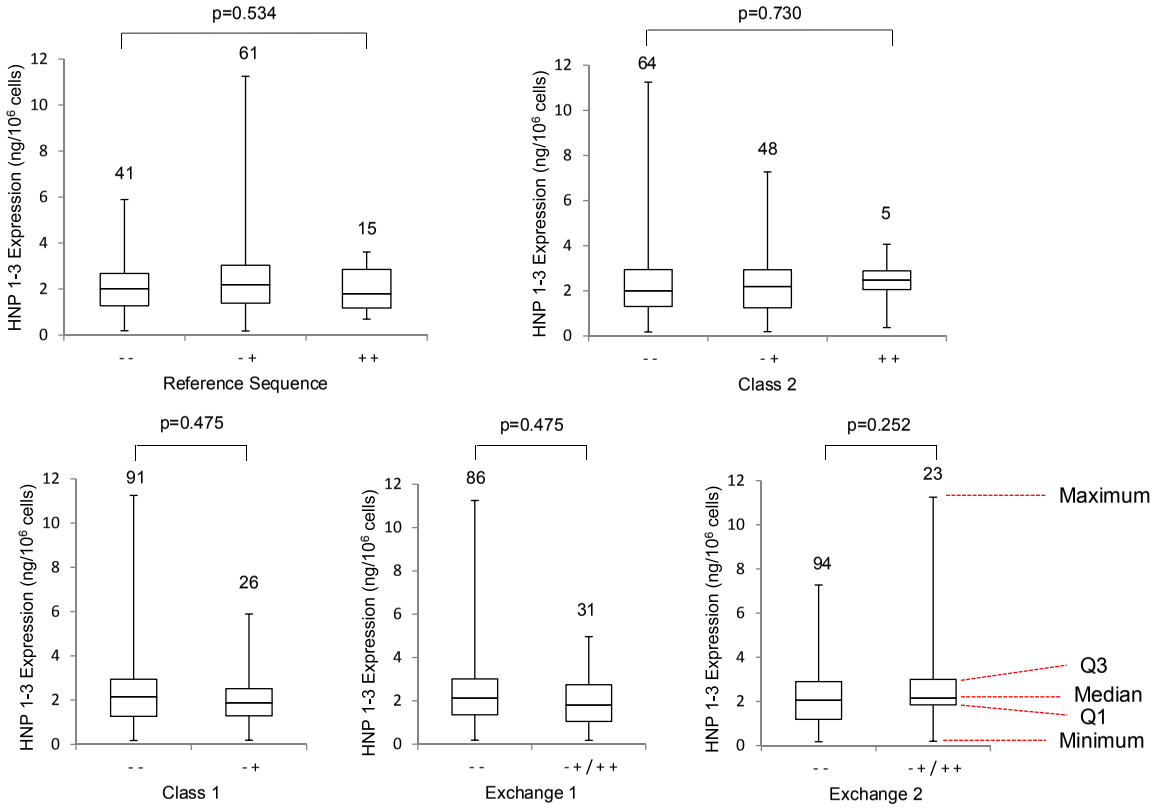


Figure 6.7: Box plots showing HNP1-3 expression for each *DEFA1A3* haplotype class genotype for 117 UK individuals. ++ = homozygous positive; +- = heterozygous; - - = homozygous negative. Number above each box plot indicates number of individuals represented in that category.

This indicates that there is no association between *DEFA1A3* haplotype class and HNP1-3 expression, as no class shows an association with either a high or low level of HNP1-3 expression.

6.4 Conclusions

In conclusion, EHF-PCR provides an effective method for obtaining structural information at multiallelic CNV loci. In combination with phased haplotype and sequencing information, it has allowed complete phasing for the *DEFA1*/*DEFA3* and Indel5 variants across 61 independent European *DEFA1A3* haplotypes and for 25 of those haplotypes with regards to the *DEFA1A3* microsatellite variant. The separation of the *DEFA1A3* full and partial repeats using a *Kpn*I digest provided some spatial information, but this method had many limitations. Firstly, it provided the most information for the centromeric partial repeat and, when two different alleles were found within the full repeats, the spatial arrangement of the alleles could not be deduced. In addition, it is a DNA-costly technique, compared to EHF-PCR. Also, the gel electrophoresis step of the protocol suffers similar issues to that of PFGE, in that DNA migrates to positions of the gel not related to the size expected, distorting the ratios from their true value and making it hard to identify cases where one allele is absent from either the full or partial repeats. Again, contamination from other *DEFA1A3* PCR products will have to the error observed. Also, in some cases, it was necessary to sample the same haplotype in different contexts in order to determine the positioning of variants. In addition, given the post-PCR analysis method, it only provides information about a single variant position, either the 7bp duplication or *DEFA1*/*DEFA3* status, whereas EHF-PCR can look at multiple variant positions in a single

reaction, given the use of post-fusion sequencing. Therefore, EHF-PCR is more useful for providing spatial information at multiallelic CNV loci.

Haplotypes within each *DEFA1A3* class display high structural similarity, despite differences in copy number. This is combined with between-class differences in structure. For example, *DEFA3* is always found at the telomeric-most repeat on Class 2 haplotypes, whether they have two or three copies of *DEFA1A3*, but at the centromeric-most repeat on Reference Sequence haplotypes. This supports the observation by Aldred *et al.* that there is a bias for *DEFA3* to be found at the centromeric-most repeat of the *DEFA1A3* array [55]; of the two classes on which *DEFA3* is commonly found, the Reference Sequence class is more common than Class 2 in the European population and has *DEFA3* at the centromeric-most repeat. However, in the vast majority of cases, EHF-PCR has shown that *DEFA3* only occupies the outer-most repeats of the array, which is in contrast to the observation by Aldred *et al.* that it can occupy all positions [55]. A study of haplotypes with a more variable *DEFA3* content would be required to further analyse this, as many of those studies using EHF-PCR had only a single copy of *DEFA3*. In addition, EHF-PCR has shown that the repeat units across a haplotype are often highly similar, which is unsurprising given that all five classes are associated with either a high or low frequency of both *DEFA3* and the Indel5 insertion. In some cases, this repeat unit similarity is also demonstrated by the *DEFA1A3* microsatellite alleles.

The similarity of features across a haplotype suggests that repeat units on a haplotype share high sequence similarity; this is likely to promote NAHR events. Combined with evidence that the *DEFA1A3* locus falls within a region of high LD, it suggests that the predominant mecha-

nism for rearrangements at the *DEFA1A3* locus is intra-allelic changes in *DEFA1A3* copy number. This means NAHR is more likely to occur between haplotypes from the same class. This is highlighted by Exchange 1, in which repeat units show high similarity at all variants studied, and this class has been observed with between 2-7 copies, despite being absent from Asian populations and extremely rare in African populations, suggesting it is relatively young. This also highlights the possibility of varying mutation rates between classes, where Exchange 1 haplotypes, with a variable number of highly similar repeats, likely experience deletion and duplication events more frequently than classes with haplotypes that display more variable structures. Inter-allelic (i.e. between class) rearrangements will also occur, but at a lesser frequency and, as the region displays high LD, it is likely that inter-allelic rearrangements result in gene conversion rather than crossover. This is supported by the identification of three gene conversion events at *DEFA1A3*- Exchange 1, Exchange 2 and the telomeric replacement polymorphism [38].

For each *DEFA1A3* haplotype class, there are one or two common, but related, structures. As the HapMap CEU1 samples are representative of the wider European population, in terms of their haplotype features (chapter 5), then it is likely that the common structures observed for each class are also applicable to the wider European population. As the five classes are tagged by a combination of four SNPs, these four SNP genotypes should be a sufficient indicator of *DEFA1A3* haplotype structure for haplotypes of European ancestry.

There are some limitations to the abilities of EHF-PCR for obtaining spatial information at *DEFA1A3*. As Sanger sequencing allows the detection of a maximum of four copies per haplotype, it limits the haplotype copy

number to which this is applicable to five copies, unless a haplotype contains only one or two variable alleles. This has led to some of the more unusual haplotypes being excluded- for example, a 7-copy Exchange 1 haplotype with an Indel5 ratio of 4:3. However, for each class, there are one or two common haplotypes, from which the majority of structures have been sampled. Those excluded tend to be rarer haplotypes. In addition, the differences in structures have led to differential representation of each class in the final dataset, with more Class 1 and Exchange 1, and less Reference Sequence and Class 2 haplotypes, relative to their frequencies in the CEU1 population. However, structures have been obtained for 64% of the 96 HapMap CEU1 haplotypes for which haplotype copy numbers are known and this is likely to provide a representative sample.

Initial analysis suggests that *DEFA1A3* haplotype class and, in turn, *DEFA1A3* haplotype structure, does not affect *DEFA1A3* expression. There are many possible reasons for this. Firstly, the sample size was small and therefore, very few homozygotes were observed for each class, such that any possible effects may be masked when the haplotype is observed in heterozygous form. It is also possible that steps in the preparation of the neutrophil cells meant the HNP1-3 level measured was not representative of neutrophil cells *in vivo*. However, measuring the peptide content of neutrophils is more suitable than measuring the *DEFA1A3* mRNA content of neutrophils, as the mRNA expression takes place in promyelocytes and, as such, will have been degraded once the cell becomes a mature neutrophil. In addition, very little is known about the mechanism of *DEFA1A3* expression. It is possible that not all copies of *DEFA1A3* are expressed and expression may be restricted to the first repeat unit on each haplotype, for example. Analysis of the expression of

HNP-1 and HNP-3 separately could provide a useful method by which to assess this, as it could allow us to determine whether *DEFA3* is always expressed, regardless of its position in the repeat array. An initial estimate suggests that humans, on average, produce 150-250mg of HNP1-3 per day [152–154] (Danielle Carpenter, personal communication); this is a significant contribution to an immune cell peptide and suggests that HNP1-3 must perform an essential function. However, investigations into the relationship between *DEFA1A3* copy number and HNP1-3 expression are still in the early stages. Additional analysis of the peptide composition, combined with sequence information across promoter regions of the *DEFA1A3* locus, will provide a better understanding of HNP1-3 expression. This will be aided by the ability of four SNPs to tag the structure of haplotypes of European ancestry.

7 Discussion

7.1 Haplotype phasing and structures of the *DEFA1A3* locus

In order to truly understand the nature of multiallelic CNV loci, it is necessary to combine sequence and positional information. This allows an understanding of how haplotypes are related to each other, the mechanisms by which variation is generated and how copy number variation is related to the expression of CNV genes. At the *DEFA1A3* locus, detailed haplotype-specific information was available for haplotypes at the *DEFA1A3* locus [50]; however, from this information alone, it was unclear what features were shared between related haplotypes.

Through the use of flanking sequence variation, it is clear that there are five common classes of haplotype at *DEFA1A3*. These classes were initially identified in haplotypes of European ancestry, but the conserved LD between the SNPs tagging these five classes has allowed the identification of the same classes in other worldwide populations. Therefore, the identification of common haplotype classes has clarified which haplotypes are most closely related.

For many *DEFA1A3* haplotypes of European ancestry, we know not only their copy number, but the ratio of *DEFA1 vs. DEFA3*, the ratio of alleles for two variables Indels, Indel5 and 7bp duplication, and now also the allele sizes of a highly variable microsatellite. In haplotypes of European ancestry, this has allowed the identification of many signif-

icant associations between *DEFA1A3* haplotype class and features of the *DEFA1A3* locus, demonstrating that related haplotypes share common features across their *DEFA1A3* haplotype. However, the associations between *DEFA1A3* haplotype class and *DEFA1A3* copy number differ between populations. This accounts for the association between the SNP rs4300027 genotype and *DEFA1A3* copy number in populations of European ancestry, but not in other populations. Therefore, this indicates that each *DEFA1A3* haplotype class will be associated with different features in different populations. As such, further studies at the *DEFA1A3* locus should use a population-specific approach.

Structural information at the *DEFA1A3* locus has been gained through the use of EHF-PCR. The methods described by Tyson and Armour [131] have been developed to include additional variant positions across a *DEFA1A3* haplotype, providing comprehensive spatial information for haplotypes of European ancestry. It has identified that there are one or two common structures associated with each haplotype class. Therefore, *DEFA1A3* alleles have been identified- haplotypes which share the same flanking sequence (i.e. class) and display the same spatial arrangement of variants across the *DEFA1A3* CNV region. Each *DEFA1A3* haplotype class appears to include one or two common alleles. This is an important discovery, as it clarifies how haplotypes are related to each other. The definition of alleles in the European population, which are tagged by four SNPs, aids future studies on how *DEFA1A3* copy number influences HNP1-3 expression.

Whilst the knowledge of haplotype structures at the *DEFA1A3* locus has been advanced, there are still unanswered questions. Firstly, it is clear that the *DEFA1A3* haplotype copy numbers associated with each class

differs between populations and it is possible that the associations with other features of the locus (e.g. *DEFA3* frequency) differ also. This would in turn lead to structural differences. In populations of non-European ancestry, very little haplotype-specific information is available, due to the lack of a three-generation DNA resource. However, the *DEFA1A3* microsatellite, in combination with EHF-PCR, could provide *DEFA1A3* haplotype copy numbers in a mother-father-child trio. In order to truly understand the between-population differences at the *DEFA1A3* locus, a more detailed analysis is required in other populations. In addition, the structural haplotypes that have been obtained are limited to the position of a few specific variants, which are being used as proxies to indicate the sequence nature of the repeat unit. Ideally, a single sequence read from one end of the *DEFA1A3* locus to the other would be generated, providing absolute sequence and structural information. However, this is not feasible given current sequencing technologies. Despite this, there have been recent advances in single molecule sequencing technologies, such as Pacific Biosciences and Oxford Nanopore technology [132, 133, 155], which may provide sufficiently long reads to allow complete phasing across a locus as complex as *DEFA1A3*.

7.2 Evolution at the *DEFA1A3* locus

The generation of structural information at *DEFA1A3* has shown that related haplotypes share not only similar structures, but similar repeat units. Combined with the information that *DEFA1A3* falls within a region of high LD, this has identified intra-allelic rearrangements as the predominant mechanism for copy number change at *DEFA1A3*. This will be combined with between-class gene conversion and crossover events;

three gene conversion events at *DEFA1A3* have already been identified. Gene conversion will act to homogenise the repeat units and has previously been observed at other variable number tandem repeat loci [156–160], which are similar to *DEFA1A3* in that they contain a variable number of copies of a highly similar sequence arranged in tandem. This homogenisation will inevitably allow further rearrangements to occur, by preserving the sequence similarity across the locus. Therefore, given that the *DEFA1A3* locus appears to be very dynamic, it seems unwise to give fixed boundaries to repeat units at the *DEFA1A3* locus, as, given its high sequence similarity, breakpoints are likely to be variable. It is more sensible to say it has a 19kb repeat unit, with unfixed boundaries. However, the use of single molecule sequencing technologies would clarify this, as it would identify possible breakpoints by comparing the exact sequence of closely related haplotypes.

A confirmation of intra-allelic rearrangements as the proposed mechanism for copy number change at *DEFA1A3* would come from looking at different populations. The structural data we have is for haplotypes of European ancestry and the landscape may be very different in other populations. In addition, the use of flanking sequence data from individuals with non-European ancestry may help to clarify the evolutionary relationship between the different classes of *DEFA1A3* haplotypes. It is hard to reconstruct a phylogenetic tree of *DEFA1A3* haplotype classes, due to the presence of gene conversion events and the lack of intermediate haplotypes.

The *DEFA1A3* locus does not show evidence of selection on any of the *DEFA1A3* haplotype classes. Given that the *DEFA1A3* haplotype classes are conserved worldwide, it would not be expected that these

haplotypes display extensive EHH, as the variants tagging these classes are relatively old and, as such, new mutations will have arisen on each class background. Exchange 1 is the only variant which possibly displays evidence of selection, but only in the HapMap CEU1 population and, even then, is only weakly significant. It is possible that it is the variability of the locus, and not the presence of a single allele, that is advantageous at this locus. Given it is an immune response gene, it could be that varying challenges faced by the immune response have led to its highly variable status. Alternatively, it could be that an advantageous allele is found on two different *DEFA1A3* class backgrounds or is only found on a subset of haplotypes within a *DEFA1A3* class. In either of these scenarios, the method used would have failed to detect selection and a more detailed analysis would be required to identify this.

7.3 Relating *DEFA1A3* copy number to HNP1-3 expression

One aspect still to be addressed is the relationship between *DEFA1A3* copy number and HNP1-3 expression and, ultimately, the functional effect of this. It appears that there is no linear association between *DEFA1A3* copy number and HNP1-3 expression (Danielle Carpenter, personal communication). Therefore, a comparison between HNP1-3 expression and *DEFA1A3* haplotype structure was made. This initial study also failed to identify an association.

However, the difficulty in interpreting these results is the lack of understanding of the regulation of expression at *DEFA1A3*. It may be possible that not all copies of *DEFA1A3* are expressed; for example, whilst it is assumed each *DEFA1A3* gene has an active promoter element upstream,

it may be possible that enhancer sequences, located at a distance from the *DEFA1A3* repeat array, may only act upon the first copy on each haplotype. In addition, very little is known about how promoter region variants influence expression. This region is known to be highly variable (Omniah Mansouri, personal communication), but the consequences of this have yet to be determined.

In terms of understanding the regulation of expression at *DEFA1A3*, it would be interesting to determine whether the ratio of *DEFA1* to *DEFA3* is maintained. This could be achieved using mass spectroscopy [161, 162], as ELISA methods previously used to study HNP1-3 expression are unable to separately quantify HNP-1, HNP-2 and HNP-3. As *DEFA3* is much less common than *DEFA1* and its position differs between haplotype classes (telomeric-most repeat on Class 2 haplotypes and centromeric-most repeat on Reference Sequence haplotypes), it would be possible to determine whether *DEFA3* is expressed regardless of position. This would go some way to understanding whether all copies of the *DEFA1A3* repeat array are expressed.

A daily HNP1-3 production by the body of 150-250mg, combined with a lack of coding variants in the *DEFA1* and *DEFA3* genes (Omniah Mansouri, personal communication), suggests that HNP1-3 must form an essential part of the immune response. In addition, the association of the *DEFA1A3* locus with IgA nephropathy in the Han Chinese population suggests that variation at the *DEFA1A3* locus is functionally important. Whilst the basis of this associations is not yet understood, it is clear that spatial information at the *DEFA1A3* locus will be important in future interpretation of results at *DEFA1A3*, aided by the identification of alleles which are tagged by flanking SNPs. In summary, knowledge of the struc-

ture of CNVRs and the differences between populations is essential for an accurate understanding of and interpretation of data for complex loci like *DEFA1A3*.

References

[1] 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.

[2] ENCODE Project Consortium, Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

[3] Shastry, B. S. (2002). SNP alleles in human disease and evolution. J Hum Genet *47*, 561–566.

[4] Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. N Engl J Med *363*, 166–176.

[5] Hindorff, L., MacArthur, J., Morales, J., Junkins, H., Hall, P., Klemm, A., and Manolio, T. (2014). A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies.

[6] Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet *14*, 125–138.

[7] Girirajan, S., Campbell, C. D., and Eichler, E. E. (2011). Human copy number variation and complex genetic disease. Annu Rev Genet *45*, 203–226.

[8] Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet *10*, 451–481.

[9] Stankiewicz, P. and Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. Annu Rev Med *61*, 437–455.

[10] International HapMap Consortium. (2003). The International HapMap Project. Nature *426*, 789–796.

[11] Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res *29*, 308–311.

[12] Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature *456*, 18–21.

[13] Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. Nat Rev Genet *7*, 85–97.

[14] Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res *16*, 1182–1190.

[15] Fan, H. and Chu, J.-Y. (2007). A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics *5*, 7–14.

[16] Li, Y.-C., Korol, A. B., Fahima, T., Beiles, A., and Nevo, E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol *11*, 2453–2465.

[17] Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet *13*, 36–46.

[18] International HapMap Consortium. (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

[19] 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

[20] 1000 Genomes Project Consortium and Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

[21] Fanciulli, M., Petretto, E., and Aitman, T. J. (2010). Gene copy number variation and common human disease. Clin Genet *77*, 201–213.

[22] Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler Smith, C., Hurles, M. E., et al. (2006). Copy number variation: new insights in genome diversity. Genome Res *16*, 949–961.

[23] Macdonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., and Scherer, S. W. (2014). The Database of Genomic Variants: a curated

collection of structural variation in the human genome. Nucleic Acids Res *42*, D986–D992.

[24] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

[25] Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira, G. (2009). Mechanisms of change in gene copy number. Nat Rev Genet *10*, 551–564.

[26] Lupski, J. R. (2007). Genomic rearrangements and sporadic disease. Nat Genet *39*, S43–S47.

[27] Kondrashov, A. S. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat *21*, 12–27.

[28] Nachman, M. W. and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics *156*, 297–304.

[29] Beckmann, J. S., Estivill, X., and Antonarakis, S. E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet *8*, 639–646.

[30] Olsson, L. M. and Holmdahl, R. (2012). Copy number variation in autoimmunity–importance hidden in complexity? Eur J Immunol *42*, 1969–1976.

[31] Woodwark, C. and Bateman, A. (2011). The characterisation of three types of genes that overlie copy number variable regions. PLoS One *6*, e14814.

[32] Chaignat, E., Yahya Graison, E. A., Henrichsen, C. N., Chrast, J., Schtz, F., Pradervand, S., and Reymond, A. (2011). Copy number variation modifies expression time courses. Genome Res *21*, 106–113.

[33] Gu, W., Zhang, F., and Lupski, J. R. (2008). Mechanisms for human genomic rearrangements. Pathogenetics *1*, 4.

[34] Liu, P., Carvalho, C. M. B., Hastings, P. J., and Lupski, J. R. (2012). Mechanisms for recurrent and complex human genomic rearrangements. Curr Opin Genet Dev *22*, 211–220.

[35] Fawcett, J. A. and Innan, H. (2013). The role of gene conversion in preserving rearrangement hotspots in the human genome. Trends Genet *29*, 561–568.

[36] Chen, J.-M., Cooper, D. N., Chuzhanova, N., Frec, C., and Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet *8*, 762–775.

[37] Mueller, M., Barros, P., Witherden, A. S., Roberts, A. L., Zhang, Z., Schaschl, H., Yu, C.-Y., Hurles, M. E., Schaffner, C., Floto, R. A., et al. (2013). Genomic pathology of SLE-associated copy-number variation at the FCGR2C/FCGR3B/FCGR2B locus. Am J Hum Genet *92*, 28–40.

[38] Black, H. A., Khan, F. F., Tyson, J., and Armour. (2014). Inferring mechanisms of copy number change from haplotype structure at the human DEFA1A3 locus. Submitted for publication.

[39] Lee, J. A., Carvalho, C. M. B., and Lupski, J. R. (2007). A DNA replication mechanism for generating nonrecurrent rear-

rangements associated with genomic disorders. Cell *131*, 1235–1247.

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

[41] Armour, J. A., Sismani, C., Patsalis, P. C., and Cross, G. (2000). Measurement of locus copy number by hybridisation with amplifiable probes. Nucleic Acids Res *28*, 605–609.

[42] Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F., and Pals, G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res *30*, e57.

[43] Tyson, J., Majerus, T. M. O., Walker, S., and Armour, J. A. L. (2010). Screening for common copy-number variants in cancer genes. Cancer Genet Cytogenet *203*, 316–323.

[44] den Dunnen, J. T. and White, S. J. (2006). MLPA and MAPH: sensitive detection of deletions and duplications. Curr Protoc Hum Genet *Chapter 7*, Unit 7.14.

[45] Aldhous, M. C., Abu Bakar, S., Prescott, N. J., Palla, R., Soo, K., Mansfield, J. C., Mathew, C. G., Satsangi, J., and Armour, J. A. L. (2010). Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease. Hum Mol Genet *19*, 4930–4938.

[46] Cantsilieris, S., Baird, P. N., and White, S. J. (2012). Molecular methods for genotyping complex copy number polymorphisms. Genomics *101*, 86–93.

[47] Cantsilieris, S., Western, P. S., Baird, P. N., and White, S. J. (2014). Technical considerations for genotyping multi-allelic copy number variation (CNV), in regions of segmental duplication. BMC Genomics *15*, 329.

[48] Armour, J. A. L., Palla, R., Zeeuwen, P. L. J. M., den Heijer, M., Schalkwijk, J., and Hollox, E. J. (2007). Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. Nucleic Acids Res *35*, e19.

[49] Walker, S., Janyakhantikul, S., and Armour, J. A. L. (2009). Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs. Genomics *93*, 98–103.

[50] Khan, F. F., Carpenter, D., Mitchell, L., Mansouri, O., Black, H. A., Tyson, J., and Armour, J. A. (2013). Accurate measurement of gene copy number for human alpha-defensin DEFA1A3. BMC Genomics *14*, 719.

[51] Hollox, E. J., Detering, J.-C., and Dehnugara, T. (2009). An integrated approach for measuring copy number variation at the FCGR3 (CD16) locus. Hum Mutat *30*, 477–484.

[52] Carpenter, D., Dhar, S., Mitchell, L. M., Fu, B., Tyson, J., Shwan, N., Yang, F., Thomas, M., and Armour, J. A. L. (2014). Odd and even: global variation in structure of the human amylase gene cluster. Submitted for publication.

[53] Le Scouarnec, S. and Gribble, S. M. (2012). Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics. Heredity *108*, 75–85.

[54] Perry, G. H., Tchinda, J., McGrath, S. D., Zhang, J., Picker, S. R., Cceres, A. M., Iafrate, A. J., Tyler Smith, C., Scherer, S. W., Eichler, E. E., et al. (2006). Hotspots for copy number variation in chimpanzees and humans. Proc Natl Acad Sci U S A *103*, 8006–8011.

[55] Aldred, P. M. R., Hollox, E. J., and Armour, J. A. L. (2005). Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3. Hum Mol Genet *14*, 2045–2052.

[56] Abu Bakar, S., Hollox, E. J., and Armour, J. A. L. (2009). Allelic recombination between distinct genomic locations generates copy number diversity in human beta-defensins. Proc Natl Acad Sci U S A *106*, 853–858.

[57] Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science *315*, 848–853.

[58] Tang, Y.-C. and Amon, A. (2013). Gene copy-number alterations: a cost-benefit analysis. Cell *152*, 394–405.

[59] Lupski, J. R., Wise, C. A., Kuwano, A., Pentao, L., Parke, J. T., Glaze, D. G., Ledbetter, D. H., Greenberg, F., and Patel, P. I. (1992). Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A. Nat Genet *1*, 29–33.

[60] Vittori, A., Breda, C., Repici, M., Orth, M., Roos, R. A. C., Outeiro, T. F., Giorgini, F., Hollox, E. J., and the R. E. G. I. S. T. R. Y investigators of the European Huntington's Disease Network.

(2014). Copy-number variation of the neuronal glucose transporter gene SLC2A3 and age of onset in Huntington's disease. Hum Mol Genet. Advanced Online Publication: doi: 10.1093/hmg/ddu022.

[61] Willcocks, L. C., Lyons, P. A., Clatworthy, M. R., Robinson, J. I., Yang, W., Newland, S. A., Plagnol, V., McGovern, N. N., Condliffe, A. M., Chilvers, E. R., et al. (2008). Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake. J Exp Med *205*, 1573–1582.

[62] Mueller, M., Barros, P., Witherden, A. S., Roberts, A. L., Zhang, Z., Schaschl, H., Yu, C.-Y., Hurles, M. E., Schaffner, C., Floto, R. A., et al. (2013). Genomic pathology of SLE-associated copy-number variation at the FCGR2C/FCGR3B/FCGR2B locus. Am J Hum Genet *92*, 28–40.

[63] Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J., et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science *307*, 1434–1440.

[64] McKinney, C., Merriman, M. E., Chapman, P. T., Gow, P. J., Harrison, A. A., Highton, J., Jones, P. B. B., McLean, L., O'Donnell, J. L., Pokorny, V., et al. (2008). Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. Ann Rheum Dis *67*, 409–413.

[65] Field, S. F., Howson, J. M. M., Maier, L. M., Walker, S., Walker, N. M., Smyth, D. J., Armour, J. A. L., Clayton, D. G., and Todd,

J. A. (2009). Experimental aspects of copy number variant assays at CCL3L1. Nat Med *15*, 1115–1117.

[66] Aklillu, E., Odenthal Hesse, L., Bowdrey, J., Habtewold, A., Ngaimisi, E., Yimer, G., Amogne, W., Mugusi, S., Minzi, O., Makonnen, E., et al. (2013). CCL3L1 copy number, HIV load, and immune reconstitution in sub-Saharan Africans. BMC Infect Dis *13*, 536.

[67] Carpenter, D., Walker, S., Prescott, N., Schalkwijk, J., and Armour, J. A. (2011). Accuracy and differential bias in copy number measurement of CCL3L1 in association studies with three autoimmune disorders. BMC Genomics *12*, 418.

[68] Hollox, E. J., Hüffmeier, U., Zeeuwen, P. L. J. M., Palla, R., Lascorz, J., Rodijk Olthuis, D., van de Kerkhof, P. C. M., Traupe, H., de Jongh, G., den Heijer, M., et al. (2008). Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet *40*, 23–25.

[69] Stuart, P. E., Hüffmeier, U., Nair, R. P., Palla, R., Tejasvi, T., Schalkwijk, J., Elder, J. T., Reis, A., and Armour, J. A. L. (2012). Association of β-defensin copy number and psoriasis in three cohorts of European origin. J Invest Dermatol *132*, 2407–2413.

[70] Bentley, R. W., Pearson, J., Gearry, R. B., Barclay, M. L., McKinney, C., Merriman, T. R., and Roberts, R. L. (2010). Association of higher DEFB4 genomic copy number with Crohn's disease. Am J Gastroenterol *105*, 354–359.

[71] Fellermann, K., Stange, D. E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C. L., Reinisch, W., Teml, A., Schwab, M.,

Lichter, P., et al. (2006). A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am J Hum Genet *79*, 439–448.

[72] Wellcome Trust Case Control Consortium, Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D. F., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature *464*, 713–720.

[73] 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.

[74] Ganz, T. (2003). Defensins: antimicrobial peptides of innate immunity. Nat Rev Immunol *3*, 710–720.

[75] Chen, H., Xu, Z., Peng, L., Fang, X., Yin, X., Xu, N., and Cen, P. (2006). Recent advances in the research and development of human defensins. Peptides *27*, 931–940.

[76] Patil, A., Hughes, A. L., and Zhang, G. (2004). Rapid evolution and diversification of mammalian alpha-defensins as revealed by comparative analysis of rodent and primate genes. Physiol Genomics *20*, 1–11.

[77] Lynn, D. J. and Bradley, D. G. (2007). Discovery of alpha-defensins in basal mammals. Dev Comp Immunol *31*, 963–967.

[78] Ballana, E., Gonzlez, J. R., Bosch, N., and Estivill, X. (2007). Inter-population variability of DEFA3 gene absence: correlation

with haplotype structure and population variability. BMC Genomics *8*, 14.

[79] Nguyen, T. X., Cole, A. M., and Lehrer, R. I. (2003). Evolution of primate theta-defensins: a serpentine path to a sweet tooth. Peptides *24*, 1647–1654.

[80] Li, D., Zhang, L., Yin, H., Xu, H., Trask, J. S., Smith, D. G., Li, Y., Yang, M., and Zhu, Q. (2014). Evolution of primate α and θ defensins revealed by analysis of genomes. Mol Biol Rep.

[81] Das, S., Nikolaidis, N., Goto, H., McCallister, C., Li, J., Hirano, M., and Cooper, M. D. (2010). Comparative genomics and evolution of the alpha-defensin multigene family in primates. Mol Biol Evol *27*, 2333–2343.

[82] Lynn, D. J., Lloyd, A. T., Fares, M. A., and O'Farrelly, C. (2004). Evidence of positively selected sites in mammalian alpha-defensins. Mol Biol Evol *21*, 819–827.

[83] Linzmeier, R., Ho, C. H., Hoang, B. V., and Ganz, T. (1999). A 450-kb contig of defensin genes on human chromosome 8p23. Gene *233*, 205–211.

[84] Mars, W. M., Patmasiriwat, P., Maity, T., Huff, V., Weil, M. M., and Saunders, G. F. (1995). Inheritance of unequal numbers of the genes encoding the human neutrophil defensins HP-1 and HP-3. J Biol Chem *270*, 30371–30376.

[85] Linzmeier, R. M. and Ganz, T. (2005). Human defensin gene copy number polymorphisms: comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23. Genomics *86*, 423–430.

[86] Chertov, O., Yang, D., Howard, O. M., and Oppenheim, J. J. (2000). Leukocyte granule proteins mobilize innate host defenses and adaptive immune responses. Immunol Rev *177*, 68–78.

[87] Jones, D. E. and Bevins, C. L. (1992). Paneth cells of the human small intestine express an antimicrobial peptide gene. J Biol Chem *267*, 23216–23225.

[88] Porter, E. M., Liu, L., Oren, A., Anton, P. A., and Ganz, T. (1997). Localization of human intestinal defensin 5 in Paneth cell granules. Infect Immun *65*, 2389–2395.

[89] Ganz, T., Selsted, M. E., Szklarek, D., Harwig, S. S., Daher, K., Bainton, D. F., and Lehrer, R. I. (1985). Defensins. Natural peptide antibiotics of human neutrophils. J Clin Invest *76*, 1427–1435.

[90] Valore, E. V. and Ganz, T. (1992). Posttranslational processing of defensins in immature human myeloid cells. Blood *79*, 1538–1544.

[91] Linzmeier, R., Michaelson, D., Liu, L., and Ganz, T. (1993). The structure of neutrophil defensin genes. FEBS Lett *321*, 267–273.

[92] Selsted, M. E., Harwig, S. S., Ganz, T., Schilling, J. W., and Lehrer, R. I. (1985). Primary structures of three human neutrophil defensins. J Clin Invest *76*, 1436–1439.

[93] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996–1006.

[94] Valore, E. V., Martin, E., Harwig, S. S., and Ganz, T. (1996). Intramolecular inhibition of human defensin HNP-1 by its propiece. J Clin Invest *97*, 1624–1629.

[95] Lehrer, R. I., Barton, A., Daher, K. A., Harwig, S. S., Ganz, T., and Selsted, M. E. (1989). Interaction of human defensins with Escherichia coli. Mechanism of bactericidal activity. J Clin Invest *84*, 553–561.

[96] Ganz, T., Selsted, M. E., and Lehrer, R. I. (1990). Defensins. Eur J Haematol *44*, 1–8.

[97] Schneider, J. J., Unholzer, A., Schaller, M., Schfer Korting, M., and Korting, H. C. (2005). Human defensins. J Mol Med (Berl) *83*, 587–595.

[98] Daher, K. A., Selsted, M. E., and Lehrer, R. I. (1986). Direct inactivation of viruses by human granulocyte defensins. J Virol *60*, 1068–1074.

[99] Ericksen, B., Wu, Z., Lu, W., and Lehrer, R. I. (2005). Antibacterial activity and specificity of the six human alpha-defensins. Antimicrob Agents Chemother *49*, 269–275.

[100] Yang, D., Biragyn, A., Kwak, L. W., and Oppenheim, J. J. (2002). Mammalian defensins in immunity: more than just microbicidal. Trends Immunol *23*, 291–296.

[101] Grigat, J., Soruri, A., Forssmann, U., Riggert, J., and Zwirner, J. (2007). Chemoattraction of macrophages, T lymphocytes, and mast cells is evolutionarily conserved within the human alpha-defensin family. J Immunol *179*, 3958–3965.

[102] Yang, D., Chen, Q., Chertov, O., and Oppenheim, J. J. (2000). Human neutrophil defensins selectively chemoattract naive T and immature dendritic cells. J Leukoc Biol *68*, 9–14.

[103] Lillard, J., Jr, Boyaka, P. N., Chertov, O., Oppenheim, J. J., and McGhee, J. R. (1999). Mechanisms for induction of acquired host immunity by neutrophil peptide defensins. Proc Natl Acad Sci U S A *96*, 651–656.

[104] Selsted, M. E. and Ouellette, A. J. (2005). Mammalian defensins in the antimicrobial immune response. Nat Immunol *6*, 551–557.

[105] Hazlett, L. and Wu, M. (2011). Defensins in innate immunity. Cell Tissue Res *343*, 175–188.

[106] 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.

[107] Chen, Q., Hakimi, M., Wu, S., Jin, Y., Cheng, B., Wang, H., Xie, G., Ganz, T., Linzmeier, R. M., and Fang, X. (2010). Increased genomic copy number of DEFA1/DEFA3 is associated with susceptibility to severe sepsis in Chinese Han population. Anesthesiology *112*, 1428–1434.

[108] Cheng, F.-J., Zhou, X.-J., Zhao, Y.-F., Zhao, M.-H., and Zhang, H. (2013). Alpha-defensin DEFA1A3 gene copy number variation in Asians and its genetic association study in Chinese systemic lupus erythematosus patients. Gene *517*, 158–163.

[109] Nuytten, H., Wlodarska, I., Nackaerts, K., Vermeire, S., Vermeesch, J., Cassiman, J.-J., and Cuppens, H. (2009). Accurate determination of copy number variations (CNVs): application to the alpha- and beta-defensin CNVs. J Immunol Methods *344*, 35–44.

[110] 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., et al. (2012). A genome-wide association study in Han Chinese identifies multiple susceptibility loci for IgA nephropathy. Nat Genet *44*, 178–182.

[111] Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J., and Schork, N. J. (2011). The importance of phase information for human genomics. Nat Rev Genet *12*, 215–223.

[112] Petersdorf, E. W., Malkki, M., Gooley, T. A., Martin, P. J., and Guo, Z. (2007). MHC haplotype matching for unrelated hematopoietic cell transplantation. PLoS Med *4*, e8.

[113] Lemmers, R. J. L. F., Wohlgemuth, M., van der Gaag, K. J., van der Vliet, P. J., van Teijlingen, C. M. M., de Knijff, P., Padberg, G. W., Frants, R. R., and van der Maarel, S. M. (2007). Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. Am J Hum Genet *81*, 884–894.

[114] Lupski, J. R., Reid, J. G., Gonzaga Jauregui, C., Rio Deiros, D., Chen, D. C. Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D. A., et al. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med *362*, 1181–1191.

[115] Simunovic, M. P. (2010). Colour vision deficiency. Eye *24*, 747–755.

[116] International HapMap Consortium. (2003). The International HapMap Project. Nature *426*, 789–796.

[117] International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

[118] Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., and Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet *38*, 1251–1260.

[119] Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.

[120] Yang, H., Chen, X., and Wong, W. H. (2011). Completely phased genome sequencing through chromosome sorting. Proc Natl Acad Sci U S A *108*, 12–17.

[121] Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K., and Song, Q. (2010). Direct determination of molecular haplotypes by chromosome microdissection. Nat Methods *7*, 299–301.

[122] Browning, S. R. and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. Nat Rev Genet *12*, 703–714.

[123] Duitama, J., McEwen, G. K., Huebsch, T., Palczewski, S., Schulz, S., Verstrepen, K., Suk, E.-K., and Hoehe, M. R. (2012). Fosmid-based whole genome haplotyping of a HapMap trio child: evalu-

ation of Single Individual Haplotyping techniques. Nucleic Acids Res *40*, 2041–2053.

[124] Suk, E.-K., McEwen, G. K., Duitama, J., Nowick, K., Schulz, S., Palczewski, S., Schreiber, S., Holloway, D. T., McLaughlin, S., Peckham, H., et al. (2011). A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res *21*, 1672–1685.

[125] Kirkness, E. F., Grindberg, R. V., Yee Greenbaum, J., Marshall, C. R., Scherer, S. W., Lasken, R. S., and Venter, J. C. (2013). Sequencing of isolated sperm cells for direct haplotyping of a human genome. Genome Res *23*, 826–832.

[126] Selvaraj, S., R Dixon, J., Bansal, V., and Ren, B. (2013). Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat Biotechnol *31*, 1111–1118.

[127] Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y. T., Haas, J., et al. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature *487*, 190–195.

[128] Kaper, F., Swamy, S., Klotzle, B., Munchel, S., Cottrell, J., Bibikova, M., Chuang, H.-Y., Kruglyak, S., Ronaghi, M., Eberle, M. A., et al. (2013). Whole-genome haplotyping by dilution, amplification, and sequencing. Proc Natl Acad Sci U S A *110*, 5552–5557.

[129] Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M., and Snyder, M. (2014). Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol.

[130] Michalatos Beloin, S., Tishkoff, S. A., Bentley, K. L., Kidd, K. K., and Ruano, G. (1996). Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res *24*, 4841–4843.

[131] Tyson, J. and Armour, J. A. L. (2012). Determination of haplotypes at structurally complex regions using emulsion haplotype fusion PCR. BMC Genomics *13*, 693.

[132] Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., et al. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. Advanced Online Publication: doi:10.1101/gr.168450.113.

[133] Pacific Biosciences (2013). PacBio RSII Sequencing System. http://files.pacb.com/pdf/PacBio_RS_II_Brochure.pdf.

[134] Crawford, D. C. and Nickerson, D. A. (2005). Definition and clinical importance of haplotypes. Annu Rev Med *56*, 303–320.

[135] Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J. M., Gough, S. C. L., de Smith, A., Blakemore, A. I. F., et al. (2007). FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet *39*, 721–723.

[136] Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., and Rozen, S. G. (2012). Primer3–new capabilities and interfaces. Nucleic Acids Res *40*, e115.

[137] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403–410.

[138] Hall, T. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium *41*, 95–98.

[139] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res *31*, 3497–3500.

[140] Lander, E. S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A *84*, 2363–2367.

[141] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and , . G. P. D. P. S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

[142] Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

[143] 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.

[144] Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. PLoS Biol *4*, e72.

[145] Gautier, M. and Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics *28*, 1176–1177.

[146] Turner, D. J. and Hurles, M. E. (2009). High-throughput haplotype determination over long distances by haplotype fusion PCR and ligation haplotyping. Nat Protoc *4*, 1771–1783.

[147] Wetmur, J. G., Kumar, M., Zhang, L., Palomeque, C., Wallenstein, S., and Chen, J. (2005). Molecular haplotyping by linking emulsion PCR: analysis of paraoxonase 1 haplotypes and phenotypes. Nucleic Acids Res *33*, 2615–2619.

[148] Wetmur, J. G. and Chen, J. (2011). Linking emulsion PCR haplotype analysis. Methods Mol Biol *687*, 165–175.

[149] IBM Corporation (2012). IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM.

[150] Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. Theor Popul Biol *3*, 87–112.

[151] Kimmel, M. and Chakraborty, R. (1996). Measures of variation at DNA repeat loci under a general stepwise mutation model. Theor Popul Biol *50*, 345–367.

[152] Dancey, J. T., Deubelbeiss, K. A., Harker, L. A., and Finch, C. A. (1976). Neutrophil kinetics in man. J Clin Invest *58*, 705–715.

[153] von Vietinghoff, S. and Ley, K. (2008). Homeostatic regulation of blood neutrophil counts. J Immunol *181*, 5183–5188.

[154] Ihi, T., Nakazato, M., Mukae, H., and Matsukura, S. (1997). Elevated concentrations of human neutrophil peptides in plasma, blood, and body fluids from patients with infections. Clin Infect Dis *25*, 1134–1140.

[155] Oxford Nanopore Technologies (2014). Oxford Nanopore Technologies. https://www.nanoporetech.com/.

[156] Jeffreys, A. J., Tamaki, K., MacLeod, A., Monckton, D. G., Neil, D. L., and Armour, J. A. (1994). Complex gene conversion events in germline mutation at human minisatellites. Nat Genet *6*, 136–145.

[157] Buard, J., Bourdet, A., Yardley, J., Dubrova, Y., and Jeffreys, A. J. (1998). Influences of array size and homogeneity on minisatellite mutation. EMBO J *17*, 3495–3502.

[158] Buard, J., Shone, A. C., and Jeffreys, A. J. (2000). Meiotic recombination and flanking marker exchange at the highly unstable human minisatellite CEB1 (D2S90). Am J Hum Genet *67*, 333–344.

[159] Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., Kallicki, J., Kaul, R., Wilson, R. K., and Eichler, E. E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell *143*, 837–847.

[160] Nuttle, X., Huddleston, J., O'Roak, B. J., Antonacci, F., Fichera, M., Romano, C., Shendure, J., and Eichler, E. E. (2013). Rapid

and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions. Nat Methods *10*, 903–909.

[161] Lundy, F. T., Orr, D. F., Shaw, C., Lamey, P.-J., and Linden, G. J. (2005). Detection of individual human neutrophil alpha-defensins (human neutrophil peptides 1, 2 and 3) in unfractionated gingival crevicular fluid–a MALDI-MS approach. Mol Immunol *42*, 575–579.

[162] Thompson, L., Turko, I., and Murad, F. (2006). Mass spectrometry-based relative quantification of human neutrophil peptides 1, 2, and 3 from biological samples. Mol Immunol *43*, 1485–1489.