GENERATION OF DIVERSITY AT THE HUMAN BETA-DEFENSIN COPY NUMBER

Suhaili Abu Bakar

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

June 2010

ABSTRACT

Submicroscopic structural genomic variation includes copy number variation (CNV) that can result changes in DNA dosage, and the impacts can be observed on common disease, metabolism, and heritable traits such as colour vision and rhesus blood group. Human beta-defensins form a cluster of at least seven genes on human chromosome 8p23.1, with a diploid copy number commonly ranging between 2 and 7 copies. They encode small secreted antimicrobial peptides with cytokine-like properties which are found expressed at high levels in psoriasis patients, and copy number at this locus has been found to be associated with inflammatory bowel disease, particularly colonic Crohn's disease.

The focus of this thesis has been divided into two studies; looking for the origin of diversity at the human beta-defensins copy number, and development of a multiplex PRT measurement system for accurately typing the beta-defensin region in large case association study.

The origin of diversity at the human beta-defensin copy number has been followed by using segregation in CEPH families. Three out of 416 meiotic transmissions changed the copy number by simple allelic recombination between two distinct loci for these genes. Deducing haplotype copy number from microsatellite and multiallelic length polymorphism have allowed this study to map the beta-defensin repeats in two locations at the original location distally REPD and about 5 Mb away at proximally REPP. We have demonstrated our multiplex PRT system is a powerful technique to determine the association of the beta-defensin genes in Crohn's disease even though we did not produce any convincing support for associations reported from previous studies.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my supervisor, Professor John Armour for all his guidance and support throughout this demanding research process. I would also like to thank Professor John Brookfield as my internal advisor who has given me great support throughout my study. Raquel Palla who is the only and the best mate in the beta-defensins group; thank you for sharing together all the human beta-defensin work. All the members of C10 laboratory at the Institute of Genetics, Jess Tyson, Tamsin Majerus, Dannie Carpenter, Fayeza Khan, Somwang Jayakhantikul, Sughanda Dhar, Ladas Ionnis and the former PhD and Mres student, Susan Walker and Emma Dannhausser have also helped and kept me relatively sane over the past four years They also have made coming to work most enjoyable.

I would also like to say thank you to the Academic Unit of Institute of Genetics and School of Biology, both past and present, for all their technical and much needed moral support. I would also like to thank friends outside the department especially for my Malaysian friends in Nottingham for flourishing my life; our friendship will last forever. Thank you to my employee, the Universiti Putra Malaysia and Ministry of Higher Education, for the sponsorship of my education and my family in Nottingham.

Last but not least, a special compliment to my husband, Taufik, and my three boys, Haiman, Ubai and little Hafiz, who are keeping me alive to make this whole adventure possible. I would like to thank my parents for their continuous encouragement and passion for my education. This is for you mum and dad, Siti Besah Ismail, and Abu Bakar Abdul Salam, with all my love.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

a-CGH	Array Comparative Genomic Hybridisation
BAC	Bacterial Artificial Chromosome
BSA	Bovine Serum Albumin
CEPH	Centre de'Etude du Polymorphisme Humain
CD	Crohn's Disease
CGH	Comparative Genomic Hybridization
CNPs	Copy Number Polymorphisms
CNV	Copy Number Variations
CNVs	Copy Number Variants
ddNTPs	dideoxyNucleotides
DNA	DeoxyNucleic Acid
dNTP	deoxyNucleotides
FISH	Fluorescence *In Situ* Hybridisation
Indels	Insertions and Deletions
kDa	kiloDalton
LCRs	Low Copy Repeats
LCVs	Large-scale Copy Number Variations
LD	Low DNTP
MAPH	Multiplex Amplifiable Probe Hybridization
MLPA	Multiplex Probe Ligation Assay
MSVs	MultiSite Variants
NAHR	Non-Allelic Homologous Recombination
PCR	Polymerase Chain Reaction
PEM	Paired-End Mapping
PPRT	Pyrosequencing-based Paralog Ratio Test
PRT	Paralogue Ratio Test
QF-PCR	Quantitative Fluorescent PCR
REPD	REPeat Distal
REPP	REPeat Proximal

REVDR	Restriction Enzyme Digest Variant Ratios
RNA	RiboNucleic Acid
ROMA	Representational Oligonucleotide Microarray
RT-qPCR	Real Time Quantitative PCR
SDs	Segmental Duplications
SKY	Spectral Karyotyping
SNP	Single Nucleotide Polymorphism
STRs	Short Tandem Repeats

PUBLICATIONS RESULTING FROM THIS THESIS

Work from this thesis is reported in the following publication:

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.
Proceedings of the National Academy of Sciences **106**(3): 853-858.

Chapter 1: Introduction

1.1: Genome variation

Genome variations are differences in the sequence or copy number of DNA from one individual to another. Human genetic variation can be catalogued into large-scale variation involving more than 3 Mb which may be detectable at a microscopic (cytogenetic) level, and smaller (submicroscopic) scale variation which involves less than 3 Mb of DNA and includes the highly abundant copy number variants (CNVs) and single nucleotide polymorphism (SNP) variants. This field began from early research on human genetic diversity through cytogenetic approaches, which involved large-scale microscopic chromosome number and structural variation. Nowadays, as technology has advanced to resolve different classes of submicroscopic structural variation, research into variation has become one of the major efforts in human genetics, since these genetic differences are also found to play a crucial role in susceptibility to both inherited and infectious diseases. Furthermore, study on genetic diversity has allowed scientists to examine how different individuals can have a different response when affected by a disease, which will facilitate development of a specific treatment. Consequently, knowledge of human genetic diversity is a fundamental key to establish a link between genomic variation and phenotypic differences.

1.1.1: Numerical chromosomal variation and its consequences

Genetic differences in the human genome were first observed in the quantity and structure of chromosomes. These chromosomal variations were often associated with phenotypes in specific human disorders and could be seen by light microscopy in conventional cytogenetics, subsequently developed into the high-resolution chromosome banding techniques and fluorescence *in situ* hybridisation (FISH). These abnormalities can involve either an extra haploid set (n or 23 chromosomes) or sets of the chromosome to the normal diploid set (2n or 46 chromosomes) which is known as polyploidy, or an abnormal number of an individual chromosome (aneuploidy). Triploidy (3n) is one of the commonest chromosomal abnormalities seen in humans and is usually lethal during development, whereas tetraploidy (4n) is rarer than triploidy and always lethal. Polyploidy could be caused by the failure of meiosis in the germ cell (triploidy) or failure of the first cleavage division during mitosis which the DNA has replicated to give four chromosomes, but cell division failed to function normally (Jack, 1999; Peter and Ellard, 2007). The numerical chromosomal abnormalities are a relatively common occurrence in which the most obvious is trisomy of chromosome 21 (Down's syndrome) (Lejeune *et al.*, 1959) or of the sex chromosomes (for example 47,XXX, 47,XXY, or 47,XYY) (Jacobs *et al.*, 1997).

Aneuploidy may involve loss of a chromosome pair (nullisomy) or single chromosome (monosomy); or gain of a chromosome pair (tetrasomy) or

single chromosome (trisomy). Whole chromosome trisomy commonly results from meiotic or mitotic non-disjunction events. Examples of a condition caused by trisomy of autosomal chromosomes are Down's syndrome (Lejeune *et al.*, 1959), also known as Trisomy 21, (an individual with Down's syndrome has three copies of chromosome 21, rather than two); trisomy of chromosome 18 (Edward's syndrome) (Edwards *et al.*, 1960) and trisomy chromosome 13 which is known as Patau's syndrome (Patau *et al.*, 1960). Down's syndrome is the most common and compatible with survival to adulthood; it occurs in about every 700 live births and is much more common with older mothers. Individuals with this syndrome are characterized by marked hypotonia (low muscle tone) and a constellation of characteristic facial features. They tend to have a lower than average cognitive ability, often ranging from mild learning difficulties to severe mental disability, and a high frequency of congenital heart disease (Korenberg *et al.*, 1994).

Furthermore, in contrast to the autosomes, having an additional sex chromosome as found in syndromes such as Klinefelter's syndrome (47,XXY) (Jacobs and Strong, 1959), causes relatively few problems and is compatible with normal lifespan. Autosomal monosomies are usually lethal during early embryonic development, but loss of a sex chromosome is less deleterious. Turner's syndrome is an example of monosomy, in which an individual is female phenotypically but is born with only one sex chromosome, an X chromosome which is maternally derived with one of

#the paternal X or Y chromosome (45,X) lost either in meiosis or early embryogenesis (Jacobs *et al.*, 1997). Klinefelter's syndrome is the most common disorder affecting the sex chromosomes described in humans, and occurs in one of every 500 male births. Most affected males are infertile and have some reduction in speech and language ability. Meanwhile the typical clinical features of a female patient with Turner's syndrome are congenital webbed neck and deformity of the elbow; they have a variable phenotype, which includes short stature, ovarian failure, specific neurocognitive deficits and anatomical abnormalities (Sybert and McCauley, 2004). The incidence of this syndrome is found in one out of 3000 live female births. It is likely that the characteristic clinical phenotypes associated with expand monosomies and trisomies of chromosomes relate to a relative gene dosage effect.

1.1.2: Structural chromosomal variation and its consequences

A number of cytogenetically visible chromosomal rearrangements (structural chromosomal variation) have been identified. An example of a chromosome structural abnormality is a rearrangement resulting in a translocation event between two acrocentric chromosomes that fuse near the centromere region with loss of the short arms; such Robertsonian translocations have been seen involving all combinations of acrocentric chromosomes (chromosomes 13, 14, 15, 21 and 22). Reciprocal translocation may result from single breaks in any of two non-homologous chromosomes. The karyotype of Robertsonian translocations is left only

with 45 chromosomes since two chromosomes have fused together. There may be a balanced translocation in which there is no overall loss or gain of genetic material so that carriers of this balanced translocation typically have no adverse effects except relating to reproduction where infertility, recurrent spontaneous abortion, and risk of chromosomal imbalance among offspring can occur. These structural abnormalities may be constitutional involving for example recombination between mispaired chromosomes in the meiotic steps preceding gamete formation, or be acquired and arise in somatic cells secondary to exogeneous or endogeneous agents causing double-stranded DNA breaks.

Other large scale structural variations discovered within a single chromosome or homologous pair of chromosome could be classified into deletions, microdeletions, duplications, inversions, isochromosomes and marker chromosomes. Genomic disorders are a group of genetic diseases which can arise from such chromosomal rearrangements which results in a gain, loss or disruption of dosage-sensitive genes (Lupski, 1998). There is clinical and phenotypic diversity for genomic disorders based on the actual genes and the number of genes that are deleted, and most are involved with mental retardation and congenital malformations. Cri du Chat and Wolf-Hirschhorn syndromes are two common examples of deletions involving chromosome 5p and 4p respectively. Microdeletion from paternal 15q11.2-q13 is a main cause of Prader-Willi syndrome (PWS) and Angelman syndrome (AS) involves the maternal deletion of chromosome

15q11.2-q13 (Nicholls, 1993). PWS and AS represent the classical imprinting disorders in humans due to deletion of the paternal copies of the imprinted *SNRPN* gene and of maternal copies of the imprinted *UBE3A* gene respectively (Glenn *et al.*, 1997).

Inversion is also one of the possible structural variations in our genome and is defined as a variant in which DNA is reversed in orientation with respect to the rest of chromosome (Figure 1). Pericentric inversions include the centromere, whereas paracentric inversions do not involve the structure of the centromere. These abnormalities may not be associated with any phenotypic consequences but inversions of segmentally duplicated olfactory receptor genes at 4p16.1 and 8p23.1 have been associated with having offspring carrying a recurrent unbalanced translocation (Giglio *et al.*, 2002). Isochromosomes occur when the centromere is divided transversely, instead of longitudinally in which one arm is lost and replaced with an exact copy of the other arm. It is found in Turner's syndrome as a mosaic cell line along with a 45,X cell line. A marker chromosome (mar) is a structurally abnormal chromosome of which the origin is not clear. The significance of a marker is hugely variable as it depends on what material is contained within the marker. Nevertheless, structurally abnormal variants as described above including translocations, inversions, deletions and duplications at the chromosome level, are identified less frequently than aneuploidies. Thus, their numbers might be

underestimated particularly with respect to the submicroscopic structural variation.

1.1.3: Submicroscopic structural variation and its consequences

Besides changes at the cytogenetic level, human genetic variation also involves structure of the DNA sequence on the chromosome known as submicroscopic structural variation. Extensive studies have been performed on polymorphic variation in the genome structure because of the ability to analyze duplication and deletions using hybridization techniques such as Southern blot hybridisation. Variation also can involve either only one base or many bases and can occur at any point in a DNA sequence. From these changes, there are numerous types of DNA variation, ranging from Single Nucleotide Polymorphisms (SNPs) to larger structural alterations such as copy number variants and inversions (Iafrate *et al.*, 2004; Sebat *et al.*, 2004; Khaja *et al.*, 2006). If the two strands of a chromosome are thought of as nucleotides threaded on a string, then, for example, a string can break, and the order of the beads can vary. One or more nucleotides may be changed, added, or removed. These structural rearrangements can be called polymorphisms if they are sufficiently frequent in the population, and can in principle involve insertions, deletions, copy number variants (CNVs) due to deletion or duplication, inversion or translocation, based on the changes of DNA nucleotides with respect to the reference genome as illustrated in figure 1. Most cases of genomic disorders result from such submicroscopic structural variation, but some of

these structural variants arise among healthy individuals and some may contribute to susceptibility to disease.

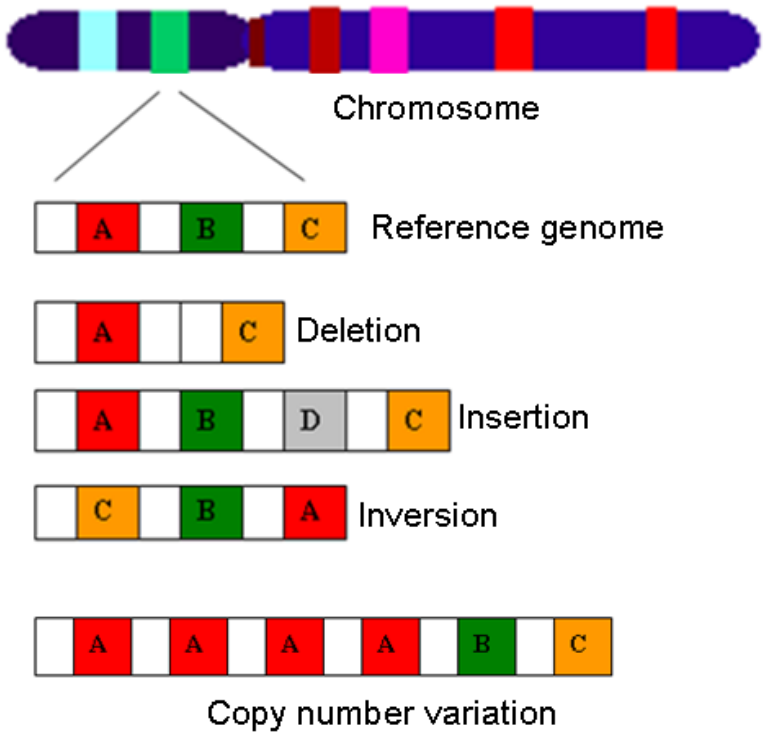


Figure 1: Illustration of the examples of DNA variation involving structural rearrangements of one or more regions.

Among types of small-scale genetic variation, insertions and deletions (indels) are second commonest in the human genome. Indels have collectively been described as gain or loss of one or more contiguous nucleotides in genomic sequence when compared to a reference or ancestral sequence. The extent of indel polymorphism is becoming apparent with an estimated 1.5 million indels in human populations (Mills *et al.*, 2006). Furthermore, two studies have used the HapMap data on family

trios to identify copy number polymorphisms (CNPs) (Conrad *et al.*, 2006; McCarroll *et al.*, 2006). In addition, Conrad *et al*. (2006) have found that deletions of 5 kb and larger are extremely widespread in the human genome and when a deletion is transmitted from parent to offspring, the child will show a null genotype or a genotype violating the rules of mendelian inheritance (Conrad *et al.*, 2006). Other investigators have used an array comparative genomic hybridization approach to compare 24 unrelated individuals with the reference human genome sequence (Hinds *et al.*, 2006). Hinds and colleagues (2006) found 215 deletions ranging in size from 70 bp to 7 kb.

Tandem repeat DNAs, for example minisatellites and microsatellites with repeat size ranging from 2 bp to 1000 bp also contribute to variation in the human genome. Variable length, which is based on numbers of repeating units, makes them highly informative as genetic markers of diversity. Satellite DNA comprises very long arrays of tandem repeats typically 100 kb to several megabases in size and mainly forms the component of functional centromere, heterochromatin in pericentromeric and telomeric regions, and in the short arms of the acrocentric chromosomes. Minisatellite DNA arrays are of intermediate size and typically span between 100 bp and 20 kb with each repeat unit between 6 and 100 bp in length. Meanwhile, microsatellite DNA comprises short arrays less than 100 bp in size, made up of simple tandem repeats 1 – 6 bp in length. Each variant is inherited allowing them to be used for personal or parental

identification. Therefore, analysis of their length is useful in genetics and biology research, forensics, and DNA fingerprinting. Tandem repeats including minisatellite and microsatellite have been associated with a number of diseases and phenotypic conditions (Armour, 2006). For example polymorphism in expanded arrays of the triplet (CCG) upstream of the *FMR1* gene causes fragile-X syndrome by reducing *FMR1* transcription (Pieretti *et al.*, 1991) and Huntington's disease is caused by the expansion of tandemly repeated CAG (glutamine) codons in the coding sequence, associated with gain of function (Macdonald *et al.*, 1993). Polymorphism of tandem repeats within transcribed sequences for example, interactions of the expanded (CUG)n in RNA molecules transcribed from the *DMPK* gene with specific RNA binding proteins is known to modulate disease and can effect changes in the protein products of genes, leading to diseases such as myotonic dystrophy (Machuca-Tzili *et al.*, 2005). Meanwhile, one of the earliest and best known minisatellite polymorphisms was found upstream of the insulin gene, where shorter alleles were associated with increased risk of type 1 diabetes (Bennett *et al.*, 1995). These polymorphisms can arise from events such as unequal crossover, replication slippage or double-strand break repair.

Certain features of genomic architecture, notably highly homologous low copy repeats (LCRs), predispose to chromosomal rearrangements through non-allelic homologous recombination (NAHR) and are the origin of a diverse group of genetic diseases (genomic disorders) (Lupski, 1998). The

term low copy repeats (LCRs) was used by Bernice Morrow following her studies of DiGeorge syndrome rearrangement breakpoints (Edelmann *et al.*, 1999). Subsequently, the term “segmental duplications” (SDs) was first introduced by Evan Eichler (Bailey *et al.*, 2001; Bailey *et al.*, 2002) to explain his observations from genome wide studies. SDs are blocks of DNA ranging from 1-400 kb in length that occur at multiple sites within the genome and typically share a high level (>90%) of sequence identity (Eichler, 2001) and in total comprise about 5% of the human genome (Bailey *et al.*, 2002). Some of the block sequences of SDs may be duplicated to multiple locations within a single chromosome, termed intrachromosomal duplication, and others might be located on non-homologous chromosomes termed inter or transchromosomal duplication (Sharp *et al.*, 2006). SDs are found interspersed throughout the genome and tend to cluster at pericentromeric and subtelomeric regions (Sharp *et al.*, 2006). These features of SDs have considerable implications for human disease, evolution and other structural variation, as they provide a substrate for structural rearrangements resulting in the deletion, duplication or inversion of the intervening sequence by non-allelic homologous recombination (NAHR) (Sharp *et al.*, 2006). Many studies have suggested that segmental duplications frequently mediate polymorphic rearrangement of intervening sequences (Iafrate *et al.*, 2004; Sebat *et al.*, 2004; McCarroll *et al.*, 2006), and in addition they are often themselves variable in copy number (Sharp *et al.*, 2005), which could certainly be generated by NAHR.

Colour vision variation (Nathans *et al.*, 1986; Vollrath *et al.*, 1988) and Rhesus (Rh) blood groups were shown to result from this kind of rearrangement (Sharp *et al.*, 2006). High sequence identity between those duplicated genes, for example in *OPN1LW* ('opsin 1 long-wave-sensitive' for the red photopigment) and *OPN1MW* ('opsin 1 medium-wave-sensitive' for the green photopigment) genes, has predisposed to frequent NAHR leading to red-green colour vision defects of variable severity affecting up to 8% of male north Europeans (Deeb, 2004). Unequal crossing over which involves two highly homologous regions flanking the RHD gene known as rhesus boxes has also mediated the gene rearrangement for this gene (Wagner and Flegel, 2000). A further example is the alpha globin gene cluster, containing two highly homologous genes *HBA1* and *HBA2*, which also highlights the occurrence of gene duplication. Unequal crossing over is thought to be involved with the occurrence of duplication or deletion of the *HBA* genes and then provides examples of copy number variation. Normal humans usually have four alpha-globin genes (αα/αα); two on each chromosome. Individuals from different ethnic backgrounds have been reported having one chromosome with triplicated alpha-globin loci (Goossens *et al.*, 1980). Most alpha thalassemia disease involves with deletion of one or both alpha-globin genes (Higgs *et al.*, 1989) which results from unequal crossing over at meiosis due to extremely similar sequences within the gene cluster.

Further examples of recurrent submicroscopic structural variants that are involved with genomic disorders include Charcot-Marie-Tooth disease type 1A (*CMT1A*) and hereditary neuropathy with liability to pressure palsy (*HNPP*) resulting from reciprocal duplication and deletion, respectively (Lupski, 2009). Individuals with three copies of the *PMP22* gene (encoding a myelin protein) have CMT type 1A disease, and deletion of the gene is associated with HNPP. This region is flanked by complex low copy repeat sequences where reciprocal recombination events can occur if misalignment of the homologous chromosomes involving the region happened (Reiter *et al.*, 1996; Reiter *et al.*, 1998). Pelizaeus-Merzbacher disease is one of the examples of non-recurrent duplication, which is seen associated with the dosage sensitive proteolipid protein 1 gene (*PLP1*) at chromosome Xq22 (Lee *et al.*, 2007).

An example of large inversion that arises from non-allelic homologous recombination involving highly homologous low copy repeat regions (Shaw and Lupski, 2004) is on the long arm of chromosome 17. Stefansson and colleagues found a large inversion polymorphism on chromosome 17q21.31 that was common among Europeans with an inversion frequency of 21% but rarer in Africans (6%) and even rarer among Asians (1%) (Stefansson *et al.*, 2005). Analysis on H2 lineage of chromosome 17q21.31 inversion showed that the carrier females of this inversion polymorphism have more children and higher recombination rates (Stefansson *et al.*, 2005).

1.1.3.1: Copy Number Variation (CNV) and its consequences

DNA copy number variation (CNV) has long been known as a source of genetic variation, but its importance has only been appreciated recently from studies using new technologies (Iafrate *et al.*, 2004; Sebat *et al.*, 2004; Redon *et al.*, 2006). These studies have identified a previously uncharacterized prevalence of structural variants of DNA along chromosomes in the size range of 1kb or greater in the human genome, termed submicroscopic copy number variations. Copy number variation (CNV) is defined as change of DNA dosage of DNA segments greater than 1 kb in size (and typically less than 3 Mb) (Scherer *et al.*, 2007). CNV is characterized by variable numbers of copies of a DNA segment (Feuk *et al.*, 2006), and large copy number variants tend to be flanked by segmental duplications. New copy number variation may arise from NAHR as discussed earlier but can arise even when there is no sequence similarity between flanking sequences, via simple non-homologous end-joining to repair double-stranded DNA breaks that can lead to loss or gain of nucleotides from imprecise repair.

Iafrate *et al.* (2004) and Sebat *et al.* (2004) demonstrated a higher level of large scale copy number variants to be common among apparently phenotypically normal individuals than previously appreciated. Even though both of these research teams used technologies that are limited in coverage and resolution, the unexpected scale of CNV that was found promoted research interest and further high resolution analyses. Moreover,

Redon and colleagues found 1447 copy number variable regions spanning a total of 360 Mb of sequence, which comprises at least 12% of the human genome for a large panel of ethnically diverse individuals (Redon *et al.*, 2006). They studied 270 lymphoblastoid cell lines from the International HapMap project, established from individuals of African, European, and Asian origin.

More recently, several studies have shown that some genes that appear variable in copy number can have important functional consequences. Given that, it seems likely that this variation may play a fundamental role in influencing human disease susceptibility. However, not all variations in genomic structure have an effect. Among the variations that do cause effects, some are more serious than others. The outcome depends on two factors: where in the genome the change occurs (for example, in a non-coding region, coding, or regulatory region) and the exact nature of the change such as the size and boundaries of the changes. The *AMY1* gene, which encodes salivary amylase was one of the large scale CNV identified and observed in half the individuals studied by Iafrate and co-workers with relative gains and losses in equal numbers of people (Iafrate *et al.*, 2004). Further study by Perry and colleagues on this gene found that it varied in copy number from two to 15 with a mean of seven in the European-American population (Perry *et al.*, 2007). The authors compared the CNV in *AMY1A* between different population groups sampled from Africa, Asia, and Europe, and found particular groups that consume high starch diets

did indeed have more *AMY1A* copies. They also proposed that there had been positive or directional selection for CNV among these individuals, while in populations with low starch diets, there had been genetic drift and the locus had evolved neutrally (Perry *et al.*, 2007).

Another example is the *CYP2D6* gene found at chromosome 22q13.1. This copy number variable gene was first described with individual alleles between zero and thirteen copies (Johansson *et al.*, 1993). *CYP2D6* is involved in the metabolism of an estimated 20 – 25% of all marketed drugs, and thus genetic polymorphism of *CYP2D6* is responsible for much of the observed variation between individuals in enzyme activity (Ingelman-Sundberg *et al.*, 2007). For example, individuals who have more than two active functional copies of a CYP gene may demonstrate increased drug metabolism and absence of response at ordinary drug dosages (ultrarapid metabolzers) while poor metabolizers lack functional enzyme due to defective or deleted genes. Currently, there is an increasing amount of published work on the potential role of copy number variation in complex diseases including susceptibility to autism, schizophrenia, psoriasis, systemic lupus erythematosus (SLE) and HIV-1. For example, the genes *CCL3L1*, *CCL4L1,* and *TBCID3* are present in different copy numbers (between 0 and 10 copies per individual) within and between populations. In Europe, common variation is from zero to four copies, whereas in Africa, copy numbers as high as ten have been recorded. This variation is reported to be involved in susceptibility to, and progression of, infection

with HIV-1 (Townson *et al.*, 2002; Gonzalez *et al.*, 2005). For individuals with a higher number of *CCL3L1* copies, the risk was significantly reduced, while lower copy numbers were associated with faster disease progression.

Beta-defensin genes, the genes of interest in this study, are located on chromosome band 8p23.1 and were found commonly to be variable between two to seven copies, but in some individuals there were up to 12 copies present (Hollox *et al.*, 2003). These genes encode small secreted antimicrobial peptides with cytokine-like properties as described in section 1.2.2. Hollox and colleagues have highlighted the role of copy number variation on this region in susceptibility to psoriasis; when more than four copies in diploid genome were present, each extra copy increased the relative risk of disease (Hollox *et al.*, 2008). Intriguingly, two different studies at this locus demonstrated different finding of association with inflammatory bowel disease, particularly in colonic Crohn's disease (Fellermann *et al.*, 2006; Bentley *et al.*, 2009), which will be discussed in chapter four. In contrast, individuals with lower numbers of copies of total the C4 and C4A genes were found with increasing risk of the autoimmune disease SLE, and higher copy numbers were a protective factor against SLE susceptibility in European Americans (Yang *et al.*, 2007). Another example is *FCGR3B* gene which encodes the Fc fragment of IgG, low affinity IIIb receptor on chromosome band 1q23, which has copy number ranges from none, one, two, three and four diploid copies. Low copy

number of this gene is found significantly increased risk for glomerulonephritis disease (Aitman *et al.*, 2006; Fanciulli *et al.*, 2007).

Copy number variation of DNA sequences may be functionally remarkable because it can influence or correlate with the level of gene expression through effects on gene dosage in which there is loss or gain of functional gene copies (McCarroll *et al.*, 2006). It may also act indirectly through position effects, predisposition to deleterious genetic changes, or by providing substrates for chromosomal changes in evolution (Feuk *et al.*, 2006). A genome-wide analysis of the association between copy number variation and gene expression was undertaken using lymphoblastoid cell lines established from 210 unrelated individuals in the International HapMap Project from four populations (Stranger *et al.*, 2007). The authors found a significant association between copy number variation and gene expression with as much as 17.7% of the heritable variance in gene expression attributable to copy number variation.

For nearly all genes in the human genome commonly present as two copies, each copy has been inherited from each parental genome so that every diploid cell carries two copies of a gene in the nucleus. However when (for example) a copy number variable gene is measured as three copies per diploid genome, there are two possible haplotype combinations, corresponding to either 1 + 2 or 3 +0. Moreover, there is also no information of which of these haplotypes will be inherited from which parent

except by segregation analysis of copy number alleles in a pedigree. Copy number genotyping is not as simple as SNP or other basic deletion or duplication events as shown in the figure 2, adapted from Wain *et.al.,*(2009). The simplest copy number variation is the presence or absence of a gene as illustrated in figure 2A. For example in European population, the rhesus-negative allele at the main antigen locus for the rhesus blood group is commonly caused by complete deletion of the RHD gene and an individual's (diploid) genome could therefore contain two, one, or zero copies of RHD, with the zero copies corresponding to rhesus-negative and absence of D antigen expression. Genomic segments with variable copy number could encompass parts of genes, reside entirely outside genes, or, in the case of larger variants, include several known genes. Thus, a simple duplication of a genomic segment could result in diploid copy numbers of two, three, or four (figure 2B), and successive rounds of duplication could produce a wide range of diploid copy numbers, known as multiallelic copy number variants (figure 2C) which involves more complex variation. All of this information is not been explored very well yet for many CNVs in the human genome. It may contribute important knowledge to the genotype-phenotype correlation for copy number variation. In order to address the lack of studies on the origin of the copy number in the locus-specific genes, the human defensin genes, particularly of the beta-defensin class located at chromosome 8p23.1, were chosen as the region to examine in this study. This region is one of the most

fascinating examples of structural variation in the human genome and is also associated with autoimmune disease.

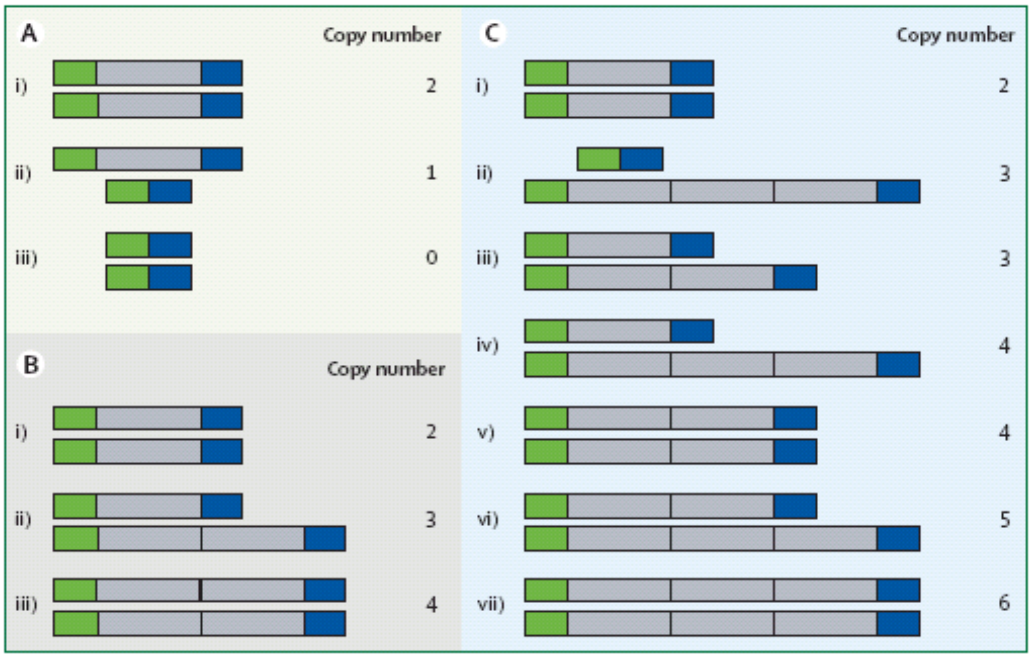


Figure 2: Diallelic and multiallelic copy number variation. Diallelic locus (grey) and flanking loci (green and blue) with variation caused by (A) deletion and (B) duplication, each showing the locus with (i) normal copy number, (ii) heterozygous modification, and (iii) homozygous modification. (C) Multiallelic locus showing (i) normal copy number, (ii) multiple rounds of duplication on one chromosome and a deletion on the homologous chromosome, (iii) duplication on one chromosome and no deletion on the homologous chromosome, (iv) two rounds of duplication on one chromosome and no deletion on the homologous chromosome, (v) one round of duplication on each chromosome, (vi) one round of duplication on one chromosome and two rounds of duplication on the homologous chromosome, and (vii) two rounds of duplication on both chromosomes.

Multiallelic assays measure diploid copy number and therefore cannot distinguish between (ii) and (iii), or (iv) and (v). Note that although duplications are usually assumed to be contiguous, this might not always be the case (figure was taken from Wain *et al.,* (2009)).

1.1.4: Human Chromosome 8p23.1

Human chromosome 8p23.1 as shown in figure 3 is a good example to study several types of structural variation. This region is known to be a frequent site of chromosomal rearrangements mediated by two large blocks of low copy repeats (LCRs) or segmental duplications (SDs). The whole 8p23.1 region, which includes the SDs and contains at least 50 genes, can extend up to 6.5 Mb. The two large set of SDs blocks are named REPeat Distal (REPD) distally and REPeat Proximal (REPP) proximally. Each of the two SDs includes olfactory receptor gene clusters and other genes such as defensins and FAM90A clusters which are found to be copy number variable (Hollox *et al.*, 2003; Aldred *et al.*, 2005; Bosch *et al.*, 2008). Thus, the sizes of the flanking SDs are variable with gaps that are not fully sequenced in the assembled genome sequence. These characteristics seem to be sufficient for this SD variability to play an important role in the distinct rearrangement affecting the 8p23.1 region (Bosch *et al.*, 2009).

Furthermore, Giglio *et al.* (2001) first reported inversion heterozygosity of the chromosome 8p23.1 region with 26% in the population of European

descent, and Sugawara *et al*. (2003) found 34% in Japanese populations. This is one of the most common submicroscopic inversions identified in autosomes assuming that the assembly of the reference sequence corresponds to the non-inverted conformation. Further studies have found an increased frequency of this inversion (around 60%) in populations of European ancestry that indicates the human reference assembly corresponds to the minor allelic orientation of this region (Chen *et al.*, 2006; Libin *et al.*, 2008; Antonacci *et al.*, 2009). The relevance of this inversion variant in the general population is its relationship with flanking segmental duplications (REPD and REPP), which may also be involved with generation of copy number variations (CNVs) during crossover. The type of rearrangements are predominantly defined by the orientation of recombining duplicons and the number of crossovers for example, deletion and reciprocal duplications result from mismatch of direct repeats and a single crossover between the distal and the proximal repeat. Inversions could result through mispairing of inverted repeats and a double recombination event between them (Small and Warren, 1998).

Interestingly, the DEF cluster is assembled at this region, the defensin gene clusters including alpha, beta and theta-defensins. The beta-defensin clusters have been identified at REPD in 8p23.1. Both alpha and beta-defensins have been found to be variable independently as described in sections 1.2.1 and 1.2.2. In the 1990s, cytogenetically visible alteration at this region was first reported but was apparently benign (Barber *et al.*,

1998) and over the next decade, the vast majority of individuals carrying additional 8p23.1 material did not have any identifiable phenotypic abnormalities (Barber *et al.*, 1998; Barber *et al.*, 2005).

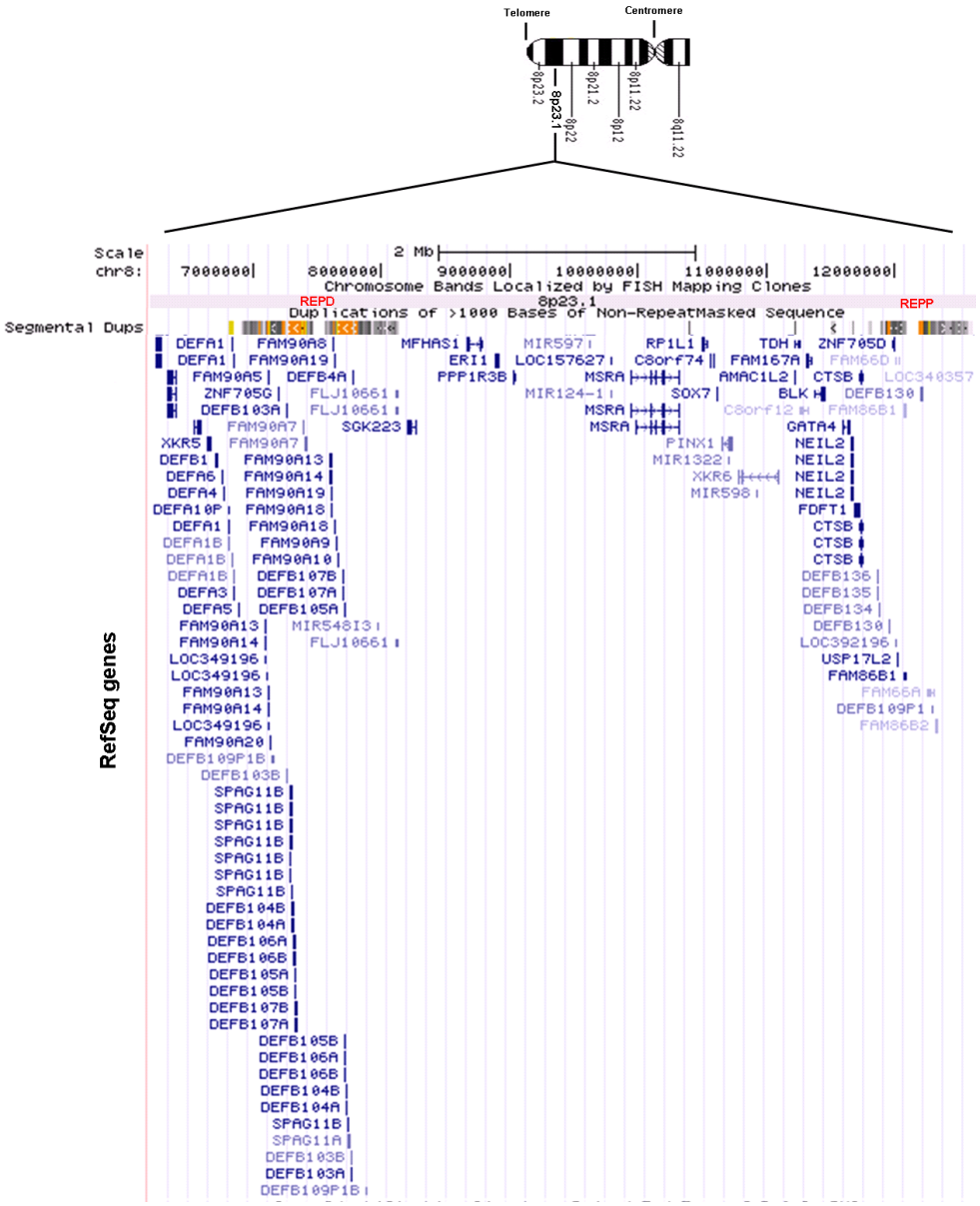


Figure 3: Chromosome 8p23.1 region flanking with two large blocks of segmental duplication named as REPD distally and REPP proximally. The beta-defensin clusters appear twice here. Figure illustrated from the UCSC Genome Browser on Human March 2006 (NCBI36/hg18) Assembly.

1.2: Human defensins

Defensins are among the most important antimicrobial peptides in the innate immune system in humans and other mammals. They have molecular masses of 3.5 – 6 kDa and encode a family of small cationic antimicrobial peptides characterized by six conserved cysteine residues cross-linked through disulphide bridges. Based on the arrangement of the cysteines and the disuphide-bonding pattern, these peptides are divided into three subfamilies: alpha-defensins, beta-defensins, and theta-defensins (Ganz, 1999; Yang *et al.*, 2001). In humans most of the genes encoding alpha and beta-defensins are located in clusters on chromosome 8p23.1 as illustrated in figure 3 but they appear as well in other clusters on chromosome 6 and 20 (Linzmeier *et al.*, 1999; Taudien *et al.*, 2004). The disulphide linkages of cysteine residues in alpha-defensins are between the first and the sixth cysteine residues Cys^1-Cys^6, Cys^2-Cys^4, and Cys^3-Cys^5, whereas in beta-defensins, the linkage are Cys^1-Cys^5, Cys^2-Cys^4, Cys^3-Cys^6. In contrast, theta-defensins have a circular structure with the cysteine residues linked as Cys^1-Cys^6, Cys^2-Cys^5, Cys^3-Cys^4 (Tang *et al.*, 1999; Klotman and Chang, 2006).

1.2.1: Alpha defensins

Alpha-defensins, which have been identified in humans, monkeys and several rodent species, are particularly abundant in neutrophils, certain macrophage populations, and Paneth cells of the small intestine. Human alpha-defensins 1, 2, 3, and 4 (*DEFA1* to *DEFA4*) are constitutively

produced by neutrophils, whereas *DEFA5* and *DEFA6* are produced in the Paneth cells. Most alpha defensin genes (*DEFA6, DEFA4, DEFA1, DEFA3,* and *DEFA5*) are found in a cluster at the telomeric end of band 8p23.1 with *DEFA1* and *DEFA3* appearing as a 19 kb tandemly repeated unit. Gene nomenclature for *DEFA1* and *DEFA3* has been replaced by *DEFA1A3*, following the recommendation of Aldred *et al.* (2005). *DEFA1* and *DEFA3* encode different peptides, human neutrophil-derived alpha-defensin 1 (HNP-1) and human neutrophil-derived alpha-defensin 3 (HNP-3) respectively. The mature HNP-1 and HNP-3 peptides differ only in their N-terminal amino acid, due to single nucleotide difference, C3400A (Ganz and Lehrer, 1995). C3400A is a paralogous sequence variant (PSV) that allows discrimination between the two gene copies. HNP1, HNP2, and HNP3 peptides are easily purified from leucocytes, and their properties has been widely studied while HNP4, HNP5 and HNP6 have been recovered only in small amounts; considerably less is known about their properties (Harwig *et al.*, 1992).

The HNP-2 peptide is identical to the last 29 amino acids of both the HNP-1 and the HNP-3 peptides. Given that no gene for DEFA2 has been discovered, it is thought that DEFA2 is a proteolytic product of one or both of *DEFA1* and *DEFA3*. Thus, HNP-2 is presumably produced from proHNP-1 and/or proHNP-3 by post-translational proteolytic cleavage (Ganz, 2003). HNP-1 peptide is likely to play a role in phagocyte-mediated host defense. These three peptides (HNP1, HNP2 and HNP3) are

constitutively produced by neutrophil cell precursors and packaged in granules before mature neutrophils are released into the blood. During phagocytosis, the defensin-containing granules fuse to phagocytic vacuoles where defensins act as antimicrobial agents (Ganz and Lehrer, 1995). Human alpha-defensin 5 (HNP5) is released as a propeptide that is processed extracellularly (Ghosh *et al.*, 2002). To date, alpha-defensins have been shown to have broad and very powerful anti-microbial activity, such as the anti HIV activity recently shown for *DEFA1* (Cole *et al.*, 2002).

1.2.2: Beta defensins

A beta-defensin was first reported as the tracheal antimicrobial peptide (TAP) of cow tongue (Diamond *et al.*, 2000). A similar functional and structural property of this peptide to those described earlier for alpha-defensin assigned TAP to the defensin family. In the beta-defensins, the three disulphide bridges are between residues 1 and 5, 2 and 4, and 3 and 6 which is different from alpha-defensin (Cys^1-Cys^6, Cys^2-Cys^4, Cys^3-Cys^5). In 1995, the first human beta-defensin, HBD-1 was isolated from the plasma of patients undergoing dialysis treatment for renal disease (Bensch *et al.*, 1995). Bensch *et al*. (1995) found that HBD-1 is expressed in epithelial cells that are directly exposed to the environment or microbial flora, for example in the lung, mammary gland, salivary gland, kidney, pancreas, and prostate. HBD-1 has also been implicated in cancer as it is lost at high frequencies in malignant prostatic tissue and has been shown to induce cytolysis and apoptosis in prostate cancer cell lines (Sun *et al.*,

2006). The second human beta-defensin, HBD-2, was originally characterized from psoriatic skin lesions and is widely expressed in epithelia (for example in respiratory tract, gastrointestinal tract, urogenital system, pancreas and skin), leucocytes and the bone marrow (Harder *et al.*, 1997). Meanwhile, another defensin, HBD-3 was nearly simultaneously isolated from lesional psoriatic scales and cloned from keratinocytes and is a broad spectrum peptide antibiotic that kills many pathogenic bacteria (Harder *et al.*, 2001).

Their antimicrobial activity, predominantly against Gram-negative bacteria and sometimes against Gram-positive bacteria, varies such that, HBD-1 is active against Gram-negative bacteria and HBD-2 has considerably greater activity compared to HBD-1. Meanwhile HBD-3 has the strongest activity. HBD-2 and HBD-3 are also induced by inflammatory stimuli such as tumour necrosis factor alpha (TNF-α). Consequently, three other beta-defensin genes (HBD-4 – 6) were discovered by using the Basic Local Alignment Search and HBD-4 was found to be highly expressed in the testis and epididymis (Yamaguchi *et al.*, 2002). More recent work has demonstrated that their functional role seems not to be limited as classic antimicrobial peptides only and further functional study may cover unexpected different roles for what is increasingly regarded as a multifunctional gene family.

The main beta-defensin gene cluster is located within chromosomal band 8p23.1 (figure 3) and is somewhat larger than the alpha-defensin cluster, spanning at least 250 kb in size of repeat unit. The exact size of the beta-defensin repeat unit is unclear and may itself be variable, although from pulse-field gel analysis it is shown to be at least about 260 kb in size (Hollox *et al.*, 2003). Subsequently, three other beta-defensin gene clusters were been identified within chromosomes 6 and 20 (Schutte *et al.*, 2002; Rodriguez-Jimenez *et al.*, 2003). A gene with strong homology to beta-defensin genes called epididymis-specific secretory protein (EP2/HE2/SPAG11) is found to be encoded by a human gene within chromosome 8p23.1 near several other beta-defensin genes (Horsten *et al.*, 2004). It is encodes an androgen-dependent molecule that is specifically expressed in the epithelium of the male reproductive tract. Most genomic structures of the beta-defensin genes contain two exons and one intron, with an exception for the *DEFB105* gene, which contains three exons and two introns. The first exon generally encodes the signal peptide and the second carries information about the mature peptide sequence preceded by a short anionic pro-peptide; this is not true for *DEFB1*, for which the first exon encodes both the signal peptide and pro-peptide segment (Liu *et al.*, 1997). HBD-2 has been demonstrated to be subject to gene copy number variation and polymorphic within the healthy population (Hollox *et al.*, 2003). Copy number variation of the *DEFB4* gene (HBD-2) has been thought to be associated in development of colonic Crohn's

disease (Fellermann *et al.*, 2006; Bentley *et al.*, 2009) and psoriasis (Hollox *et al.*, 2008).

1.2.3: Theta defensins

The only primate theta-defensin described so far are isolated from rhesus macaque leucocytes: rhesus theta-defensin 1 (RTD1), RTD2 and RTD3 (Tang *et al.*, 1999). The mature theta-defensin peptides arise by an as-yet-uncharacterized process that generates a cyclic peptide by splicing and cyclization from two of the nine amino acid segments of alpha-defensin-like precursor peptides (Ganz, 2003). The theta-defensins apparently evolved in primates, but are inactivated in humans due to premature stop codons in the genes that abort translation and subsequent peptide production (Cole *et al.*, 2002)

1.2.4: Copy number variation in human defensin genes

In 1995, CNV of DEFA1 had been observed from somatic cell hybrid mapping of chromosome 8 (Mars *et al.*, 1995). Aldred et. al (2005) and Linzmeier et. al (2005) found that the *α*-defensins *DEFA1* and *DEFA3* are variable in copy number. The other alpha-defensins (*DEFA4, DEFA5* and *DEFA6*) and the beta-defensin *DEFB1* do not show copy number variation (Hollox *et al.*, 2003; Aldred *et al.*, 2005; Linzmeier and Ganz, 2005). The 8p23.1 beta-defensin cluster including the *DEFB4, DEFB103, DEFB104, DEFB105, DEFB106, DEFB107 and SPAG11* genes, was found to be polymorphic in copy number. Most previous studies measured beta-

defensin copy number at different sequences along the repeat region, and found all the above genes varied concordantly (Hollox *et al.*, 2003; Groth *et al.*, 2008; Abu Bakar *et al.*, 2009). Furthermore, none of the studies have provided any support for extensive copy number heterogeneity, and all copy number typing methods for a given sample generally report the same copy number.

In cytogenetic analyses, chromosomes carrying high-copy alleles are visible as 8p23.1 euchromatic variants in individuals found to have nine to twelve copies of the region, whereas most other normal people have two to seven copies per diploid genome (Barber *et al.*, 1998; Hollox *et al.*, 2003; Barber *et al.*, 2005). One copy of the beta-defensin cluster per diploid genome has been observed in only one individual among more than 1500 individuals typed using DNA from blood (Hollox *et al.*, 2008). The author and colleagues suggested that the null state exists but is very rare with haplotype frequency as low as 0.2%. Additionally, the beta-defensin clusters on chromosome 20 do not show common CNV (Hollox and Armour, 2008).

1.3: Methods for detection and measurement of genome variation

At present there are several assays that have been used for the detection of human genetic variation, ranging from point mutations to the larger

genomic variation over more than 100bp. Nevertheless, every single assay has limitations in both quantitative and qualitative terms. Development of all of the techniques depends on the type of variation and the features of the methods. Early studies led to the concept of genomic variation from detection of disorders in individuals, and chromosome analysis is a method to visualize the entire human genome.

1.3.1: Chromosomal aberrations

Chromosomal aberrations include changes in the normal number of chromosomes per individual (46 chromosomes per diploid cell) either to fewer chromosomes (45 chromosomes) or more, such as 47 chromosomes. Conventional cytogenetic analysis including staining of the chromosome bands is more useful for this detection and G – banding is one of the most common methods of chromosome banding (Yunis and Sanchez, 1973; Speicher and Carter, 2005). This technique has been broadly used in the clinical laboratory for both pre- and postnatal diagnosis, and in oncology, where karyotyping allowed scientists to look for somatic chromosomal alterations. However, this technique will not always allow scientists to detect other structural abnormalities such as small-scale duplication or translocation. Due to those limitations, molecular cytogenetics has been developed from the 1980s. Hybridizing fluorescently labelled probes to chromosome preparations, fluorescent *in situ* hybridization (FISH) has allowed investigators further advances in examination of chromosome structure as discussed in the next section.

Comparative genomic hybridization (CGH) (Kallioniemi *et al.*, 1992) and spectral karyotyping (SKY) (Schrock *et al.*, 1996) were then developed to screen the whole set of chromosomes per individual. All of the above techniques from a cytogenetic viewpoint are able to identify structural variation qualitatively, and are useful for detection of some structural variation in the whole genome. Moreover, structurally abnormal variants including translocations, inversions, deletions and duplications are less frequently identified than aneuploidies at the karyotype level.

1.3.2: Structural variation including copy number variation (gene dosage)

Most of the techniques used to investigate structural variation in the human genome may divided into two groups; polymerase chain reaction based (PCR-based) and array-based comparative genomic hybridisation (CGH), with the exceptional of the cytogenetic methods as discussed above and Southern blot hybridisation. There are huge numbers of PCR-based techniques widely used as experimental tools for genetic analysis, combining a highly specific primer hybridization process with exponential amplification of the target sequence. Modifications of PCR to a quantitative process suitable for the measurement of gene dosage have been carried out by considering a number of factors in the theory and principle of the PCR technique. In analysis of copy number polymorphisms, PCR can be used to amplify abundant DNA polymorphic markers like microsatellites and SNPs which has allowed investigators to obtain gene dosage

information in some instances, for example in the detection of large deletions and duplications within chromosome 17p11.2 in Charcot-Marie-Tooth disease type 1A (Brice *et al.*, 1992). However, the preferential amplification of the smaller allele, and the stuttering effect that produces a series of minor peaks immediately adjacent to the major peak, has made microsatellite analysis more difficult to interpret. Nevertheless, the difficulties with quantitative fluorescent PCR (QF-PCR) of microsatellites are reduced by the use of tetranucleotide repeat markers, and multiplex QF-PCR is finding increasing use in the rapid diagnosis of the common aneuploidies (Armour *et al.*, 2002). SNPs which are less informative individually than microsatellites but far more abundant can be used for high throughput genotyping. A number of PCR-based techniques have been developed such as competitive PCR, differential PCR, Real-Time PCR, Long PCR, MAPH, MLPA, and PRT. Some of them are described in the following paragraph on detection and measurement of the structural variation, including copy number variants in the human genome. Meanwhile, developments of array-based CGH have encouraged its use in detection and mapping structural variations.

Previous studies have identified human genomic variation among unrelated individuals by using different approaches. The most commonly used method to detect CNV was array comparative genomic hybridization (aCGH) developed by Pinkel and colleagues. They demonstrated the ability of aCGH to measure copy number with high precision in the human

genome, and to analyse clinical specimens by obtaining new information on chromosome 20 aberrations in breast cancer (Pinkel *et al.*, 1998). The method uses an array of probes that represent the genome and differentially labelled test and reference DNA samples that are jointly hybridized to the array; at each probe, the ratio of the labels is used to identify comparative deletions or duplications of the DNA. An array could be of either bacterial artificial chromosome (BAC) clones, in which fragments of genomic DNA (80-200 kb) from across the genome can be inserted into the bacterial plasmid and replicated, or synthetic oligonucleotides, which are often uses as targets. Iafrate *et al*. (2004) also using BAC clone array CGH described large-scale copy number variations (LCVs), involving gains and losses of several kilobases to hundreds of kilobases of genomic DNA; they found twenty-four variants in more than ten percent of the 55 unrelated individuals by using similar methods (Iafrate *et al.*, 2004).

Sebat *et al*. (2004) showed that large-scale copy number polymorphisms (CNPs) (at a scale of about 100 kilobases and greater) contribute substantially to genomic variation between normal human genomes by using representational oligonucleotide microarray analysis (ROMA) of 20 individuals. The study revealed a total of 221 copy number differences representing 76 unique copy number polymorphisms (CNPs). This method is a variant of array-CGH in which the reference and test DNA samples are made into “representations” to reduce the sample complexity before

hybridization to a custom oligonucleotide array (Sebat *et al.*, 2004). However, the most common variant in the Iafrate *et al*.,(2004) study, located at 1p21.1 (*AMY1A-AMY2A*) and present in >49% of the individuals studied, was not detected by Sebat *et al.* (2004). Therefore, this might suggest that the coverage or resolution of the methods used is different between those two studies. Furthermore, data from array comparative genomic hybridization are often much harder to assess than a sequence trace and sometimes may overestimate the size of the copy number variant because signal ratio from BAC clones is mapped onto the extent of the fragment of DNA inserted into the bacterial plasmid.

By contrast, CNV sequencing-based approaches can map copy number variants with much greater accuracy than aCGH, and can detect inversions or translocations (Wain *et al.*, 2009). Currently there are two possible ways of comparison of a segment of reference and test DNA sequences which are paired-end mapping (PEM) (Korbel *et al.*, 2007) and fosmid end sequencing (Tuzun *et al.*, 2005; Kidd *et al.*, 2008). For fosmid cloning, genomic DNA is inserted into a fosmid cloning vector and transfected into cells for propagation. Only a restricted size range of DNA inserts (32 – 48 kb) can be efficiently cloned (Tuzun *et al.*, 2005) whereas PEM directly uses a test sample prepared from size-selected genomic DNA (Korbel *et al.*, 2007). In addition Tuzun and colleagues published the results of a sequence-based approach to try to produce a fine scale map of structural variation across the genome (Tuzun *et al.*, 2005). They found 297 sites of

variation including 139 insertions, 102 deletions, and 56 inversion breakpoints; 28% showed copy number variation when 57 sites were analysed and validated by array CGH. They also noted overrepresentation of structural variation for genes involved in drug detoxification, innate immunity and inflammation, surface integrity, and antigens (Tuzun *et al.*, 2005). These recent studies have shown that sequencing-based approaches have played an important part in development of maps of copy number variation, but at present, the expense of this technique prohibits their application on a genome-wide scale across many individuals.

Another method for mapping CNV is using signals from SNP arrays as a proxy, though early SNPs arrays sometimes excluded many SNPs from regions of copy number variants. New hybrid oligonucleotide microarrays may overcome this limitation, which include non-polymorphic probes to detect copy number variation (widely referred to as non-SNP probes or copy number variant probes) and probes for many more SNPs that were excluded in the early arrays. These arrays have been developed to accurately analyze SNPs and copy number variation simultaneously (McCarroll *et al.*, 2008). Mccarroll *et al*., used these arrays to map the genomic locations, allele frequencies and population-genetic properties of human copy number polymorphisms (CNPs); and to apply this knowledge to advance strategies for querying CNV in genome-wide association. Interestingly, McCarroll and co workers documented that large-scale (>100 kb) CNV affects far less of the human genome than previously reported.

However, this approach also has its own limitation; it is limited to sequences that are already present in the finished human genome sequence, and it misses many regions of high multiplicity duplication (McCarroll *et al.*, 2008).

Instead of just detection and mapping of the CNV, there are other methods that have been developed for measuring the real changes of copy number variants. Multiplex amplifiable probe hybridization (MAPH), combines hybridization as the primary step to detect copy number with PCR to amplify the hybridized probes (Armour *et al.*, 2000). However, the MAPH method requires labour-intensive preparation of probe sets for each assay but is ideally suited to apply to large numbers of samples, which require testing at a medium number of segments, such as the 23 exons of *BRCA1*. A similar method, by sharing some of the features of MAPH, is the multiplex ligation-dependent probe analysis (MLPA) developed in 2002 to quantify 40 different DNA sequences using only 20ng of human DNA (Schouten *et al.*, 2002).

Another established method is real time quantitative PCR (RT-qPCR) which works well for scoring individual deletions and duplications, but is not generally suitable for multiplexing (Barrois *et al.*, 2004; Feuk *et al.*, 2006). FISH can be used as confirmation of the larger structural variants found by some array-based experiments (Feuk *et al.*, 2006). Microsatellite analyses are likely to show multiple lengths of alleles, up to a maximum of the

number of copies of CNV. This technique is possible as a good support for other methods in measuring copy number variants by determining the allele ratios as a measure of possible integer copy number (Hollox *et al.*, 2005; Abu Bakar *et al.*, 2009).

Armour *et al*. (2007) demonstrated the Paralogue Ratio Test (PRT) as an assay that requires as little as 10ng genomic DNA which appears to be comparable in accuracy to the other methods in determination of copy number variation (Armour *et al.*, 2007). This assay has been suggested as the first one that can provide a rapid, simple, and inexpensive method for copy number analysis suitable for application to typing thousands of samples in large case-control association studies. Different studies have used this PRT assay for typing beta-defensin copy number variation (Armour *et al.*, 2007; Abu Bakar *et al.*, 2009). Walker *et al*. (2009) have developed a multiplex paralogue ratio test (PRT) that included three PRT systems, which can be used in combination to measure copy number at the *CCL3L1* locus.

1.3.3: Methods for typing defensin copy number

Recently, there are several methods used in determination of defensin copy number, yet there is still a lack of studies on this single-locus copy number variable region, perhaps due to the complicated structure of the territory. Detection of defensin copy number was first begun with the typing of beta-defensin copy variants using MAPH with several probes across the

beta-defensin region (Hollox *et al.*, 2003), in which the investigators also used microsatellite dosage and semi-quantitative FISH analysis to further clarify the beta-defensin copy number. They also found one of the MAPH probes DEFA1.2, to be of variable copy number independently of the beta-defensin cluster. Restriction enzyme digest variant ratios (REVDR), which involved amplification across a nucleotide variant or small deletion that differs between repeats, and determination of nucleotide ratio after restriction enzyme digestion of the amplification product and quantification of the resulting electrophoresis peaks, has also been used in order to determine the copy number of defensins (Aldred *et al.*, 2005).

The MAPH assay was utilized to investigate the relationship between beta-defensin gene copy number and susceptibility to psoriasis, and the higher-throughput paralogue ratio test (PRT) (Armour *et al.*, 2007) was developed as an alternative assay for typing *DEFB4* copy number (Hollox *et al.*, 2008). Hollox and colleagues declared that combining information from MAPH, PRT and ratios of multisite variants (MSVs) improved the overall accuracy of copy number determination in this study. Recently, comparison of multiplex ligation-dependent probe amplification (MLPA) which shares some of the features of MAPH ,and real-time PCR was carried out on beta-defensin gene copy numbers quantification, with a pyrosequencing-based paralog ratio test (PPRT) used as the standard of that comparison (Perne *et al.*, 2009). The authors claimed that the consistency of copy number given between MLPA and PPRT is higher than either real-time PCR/MLPA

or real-time PCR/PPRT based on the confirmed concordance of identical results in the samples. Real-time PCR has also recently been used in two different studies to examine the beta-defensin copy number and association with Crohn's disease, which came out with different conclusions as described earlier (Fellermann *et al.*, 2006; Bentley *et al.*, 2009).

Complementary methods such as microsatellite analyses have been used to aid in clarifying beta-defensin copy number, and for tracking individual copies of a repeat through a pedigree; each repeat has a high chance of carrying a different length of allele (Hollox *et al.*, 2003). Three STRs have been used in the first part of this study to distinguish ambiguous copy number diplotypes, to deduce the haploid copy numbers and to infer the individual haplotypes based on the allele segregation (Hollox *et al.*, 2005). Short insertion-deletion (indel) analysis can also act like STRs in verification of the diploid and haploid copy number, as well as allele segregation, in which the different length of alleles can be amplified and resolved by using electrophoresis. Development of an accurate technique for typing copy number variation in a complex polymorphic region is a fundamental key for the initial understanding of the phenotypic consequences.

Reliable detection of disease association of CNVs is limited by many factors as mentioned earlier including location, size and the breakpoints of

the changes (Wain *et al.*, 2009). Wain and colleagues have reviewed the public database of these variants considering that the above factors have been very imprecise. There is improvement for the mapping from newer studies but still scalable methods to characterise copy number variation in association studies have been inexact. Moreover, potential for misclassification in copy number variation still remains much higher compared to SNPs.

1.4: Aims of the study

This study is divided into two parts; the first is examination of the origin of the human beta-defensin diversity, and the second is developing a powerful and accurate assay to investigate the association of human beta-defensin copy number variation in large case-control studies. The main goal of the first work which is elaborated in the chapter three is to measure and understand beta-defensin copy numbers variants in individuals by using the paralogue ratio test (PRT), a high throughput assay for typing of copy number variation. Subsequently, the specific and immediate objectives applied in this part are to investigate the detailed nature of variation in beta-defensins following the segregation of beta-defensin copy numbers in family pedigrees, and to examine the segregation patterns for evidence of any recombination events that occur between haplotypes, and any changes in copy number of beta-defensin in transmission from parent to children.

Developing a multiplex assay involving PRT as a powerful system for beta-defensin copy number measurement is one of the major aims for the second work in this study and has been explained in chapter four. This new system is then applied to accurately determine the beta-defensin copy number in large case-control association studies to demonstrate the advantages of this system, as the final work in this part of these studies.

Chapter 2: Materials and Methods

2.1: Materials

2.1.1: DNA samples used

2.1.1.1: ECACC Human Random Control (HRC) samples

ECACC Human Random Control (HRC) panels 1 and 2 (http://www.hpacultures.org.uk/products/dna/hrcdna/hrcdna.jsp) were used in all studies for unrelated UK Caucasian control samples. The DNA samples were extracted from lymphoblastoid cell lines derived by Epstein Barr Virus (EBV) transformation of peripheral blood lymphocytes from single donor blood samples. The genomic DNA was provided as a solution at a standard concentration of 100ng/µl in 10mM Tris-HCl buffer (pH 8.0) with 1mM EDTA. Generally a 10ng/µl DNA concentration was used in this study.

2.1.1.2: CEPH samples

From the Centre de'Etude du Polymorphisme Humain (CEPH) panel (http://ccr.coriell.org/Sections/Collections/NIGMS/CEPHResources.aspx?PgId=525&coll=GM), a set of 26 multigenerational reference families with a total of 324 members was used in most studies as European test samples. The DNA samples were obtained from different populations as shown in table 1.

Table 1: The list of CEPH families and numbers in each family included in this study. The CEPH families used in this study were derived from four different European populations; French, Venezuela, Old Order Amish and Utah pedigree.

Family description	CEPH family number	Numbers in family
CEPH/FRENCH pedigree	2	9
CEPH/FRENCH pedigree	12	13
CEPH/FRENCH pedigree	23	8
CEPH/FRENCH pedigree	35	10
CEPH/FRENCH pedigree	37	6
CEPH/FRENCH pedigree	45	10
CEPH/FRENCH pedigree	66	11
CEPH/VENEZUELA pedigree	102	16
CEPH/VENEZUELA pedigree	104	14
CEPH/OLD ORDER AMISH pedigree	884	18
CEPH/UTAH pedigree	1331	17
CEPH/UTAH pedigree	1332	16
CEPH/UTAH pedigree	1333	14
CEPH/UTAH pedigree	1334	13
CEPH/UTAH pedigree	1341	10
CEPH/UTAH pedigree	1345	13
CEPH/UTAH pedigree	1346	14
CEPH/UTAH pedigree	1350	13
CEPH/UTAH pedigree	1362	17
CEPH/UTAH pedigree	1375	12
CEPH/UTAH pedigree	1408	14
CEPH/UTAH pedigree	1416	16
CEPH/UTAH pedigree	1421	6
CEPH/UTAH pedigree	1424	14
CEPH/UTAH pedigree	13292	12
CEPH/UTAH pedigree	13294	8

2.1.1.3: Disease cohort samples

Three different disease cohorts were used to illustrate the application of the triplex system in association studies of copy number variation of β-defensin with common disease. The DNA samples were kindly supplied by

the following collaborators: two different sets of Crohn's disease samples from Christopher Mathew (King's College, London) and Jack Satsangi (University of Edinburgh) and a set of Rheumatoid Arthritis samples from Jane Worthington (University of Manchester)

2.1.2: General reagents

2.1.2.1: 10X PCR Mix

10X PCR mix contained final concentrations of 50mM Tris-HCl (pH8.8), 12mM ammonium sulphate, 5mM magnesium chloride ($MgCl_2$), 125µg/ml BSA, 7.4mM 2-mercaptoethanol and 1.1mM of each dNTP. This buffer was used for some studies that required higher concentration of $MgCl_2$ and dNTPs.

2.1.2.2: 10X LD PCR Mix

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

2.1.3: Primer design

All PCR primers were designed using the reference sequences for the region of interest from the UCSC Genome Brower March 2006 assembly (http://genome.ucsc.edu/). Primer3 software (http://frodo.wi.mit.edu/) was used to check thermodynamic properties of all the primers. Then, BLAST

(http://blast.ncbi.nlm.nih.gov) was used to search the Human DNA sequence Trace Archive for the sequence of many primers to ensure that primer pairs were unique to the sequence to be amplified and that no common sequence variants were observed.

2.2: Standard Methods

2.2.1: Polymerase Chain Reaction (PCR)

2.2.1.1: General / ABgene PCR

PCR using ABGene buffer IV was used in some studies. Usually, 20µl PCR reactions were made as a master mix with the final concentrations of 1X ABgene buffer IV, 1mM magnesium chloride ($MgCl_2$) and 0.2mM each dNTP, 0.5µM each primer, 0.05U *Taq* DNA polymerase and 10ng input DNA. Products were then usually amplified using 35-37 cycles of 95˚C for 30 seconds, 60°C for 30 seconds and 70°C for 30 seconds. Exact cycle temperatures and times depended on the primers and the expected product.

2.2.1.2: PCR in 10X LD PCR or 10X PCR mix

PCR in 10X PCR or 10X LD PCR buffer were commonly used in this study using 10µl or 20µl PCR reactions. The PCR reaction was made as a master mix with 10X PCR or 10X LD PCR buffer as described in sections 2.1.2.1 and 2.1.2.2 with the final concentration as shown above in section 2.2.1.1. However, PCR in 10X LD PCR mix was used in most studies.

Products were then amplified using 30 cycles of 95˚C for 30 seconds, annealing for 30 seconds and 70°C for 30 seconds, followed by a single "chase" phase of annealing for 1 minute/70°C for 20 minutes to reduce levels of single-stranded DNA and complete terminal 3' dA addition. Precise cycle temperatures and times depended on the primers and the expected product.

2.2.1.3: Nested PCR

Nested PCR was used in the *DEFB4* sequencing strategy to reamplify specifically only the region of interest. A standard method as described in section 2.2.1.2 was used in this PCR. However, as a secondary PCR, products were then amplified only using 20 - 25 cycles of 95˚C for 30 seconds, annealing for 30 seconds and 70°C for 30 seconds.

2.2.2: Sequencing

Generally, two sequencing reactions per DNA product were set up; one with each of the two primers, using 200μl thin-walled tubes. The reaction mix per tube contained purified template, a final concentration of 1μM primer, 1X sequencing buffer (50mM Tris-HCl (pH 9.0) and 2mM $MgCl_2$), the standard BigDye® Terminator (Applied Biosystem, UK) that contained four standard deoxynucleotides (dNTPs) and the four dideoxynucleotides (ddNTPs) each labelled with a different fluorescent dye, in a total volume of 10μl. Products were usually amplified using 25 cycles of 96˚C for 30

seconds, followed by 50˚C annealing for 15 seconds and extension at 60°C for 4 minutes. As with PCR cycles, precise cycle temperatures and times also depended on the primers and the expected product.

2.2.3: DNA Electrophoresis

2.2.3.1: Agarose gel electrophoresis

To measure the yield and size of PCR products amplified, DNA was separated by non-denaturing agarose gel electrophoresis. Agarose was dissolved by boiling in an appropriate amount of 0.5X TBE containing 0.5μg/ml ethidium bromide to give a gel in the range of 0.7-2% (w/v). DNA samples to be run were prepared in a 10% solution of 10X loading buffer (0.025% bromophenol blue dye, 40% sucrose, 2.5x TBE buffer) and then loaded into the wells. Samples were run at 120V for 1 to 2 hours; bands were visualized by illumination under UV light and a photograph of the gel was kept for records.

2.2.3.2: Capillary electrophoresis

Capillary electrophoresis was carried out in this study using an ABI Genetic Analyzer 3100 instrumentation, (Applied Biosystems, UK). GeneScan analysis was performed using fluorescently labelled DNA. 1 to 2μl of PCR products were mixed with 10μl HiDi formamide, and 2μl of the internal size standard (GeneScan-ROX500, Applied Biosystems, UK) was included for precise determination of the length of the amplicons. After denaturation for

3 min at 96°C, the products were separated on POP-4 polymer (Applied Biosystem, UK) and analysed by 3100 GeneScan 3.1 software (Applied Biosystems, UK). Scanning results of fluorescent-dye-labelled PCR products (peak area) by GeneScan software were collected and transferred to an Excel spreadsheet file through Genotyper software (Applied Biosystems, UK) for further analysis.

2.3: Paralogue ratio test (PRT)

In order to measure the copy numbers of β-defensin per diploid genome, three different systems of PRT assay were designed in this study with some modification from the one previously described by (Armour *et al.*, 2007). Table 4 below shows the list of primers designed to amplify from different locations inside the β-defensin region.

Table 2: PCR primer sequences from three systems of PRT assay used to amplify different locations in the beta-defensin repeat only.

Primer Name	Sequence (5' - 3')	Size resolved in ABI capillary electrophoresis
HSPD5.8	Forward: CCAGATGAGACCAGTGTCC Labelled reverse: TTTTAAGTTCAGCAATTACAGC	After *Hae*III digestion: Chromosome 8 (302bp) Chromosome 5 (315bp)
PRT107A	Labelled forward: AGCCTCATTTAACTTTGGTGC Reverse: GGCTATGAAGCAATGGCCTA	Chromosome 8 (157bp) Chromosome 11 (155bp)
HSPD21	Forward: GAGGTCACTGTGATCAAAGAT Labelled reverse: AACCTTCAGCACAGCTACTC	Chromosome 8 (172bp) Chromosome 21 (180bp)

2.3.1: HSPD5.8

HSPD5.8 was the first system for PRT assay used in this study to measure the copy numbers of *DEFB4* (Armour et al., 2007). A pair of primers for HSPD5.8 as shown in table 2 was designed to amplify simultaneously the copy of a heat-shock protein pseudogene near *DEFB4* on chromosome 8, and one other copy on chromosome 5. Products were amplified using PCR in 10X LD PCR mix as described in section 2.2.1.2 with 30 cycles of 95°C for 30 seconds, 53°C for 30 seconds and 70°C for 30 seconds, followed by a single "chase" phase of 53°C for 1 minute/70°C for 20 minutes to reduce levels of single-stranded DNA and to allow terminal 3' dA addition to proceed to completion. Two parallel amplifications were carried out for each sample, one with a FAM label, the other with HEX. However, the (test) chromosome 8 (*DEFB4*) and (reference) chromosome 5 copies give PCR products too close in size (443 and 447bp respectively) to separate reliably by capillary electrophoresis. Therefore, digestion with restriction enzyme *Hae*III to give products of 302 and 315bp was carried out so that the two products could easily be distinguished and quantified. Therefore, 1µl of each PCR product was added, without further purification, to a 10µl digestion containing 1x ReAct 2 buffer (Invitrogen) and 5U *Hae*III. After incubation at 37˚C for 4-16 hours, 2µl were added to 10µl HiDi formamide with R-500 marker (Applied Biosystems), and analysed by capillary electrophoresis as described in section 2.2.3.2.

2.3.2: PRT107A

PRT107A was the second system of PRT assay developed to aid in clarification of *DEFB4* copy numbers from HSPD5.8. PRT107A primers were designed to amplify simultaneously a dispersed repeat copy near *DEFB107* on chromosome 8, and one other copy on chromosome 11. Although the PCR product formed by (test) chromosome 8 (*DEFB107*) and (reference) chromosome 11 copies were close in size, only differing by 2bp (157 and 155bp respectively), it was possible enough to separate them reliably by capillary electrophoresis. Products were amplified using PCR with 10X LD PCR mix as described in section 2.2.1.2 with pre-denaturation of 95˚C for 5 minutes, followed by 22 cycles of 95˚C for 30 seconds, 58°C for 30 seconds and 70°C for 1 minute, followed by a single “chase” phase of 58°C for 1 minute and 70°C for 40 minutes. As for HSPD5.8, two parallel amplifications were carried out for each sample, one with a FAM label, the other with HEX; without any further digestion 1µl from each reaction was added to 10µl HiDi formamide with R-500 marker (Applied Biosystems), and analysed by capillary electrophoresis as described in section 2.2.3.2.

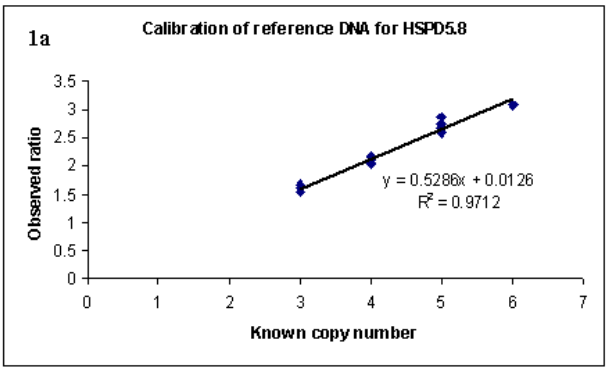
2.3.3: HSPD21

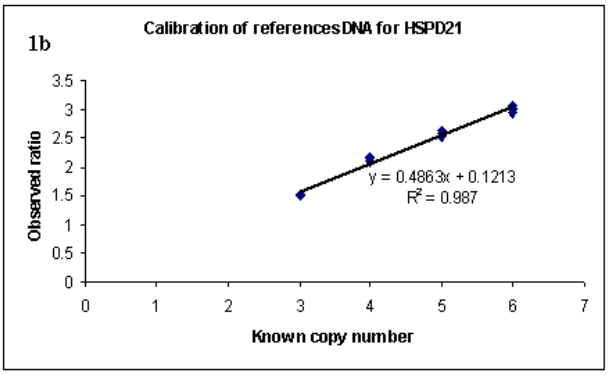
HSPD21, the third system of PRT assay, was also developed to aid in clarification of *DEFB4* copy numbers. HSPD21 primers were designed to amplify simultaneously the copy about 3kb upstream of *DEFB4* on chromosome 8 and one other copy on chromosome 21. In this study,

*HSPD*21 was carried out only in the development of triplex system as explained in 2.9. As with the other PRT systems, peak areas corresponding to the 172bp from near *DEFB4* and the 180bp from chromosome 21 were recorded for both FAM- and HEX/NED- labelled products using GeneScan and Genotyper software (Applied Biosystems).

2.3.4: Data analysis

For HSPD5.8, peak areas corresponding to the 302bp *Hae*III fragment from near *DEFB4* and the 315bp fragment from chromosome 5, peak areas corresponding to the 157bp from near *DEFB107* and the 155bp from chromosome 11 for PRT107A and peak areas corresponding to the 172bp from near *DEFB4* and the 180bp from chromosome 21 in HSPD21 were recorded using GeneScan and Genotyper software (Applied Biosystems). The ratio 302bp/315bp, 157bp/155bp or 172bp/180bp for HSPD5.8, PRT107A and HSPD21 respectively were compared between FAM- and HEX- or NED- labelled products, and in most studies results were accepted if the difference between the ratios was less than 15% of their mean; this criterion led to the rejection of about 10% of tests. If accepted, the mean of the FAM and HEX or FAM and NED ratios was used in further analysis. Mean ratios were used in conjunction with reference samples of known copy number to calibrate each experiment, and the resulting (least-squares) linear regression used to infer the copy numbers for unknown samples. Selected DNA samples giving reproducible results from several PRT assays were used as calibration standards (Figure 4a, 4b and 4c).





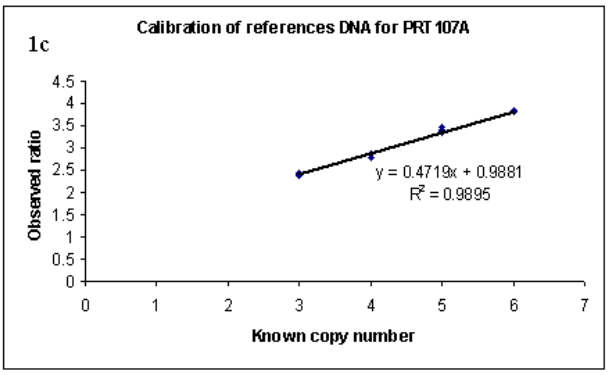


Figure 4: An example of the calibration standard on selected reference DNA samples, which give reproducible results from the PRT systems. The linear regression shown above has been used to infer the copy numbers for unknown samples. Figure 4a) was the reference DNA standard calibration for HSPD5.8, figure 4b) was the reference DNA standard calibration for HSPD21 and figure 4c) was the reference DNA standard calibration for PRT107A.

2.4: Microsatellite analysis

Whenever the PRT assays indicated the copy numbers of β-defensin in diploid genomes from CEPH family members, the segregation of variable copy number haplotypes from parents to children could be inferred by comparison with genotype data from other local markers drawn from the well-established CEPH genotype database (http://www.cephb.fr/en/index.php). Instead of just using PRT copy number measurement methods, three microsatellite analyses were chosen to clarify the segregation of the copy numbers of β-defensin for haplotype analysis, and all of the primers are shown in table 3. To genotype microsatellites EPEV1, EPEV2 and EPEV3 in the *DEFB4* region, 10ng of genomic DNA was amplified separately with three pairs of primers as shown in table 5 using PCR in 10X LD PCR mix as described in section 2.2.1.2. PCR amplification used different annealing temperatures, 58°C for EPEV1, 60°C for EPEV2 and 56°C for EPEV3 over 25 cycles of 95˚C for 1 minute, annealing temperature for 1 minute and 72°C for 1 minute, and final extension incubation of 72°C for 20 minutes to complete 3' dA addition. After that, 1.5µl were added to 10µl HiDi formamide with R-500 marker (Applied Biosystems), and analysed by capillary electrophoresis as described in section 2.2.3.2. Since labelled forward primers of EPEV1 and EPEV3 used the same fluorescent dye but produced different size ranges of alleles, both could be resolved in the same electrophoresis. For EPEV2, consideration of severe slippage as explained below meant that it was only

carried out on some of the CEPH families that needed further investigation and clarification of the recombination status.

Table 3: PCR primer sequences from three microsatellites assays used to amplify different locations in the beta-defensin repeat

Primer Name	Sequence (5' - 3')	Annealing temperature	Size resolved in ABI capillary electrophoresis
Microsatellite 1 (MSAT3/EPEV1) (Hollox et al. 2003)	Labeled forward: GGCAGTATTCCAGGATACGG Reverse: GAACAATTAGATATCCCTATGC	58°C	167bp to 191bp
Microsatellite 2 (MSAT4/EPEV2) (Hollox et al. 2005)	Labeled forward: GCACCAGAGACCTCATGTTTTC Reverse: GTAACTTACAGTTGAAAACCAC	60°C	241bp to 269bp
Microsatellite 3 (MSAT6/EPEV3) (Abu Bakar et al. 2009)	Labeled forward: GATACTGTGAACTACAGATCAC Reverse: CTGCCCTGATTCAGTATTGAAC	56°C	127bp to 151bp

2.4.1: Data analysis

The most commonly encountered problems during microsatellite analysis are poor or non-specific amplification, incomplete 3'-dA nucleotide addition and ambiguity caused by the "stutter" phenomenon caused by PCR slippage. Although in this experiment, PCR amplification produced very specific amplification in the expected size range and almost complete 3'-dA nucleotide addition, very careful attention was given to the stutter peaks caused by polymerase slippage during elongation. The stutter peaks that might occur during the PCR amplication of di-, tri- and tetranucleotide microsatellite loci appeared as minor products that are 1-10 repeat units shorter than the main allele product. In microsatellite analysis of EPEV1, EPEV2 and EPEV3 in the *DEFB4* region, successful PCR amplification

produced a stutter peak just smaller than the main allele and was generally about 10-20% of the main allele peak area. After slippage patterns were recognized in each of the EPEV1, EPEV2 and EPEV3 analyses, assuming that slippage affects all alleles to the same degree then the effect of slippage in two adjacent alleles could be accounted for by quadratic equations for 3 peaks: $x = a+1/2(b-\sqrt{(b^2-4ac)})$, $y = c+1/2(b+\sqrt{(b^2-4ac)})$ where a, b and c are the observed peak areas, and x and y the corresponding original peak areas (without slippage), as shown in figure 5. Simulations are used to reconstruct the original peak areas where there were 4 or 5 adjacent peaks.

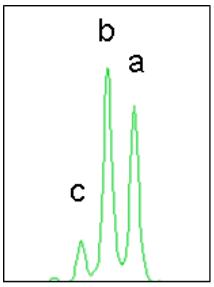


Figure 5: An example of 3 peaks from two alleles, with ‘c’ corresponding to a stutter peak. The quantities x and y are the inferred true peak areas attributable to the two alleles.

For example in EPEV1 (Figure 6), peak areas in the range 180bp and 190bp were recorded using GeneScan and Genotyper software (Applied Biosystems). The major slippage peaks were always seen 2bp (i.e. one repeat unit; 178bp and 184bp in figure 6) shorter than the main allele

peaks, as expected for this dinucleotide repeat. Second-order slippage products were neglected in this analysis. Meanwhile for EPEV3 (figure 6), there were 3 main peaks; 136bp, 140bp and 142bp were recorded and there was a minor product (138bp) from those 2 main peaks resulting from slippage.

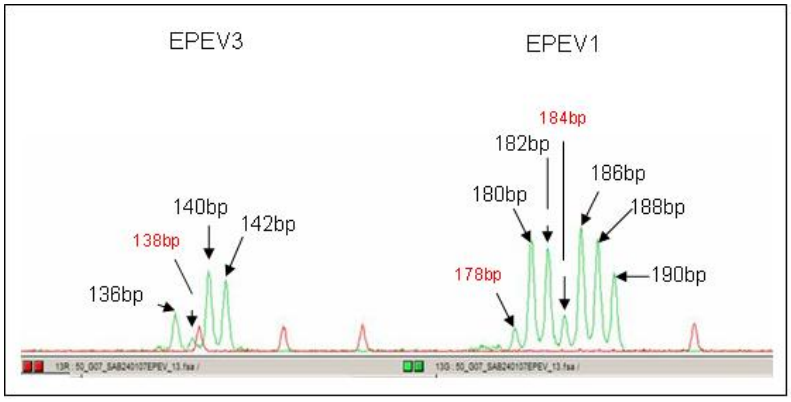


Figure 6: An example of the GeneScan electropherogram showing EPEV1 and EPEV3 resolved in the same electrophoresis. This sample had five copies of *DEFB4* from the PRT assay, and five primary microsatellite alleles are visible at EPEV1 (180bp, 182bp, 186bp, 188bp and 190bp) and there are three primary peaks at EPEV3 (136bp, 140bp and 142bp), two of which correspond to two copies each (140bp and 142bp).

2.5: Indel ratio measurements

For further clarification of the copy numbers of β-defensin in diploid genomes indicated by PRT assays, an insertion deletion (indel) ratio measurement assay was developed in this study. This assay also helped

the unambiguous inference of segregation of copy numbers from parents to children in CEPH family members.

Table 4: PCR primer sequences from the 5bp indel ratio assay (rs5889219) used to amplify different lengths of the same location in the beta-defensins. All used the same forward primer.

Primer Name	Sequence (5' - 3')	Annealing temperature	Size resolved in ABI capillary electrophoresis
5DELI	Labeled forward: AAACCAATACCCTTTCCAAG Reverse: TCTTTTGTTTCAGATTCAGATG	50°C	119bp to 125bp
5DELII	Labeled forward:AAACCAATACCCTTTCCAAG Reverse: CCCCAATTCATTAGGGTTTTT	50°C	167bp to 172bp
5DELIV	Labeled forward: AAACCAATACCCTTTCCAAG Reverse: TTCTCTTTTGTTTCAGATTCAGATG	50°C	123bp to 128bp

2.5.1: rs5889219 (5DEL)

A 2bp/5bp deletion at chr8:7,363,859-7,364,008 about 10kb downstream of *DEFB107* in the β-defensin gene cluster was chosen to verify the copy numbers and the segregation of those copies in pedigrees. Therefore, 10ng of genomic DNA was amplified with the primers shown in table 4 using PCR in 10X LD PCR mix as described in section 2.2.1.2. PCR amplification used similar annealing temperatures, 50°C over 22 cycles of 95°C for 30 seconds, annealing temperature for 30 seconds and 70°C for 30 seconds, followed by a single "chase" phase of 58°C for 1 minute and 70°C for 20 minutes to complete 3' dA addition. As for other systems, two parallel amplifications were carried out for each sample, one with a FAM

label, and the other with HEX; then 1µl from each reaction was added to 10µl HiDi formamide with R-500 marker (Applied Biosystems), and analyzed by capillary electrophoresis as described in section 2.2.3.2.

2.5.2: 9 indel

In order to verify a 9bp insertion/deletion found in *DEFB4* by sequencing, another indel ratio measurement assay was developed. Products of 308bp and 298bp in size were amplified using FAM-labeled primer 9DELF (CCAAATGGAAGAATGGCGTA) and primer 9DELR (GTCCATTGGGTTCTCAAACT) with PCR in 10X LD PCR mix as described in section 2.2.1.2. Products were amplified using 23 cycles of 95˚C for 30 seconds, 50°C for 30 seconds and 72°C for 1 minute, followed by a single "chase" phase of 50°C for 1 minute/72°C for 20 minutes to reduce levels of single-stranded DNA. There was only one amplification with a FAM label carried out for each sample; then 1µl from the reaction was added to 10µl HiDi formamide with R-500 marker (Applied Biosystems), and analyzed by capillary electrophoresis as described in section 2.2.3.2.

2.5.3: Data analysis

All peak areas corresponding to the 5DELI, 5DELII and 5DELIV, and 9DEL size ranges as mentioned earlier in table 4 and 2.5.2 respectively were recorded for both FAM- and HEX- labelled products using GeneScan and

Genotyper software (Applied Biosystems). A program to analyse copy numbers consistent with three peaks obtained from 5DEL assays was created by John Armour (Institute of Genetics, University of Nottingham) to evaluate a "squared difference score" $(R-X)^2$, where R is the measured ratio obtained from the peak area and X is the expected ratio being tested as explained in section 2.10.

2.6: Restriction fragment length polymorphism (RFLP) at rs2203837

Because of anomalous segregation in CEPH family 104 indicated by database genotypes, rs2203837 was specifically chosen to verify the genotypes in this family. PCR used forward primer CTTGGGGCTATGCATTGAGT and reverse primer TGTATGCACATACTGCCAACAC in general/ABgene PCR buffer as described in section 2.2.1.1 with 37 cycles of 95˚C for 30 seconds, 60°C for 30 seconds and 70°C for 30 seconds. After that, the product (290bp) was digested with restriction enzyme *Mse*I. The resulting DNA fragments were separated by agarose gel electrophoresis as described in the 2.2.3.1 section. Fragments of 247, 16 and 27bp appear indicate C and fragments of 187, 60, and 27bp indicate T.

2.7: Linkage analysis

Linkage analysis of markers on chromosome 8 was performed by John Armour using the CHROMPIC option of CRIMAP (Lander and Green,

1987) and corresponding physical location inferred against the non-inverted orientation of the region as represented in the UCSC March 2006 Genome Assembly (hg18, NCBI Build 36). CRIMAP with the CHROMPIC option was used in this study to determine the parental origin of the children's allele for each marker, as deduced from the available genotype information. Complete documentation about this analysis can be obtained from http://linkage.rockefeller.edu/soft/crimap/chrompic53.html.

2.8: Sequencing strategy

To follow up evidence of a dual location of the beta-defensins from the recombinations found in some CEPH families and verified by different assays, sequence analysis was carried out to compare sequences at different locations. Therefore, two gene loci in the beta-defensin cluster were chosen for sequence analysis in this study, *DEFB103* and *DEFB4*.

2.8.1: *DEFB103* locus

In order to successfully sequence the whole region of the *DEFB103* locus, primers were designed as shown in table 5. Contrasting with the *DEFB4* locus, the full sequence of *DEFB103* was amplified using only a single pair of primers to give a 1.8kb product; other internal primers were used to complete the sequence analysis. About 100ng of product was amplified using forward DEFB103F and reverse DEFB103R primers with PCR in 10X LD PCR mix as described in section 2.2.1.2. PCR amplification was carried out for 35 cycles of 95°C for 30 seconds, 60°C for 30 seconds and

70°C for 4 minutes. Products then were separated on a 0.8% agarose gel to ensure the correct size has been amplified (2.2.3.1). AMPURE® PCR purification was performed following the manufacturer's standard protocol (Agencourt Bioscience Corporation, UK) to remove unincorporated primers, dNTPs, DNA polymerases and salts used that can interfere with sequencing. Sequencing was carried out as described in section 2.2.2 using the primers in table 5. CLEANSEQ® Dye-terminator removal was carried out prior to sending for sequence analysis. The electrophoresis of automated DNA sequence reading was performed at DBS Genomic (Durham University, School of Biological and Biomedical Sciences).

Table 5: PCR and sequencing primer sequences used for sequencing of the *DEFB103* locus

Primer Name	Sequence (5' - 3')
DEFB103F	**GCAAAGGGCATATTCCAAGA**
DEFB103F2	**TTTCTTCGGCAGCATT**
DEFB103R	**AAAAGCTCCAAAGCAGCAAA**
DEFB103R2	**GGACCAAGCAGGTTTGTTGT**
DEFB103R3	**CTTTCCCCAACTCTTCAAGG**

2.8.2: *DEFB4* locus

Because a single-step PCR did not yield sufficient product from the *DEFB4* locus, two stages of PCR amplification were performed. Primary PCR amplification was carried out first to produce a 2.6kb product using forward DEFB4F and reverse DEFB4R primers. Then, two pairs of primers were

used in secondary PCR to specifically amplify two overlapping shorter regions inside the first product. The first short region was amplified using forward DEFB4Fb and reverse DEFB4R2 primers (1.8kb in size) and the other region was amplified using forward DEFB4F2 and reverse DEFB4R primers (1.6kb in size). Both of the above PCR amplifications were carried out using PCR in 10X LD mix as described in section 2.2.2.1. The first product was amplified with 35 cycles of 95°C for 30 seconds, 60°C for 30 seconds and 70°C for 4 minutes. After that, the secondary PCR was carried out as described in section 2.2.1.3. Products were then analysed on a 0.8% agarose gel. AMPURE® PCR purification and CLEANSEQ® Dye-terminator were performed as for *DEFB103* (2.8.1).

Table 6: PCR and sequencing primer sequences used for sequencing of the *DEFB4* locus

Primer Name	Sequence (5' - 3')
DEFB4F	**GCAGGTGAGTGGCAGGTTAT**
DEFB4Fb	**ATTTTCTGGTCCCAAGAGCA**
DEFB4F2	**CAGAGGCCCTGAGAACAGTC**
DEFB4R	**GGAAAATGCCGGGAATAAAT**
DEFB4R2	**TGAAGGTTAGGCATCCAGGT**
DEFB4InsF	**TTTAGACTGAGTAGACTGAATGC**
DEFB4IndR	**CTCTGGTGCCTCTTCAGAACC**
DEFBInsR	**ATTCAGTCTACTCAGTCTAAAAGTG**

2.9: Triplex system

The PRT systems developed to measure the copy number variation in this study were combined in a triplex system. Three systems had successfully measured defensin copy number: two PRT systems, PRT107A and HSPD21, and 5DELR4 as an indel measurement system. All of these three systems were chosen because they produced different sizes of products when resolved in ABI capillary electrophoresis. Both PRT systems were amplified in one PCR with slight modification made to the PCR mixture and cycle. Both products were amplified using PCR in 10X LD PCR mix as described in section 2.2.1.2 using PRT107A PCR cycle as described in section 2.3.2. Meanwhile, 5DELIV indel measurement ratio was performed as a separate PCR amplification as described in 2.5.1 section using 5DELIV primers (table 6). Routinely, two parallel amplifications were carried out for each system in every sample, one with a FAM or NED label, the other with HEX or NED; then 0.5µl to 1µl from each reaction were added together to 10µl HiDi formamide with R-500 marker (Applied Biosystems), and analyzed by capillary electrophoresis as described in section 2.2.3.2.

2.9.1: Data analysis

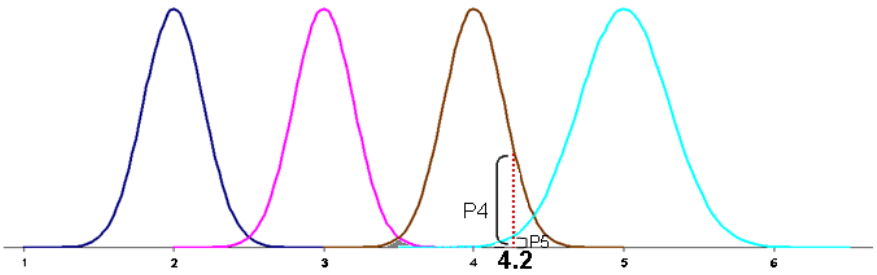
All peak areas or peak heights corresponding to HSPD21, PRT107A and 5DELIV (peaks at 172bp and 180bp, 155bp and 157bp, and 123bp to 128bp respectively) were recorded for both FAM/NED- and HEX/NED-labelled products using GeneScan and Genotyper software (Applied

Biosystems). Copy number of beta defensin was inferred using maximum likelihood analysis as described in section 2.10. Meanwhile, mean ratios for both PRT systems were used in conjunction with reference standards to calibrate each experiment as described in section 2.3.4.

2.10: Maximum Likelihood Analysis

A program written in C++ was created by John Armour (Institute of Genetics, University of Nottingham) to implement compound likelihood analysis. To run this programme, a set of results was used as an input file to evaluate the individual probability of getting the PRT or 5DEL results assuming each possible copy number from 2 to 9, and then to combine the likelihood results (by multiplying the results together for each copy number for each assay) to get the overall probability of the complete set of results for each of the possible copy numbers 2 to 9. The highest probability in each group is scaled to 1, and all other values expressed relative to that. The probabilities for each of the copy numbers 2 to 9 are calculated from PRT data assuming a normal distribution with a mean of the integer copy number and a standard deviation based on empirical observation (figure 7). However, the 5DEL data were evaluated by computing a “difference score” $(R-X)^2$, where R is the measured ratio and X is the expected ratio being tested. Where there are two ratios (i.e. three peaks), the program uses the sum of the two difference scores. The reciprocals of the difference scores (so that the lower the difference score, the higher the probability) are then taken to be proportional to the probability.

a)



b)

Relative likelihood values for

Best copy no.	Test analysis	N=2	N=3	N=4	N=5	N=6	N=7	N=8	N=9
4	**PRT 4.2**	4.35E-51	4.17E-11	1	0.001241	0.003428	5.46E-11	1.70E-11	4.52E-14

Figure 7: An example of normal distributions for copy numbers of two (blue line), three (pink line), four (brown line) and five copies (light blue line). The dotted red line shows the estimated copy number from an example PRT assay, 4.2. Based on the assumption of normal distribution for each copy number, the most probable copy number for 4.2 is 4 copies. The table below (7b) shows that a copy number of N=5 (P5) is 0.001241 times as likely (i.e. about 800 times less likely) when compared to the probability of a copy number of N=4 (P4).

Chapter 3: Origins of Diversity at Human Beta-Defensin Copy Number

3.1: Background of information

Many recent studies have given most attention to the detection of copy number variation either as a whole genome study or on certain genes varying in copy number, and use a variety of assays to measure this variation in diploid genomes. Nevertheless, very few studies involve measuring haploid copy number and investigation of the origin of this haploid copy number variation. Consequently, there is very little known about the mechanism of generation of this copy number variation. This work aims at measurement of human beta defensin copy number variation in diploid and haploid genomes, and determination and characterization of the parental origin for the copy number variants, to understand the evolutionary processes that generate and maintain this variation in human populations.

Meiotic recombination plays an important role in creating haplotype diversity in the human genome and has the potential to cause genomic rearrangements by nonallelic homologous recombination (NAHR) also known as ectopic recombination, and through changes triggered by recombination-initiation events, such as nonhomologous end joining (NHEJ) (Shaw and Lupski, 2004; Holloway *et al.*, 2006; Turner *et al.*,

2008). Giglio *et al.* (2001) reported that the beta defensin genes (at least 250kb) involved in copy number variation are found on a large genomic repeat unit within a more complex copy number variation involving retroviral elements and olfactory repeat (OR) regions, collectively known as “REPD” (for repeat distal). In addition Sugawara *et al.* (2003) found another smaller OR region called “REPP” (repeat proximal) about 5Mb proximal on 8p23.1 and sharing a high level of identity with REPD. The beta defensin cluster at chromosome 8p23.1, including *DEFB4*, *DEFB103*, *DEFB104*, *DEFB105*, *DEFB106*, *DEFB107*, but excluding *DEFB1*, are shown to vary in normal controls commonly between two to seven copies in diploid genome (Hollox *et al.*, 2003; Hollox *et al.*, 2005). Multiplex amplifiable probe hybridisation (MAPH) and semiquantitative fluorescence *in situ* hybridisation (SQ-FISH) found that cytogenetic duplications of 8p23.1, regarded as euchromatic variants (EVs) in a previous study (Barber *et al.*, 1998), had chromosomes carrying high copy number alleles between nine and twelve copies (Barber *et al.*, 2005).

Usually for any gene present in two copies as shown in figure 8, one copy has been inherited from each parental genome at fertilization. However, for a gene present as four copies such as gene A (see figure 8), deducing haplotype copy number from each parental genome is not as simple as inferring haplotype from two copies. Any combination of haploid copy number, (one and three, two and two or zero and four) corresponding to the total number of four copies could be present.

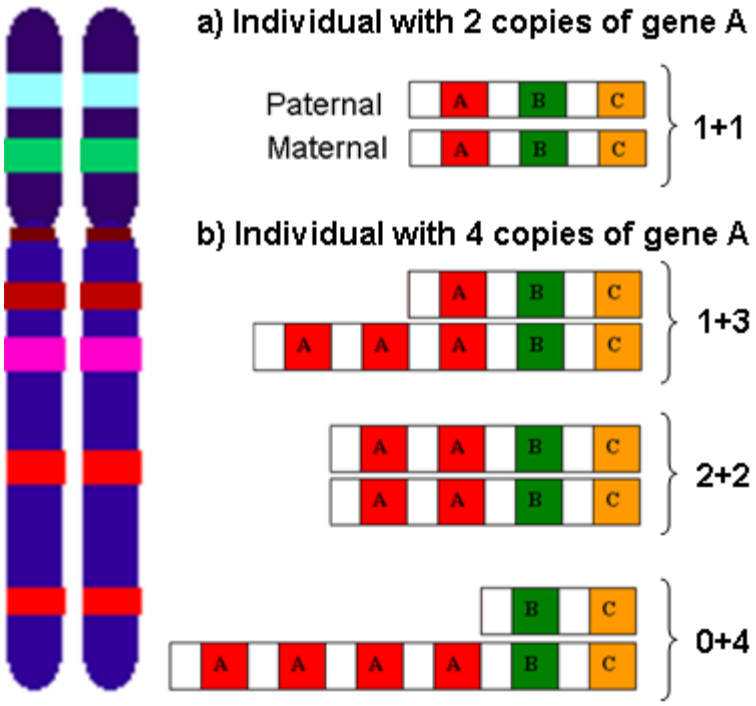


Figure 8: An example of two chromosomes that represent reference paternal and maternal chromosome. In a) the figure shows an example of an individual that has two copies for gene A (coloured in red). In the normal situation we assume that one copy has been transmitted from each parent, so for this individual, one copy has been inherited from father and one copy from mother. In b), the figure shows an example of three different possibilities of haplotype combination deduced for gene A present as 4 copies in an individual.

In order to achieve the aims of this study, twenty-six multigenerational CEPH reference families have been used as described in section 2.1.1.2. The PRT assays at two different places, HSPD5.8 and PRT107A, combined with analysis of variant ratios at microsatellites EPEV-1, EPEV-2

and EPEV-3, and the multiallelic length polymorphism including the indel rs5889219, were used to type DNA from family members. The PRT assays were designed to measure the variable copy number per diploid genome. The microsatellite and multiallelic length polymorphism analyses were used to investigate the identity and parental origin of copy number haplotypes.

3.2: Results

3.2.1: Diplotype analyses

- Paralogue ratio test (PRT assays)

An accurate assay was developed for copy number beta defensin region. Based on the principle that the test and reference amplicons should be as similar as possible, the problems of accuracy and reproducibility associated with multiplex PCR might be avoided. Deutsch *et al*. (2004) recently successfully exploited this principle to the diagnosis of trisomy when they found that some sequences present on chromosome 21 (for example) have nearly identical paralogues at another site in the genome. Therefore, a single pair of primers can be used to amplify both test and reference loci and distinguish them via minor differences of internal (reference) sequences.

From this, Armour *et al*. (2007) developed a Paralogue Ratio Test (PRT) which was able to exploit the same advantages of amplifying both test and

reference loci with a single primer pair. They designed precisely-placed primers to amplify from copies of a diverged (low copy number) repetitive sequence at the copy-variable locus and at exactly one other (reference) location. As shown as in figure 9, a heat-shock protein pseudogene of ~2kb (HSPDP3) is found ~2kb upstream of the *DEFB4* gene and at 10 locations elsewhere in the genome. Figure 10 shows a description of the PRT107A assay, amplifying a dispersed repeat copy near to *DEFB107* gene and at several locations elsewhere in the genome. Then primers that matched the copy on chromosome 8 and one other copy on the reference chromosome were designed exactly so that copies at other chromosomal locations had mismatches. As described in chapter two, section 2.3.1, 2.3.2 and 2.3.4, both products were detected without other detectable fragments under the conditions used.

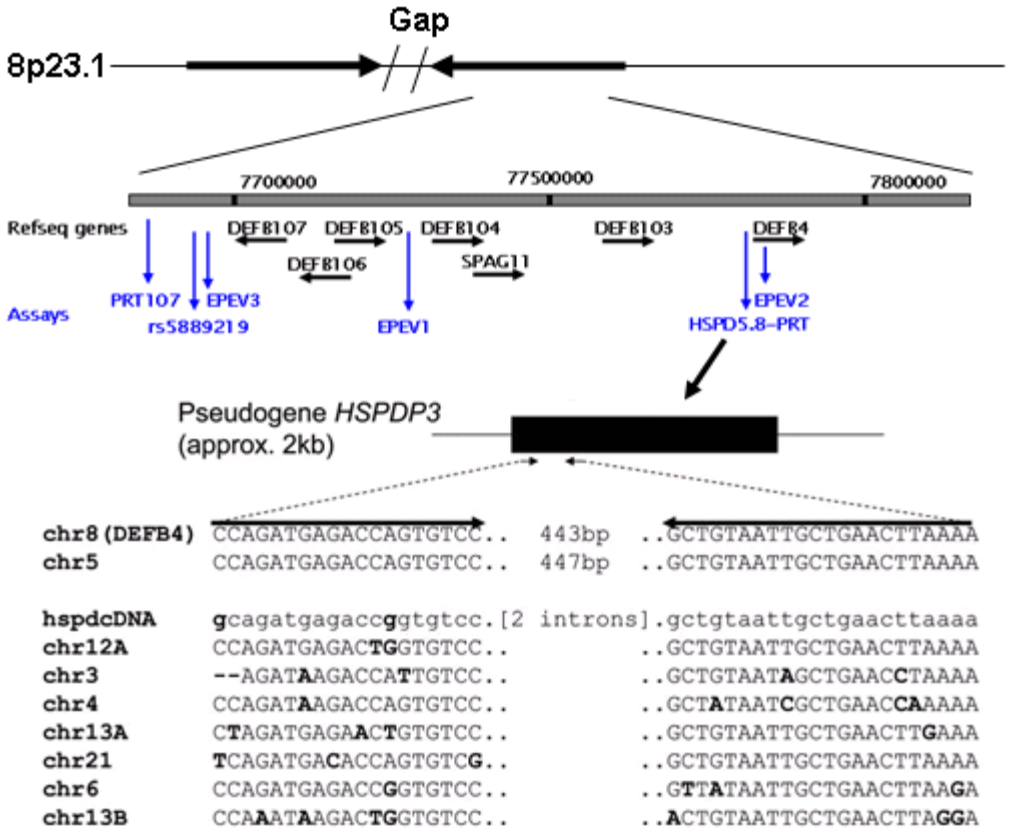


Figure 9: Principle of the HSPD5.8 assay at *DEFB4*. The top line shows the general structure of the repeat unit containing beta defensin genes (which has two inverted rather than tandem repeats in the March 2006 assembly). The middle panel shows the location of genes *SPAG11*, *DEFB4* and *DEFB103-07* (coloured in black), together with the locations of assays (coloured in blue) used in this study. In the detailed display at the bottom, the primers amplify products from *HSPDP3* pseudogene upstream of *DEFB4* on chromosome 8, and from a reference copy on chromosome 5, but have multiple mismatches with other copies of the element. In this way a single primer pair can be used to amplify two very similar products, one from near *DEFB4*, the other from chromosome 5.

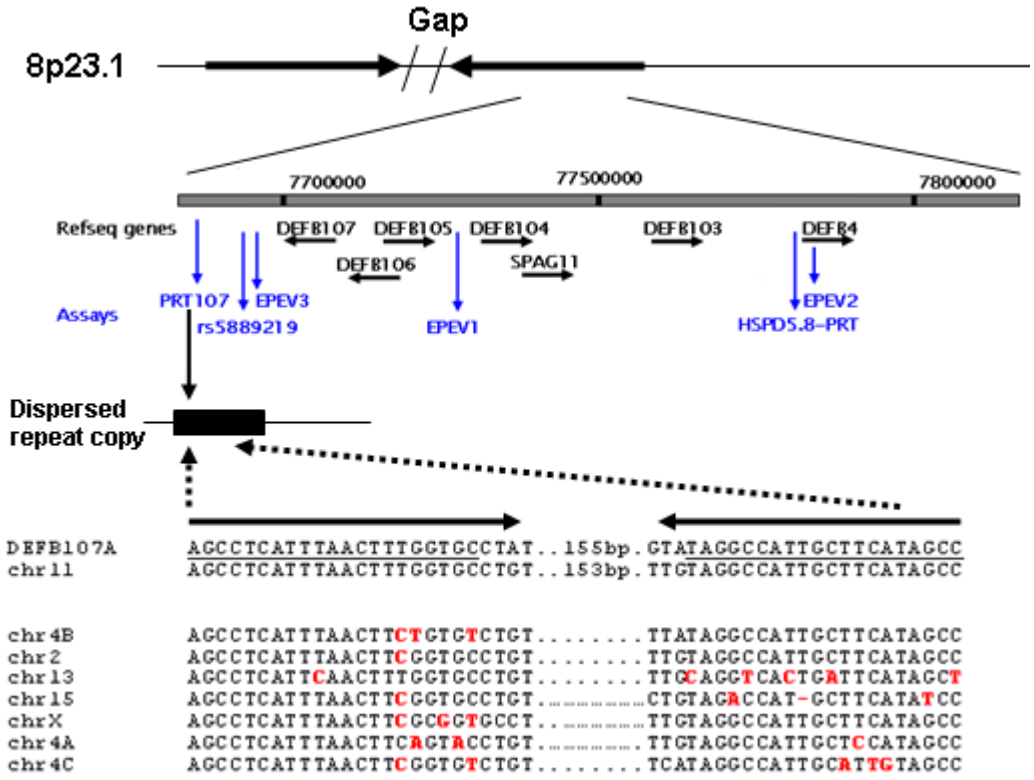


Figure 10: The principle of the PRT107A assay at *DEFB107* on chromosome 8, and from a reference copy on chromosome 11, with multiple mismatches with other copies of the element. In this way a single primer pair can be used to amplify two very similar products, one from near *DEFB107*, the other from chromosome 11.

Beta-defensin copy numbers were first determined by the HSPD5.8 PRT assay in 324 individuals from twenty-six CEPH families, including 63 unrelated individuals. As shown in figure 11, consistent with previous data (Hollox *et al.*, 2003; Hollox *et al.*, 2005), beta-defensin copy numbers were found to vary between two copies to eight copies per diploid genome with four to six copies being the most common.

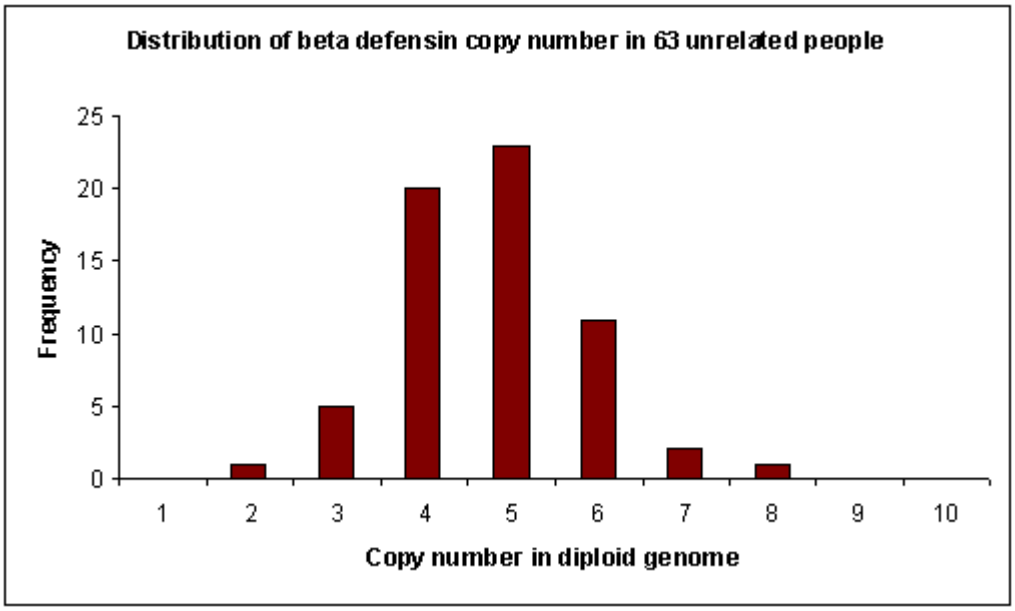


Figure 11: Distribution of beta defensin copy numbers per diploid genome measured by HSPD5.8 PRT assay in 63 unrelated individuals collected from the whole set of CEPH families.

3.2.2: Haplotype analyses

3.2.2.1: Microsatellite analysis

Microsatellites are tandemly repetitive stretches of DNA in which a short motif, from 1 to 5 nucleotides long, is repeated several times. They are common, easy to identify from genomic databases, and extremely polymorphic in the human genome. Microsatellite alleles are generally detected as DNA fragments of different sizes obtained after amplification with primers flanking the microsatellite region. Because of some problems, for example poor or non-specific amplification, incomplete 3'-A nucleotide addition and ambiguity caused by the "stutter" phenomenon caused by PCR slippage, microsatellite allele identification must be carried out after

very careful analysis. As PCR amplification produced very specific amplification in the expected size range and almost complete 3'-A nucleotide addition, very careful attention was given to the stutter peaks caused by polymerase slippage during elongation as described in the chapter two, section 2.4.1. Below is an example for complex slippage patterns in EPEV2 which is not shown in chapter 2.

For EPEV2, peak areas from 213bp to 264bp were recorded using GeneScan and Genotyper software (Applied Biosystems). However, in EPEV2, which has a complex internal sequence structure including repeat units of different length, a different approach of analysis to EPEV1 and EPEV3 has been carried out based on the peak patterns, sizes and slippage peaks. Very careful attention was taken to recognise and match the pattern because slippage peaks patterns had a complex relationship to the real peak pattern, up to 10-14bp different from the original peak as shown in figure 12.

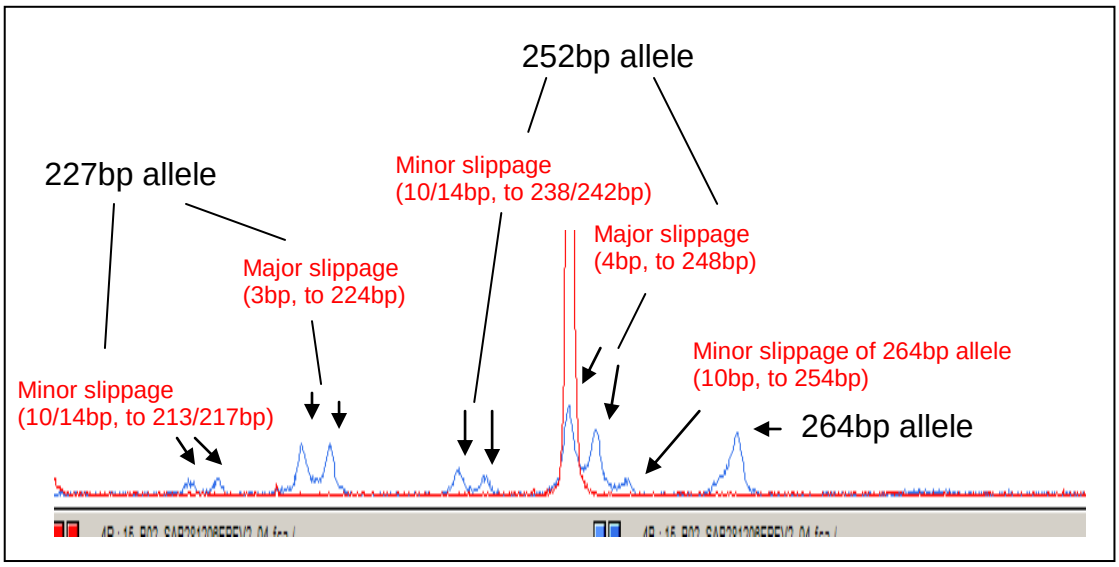


Figure 12: An example of the GeneScan electropherogram showing the stutter peaks and the main peaks in EPEV2. In this case there are four repeat units, one each containing alleles of 227bp and 264bp, and two containing 252bp alleles. The 264bp allele appears to lack a major (3/4bp) slippage product.

In this study, microsatellites within this variable region were typed in CEPH families in order to verify the beta-defensin copy numbers which were indicated from the PRT assays. Therefore, microsatellite EPEV1 and EPEV3 were typed in 324 individuals (all the CEPH families) and microsatellite EPEV2 was typed for further analysis in only selected CEPH families. Microsatellite analysis was expected to give results consistent with the copy numbers from the PRT assay. The microsatellites could be used to confirm copy number measurements; for example, amplification of four variants with yields in the ratio 1:1:2:1 strongly suggests a copy

number of 5. For an example, see the microsatellite EPEV3 profile of 141601 in figure 15. Besides acting as a verification method for the PRT assay in determining the beta defensin copy numbers, these three microsatellites were also used to distinguish different copies of beta defensin repeats in segregation from parent to children, to determine the haplotype transmission. Figure 13 below shows the distribution of beta defensin copy number per haploid genome for 63 unrelated individuals (126 haplotypes) inferred from microsatellites EPEV1 and EPEV3. The frequency of beta-defensin copy numbers were found to vary between one copy and five copies per haploid genome, with two and three copies being the most common.

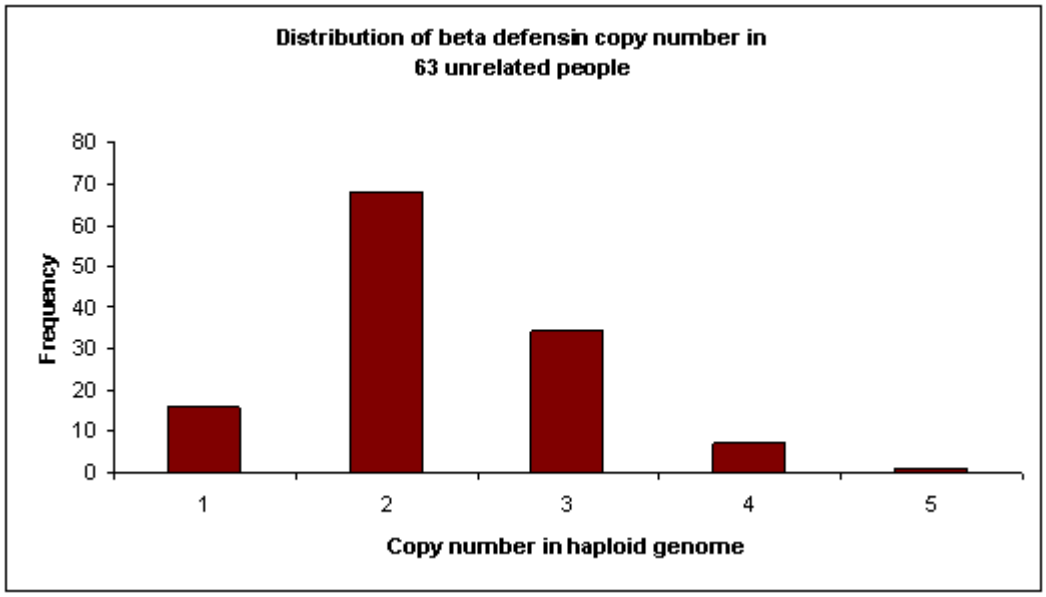


Figure 13: Distribution of beta defensin copy numbers for 63 unrelated individuals (126 haplotypes) in CEPH families per haploid genome from microsatellites EPEV1 and EPEV3.

To follow haplotypes of beta defensin copy number variants from parents to children, segregation of the corresponding region of 8p23.1 from parents to the children in a set of 26 multigenerational CEPH families (Table 1) was first inferred with information from other local genetic markers typed in these same families, using the well-established CEPH genotype database. Below is one example of pedigree analysis on the AFM304ze9 marker for CEPH family 1416 (Figure 14). Eighteen of 26 CEPH families were obtained as three-generation pedigrees and the others were two generation only, including a total of 208 offspring in this study. A total of 648 haplotypes have been examined in this study.

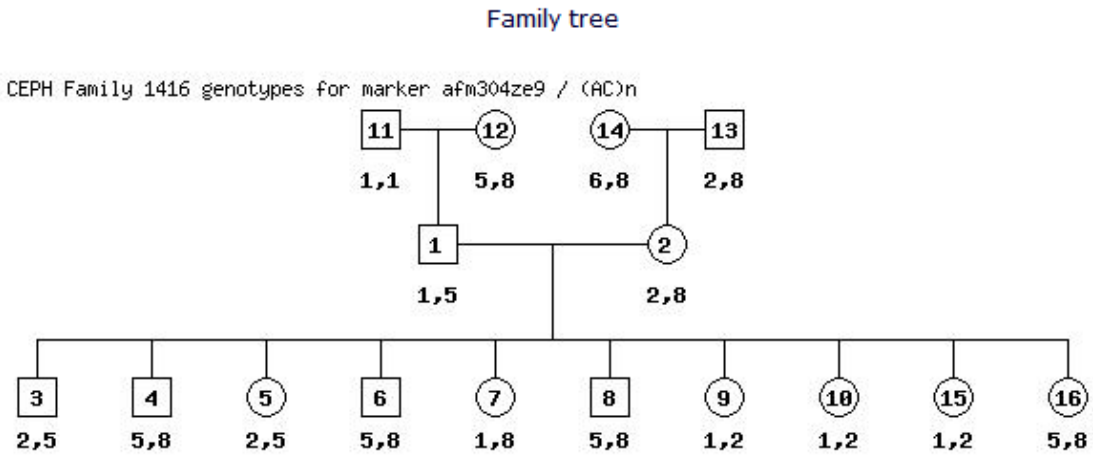


Figure 14: Example of segregation from a single-copy marker in one of the CEPH families. This figure shows individual genotypes for family 1416 on marker afm304ze9 / (AC)n. The grandparental origin of each allele in each child is unambiguous in this example

To distinguish the haplotypes corresponding to the beta defensin copy number for each member in the CEPH families, alleles produced from EPEV1 and EPEV3 in the GeneScan electropherogram were collected and arranged in each CEPH family. Analysis of alleles at the EPEV1 and EPEV3 microsatellites in this region for most of CEPH families (Figures 15, 16 and 17) showed one to three alleles per haplotype. As examples, in CEPH family 1416 (figure 15), they showed two and three alleles per haplotype, in CEPH family 1332, they only showed one and two alleles per haplotype (figure 16) and in CEPH family 1333, they showed two alleles per haplotype (figure 17). However, alleles at both EPEV1 or EPEV3 vary in copy numbers and are different between CEPH families; for example, the father in family 1416 has five copies of beta defensin from the HSPD5.8 PRT assay, one copy for each allele 137, 139 and 143 and two copies of alleles 141 from EPEV3, which indicates five copies, and one copy for each allele 169, 183 and 189 and two copies of alleles 181 are produced in EPEV1, which also indicates five copies (see Figure 15).

Meanwhile, the father in family 1332 has three copies of beta defensin from HSPD5.8 PRT assay, two copies of allele 139 and one copy of allele 141 from EPEV3, which indicates three copies, and one copy each of alleles 169, 181 and 185 from EPEV1, which also indicates three copies (see Figure 16). Furthermore the father in family 1333 who has four copies of beta defensin from HSPD5.8 PRT assay, one copy for alleles 137, 139, 141 and 143 from EPEV3 which indicates four copies, and one copy for

each allele 169 and 183 and two copies of alleles 187 from EPEV1, which also indicates four copies (see Figure 17). Therefore, the total numbers of beta defensin from both microsatellites are completely consistent with the HSPD5.8 PRT assay in most of the CEPH families used in this study and it also can begin to reveal the segregation of different copy number alleles from parents to the children as shown in figure 15, 16 and 17.

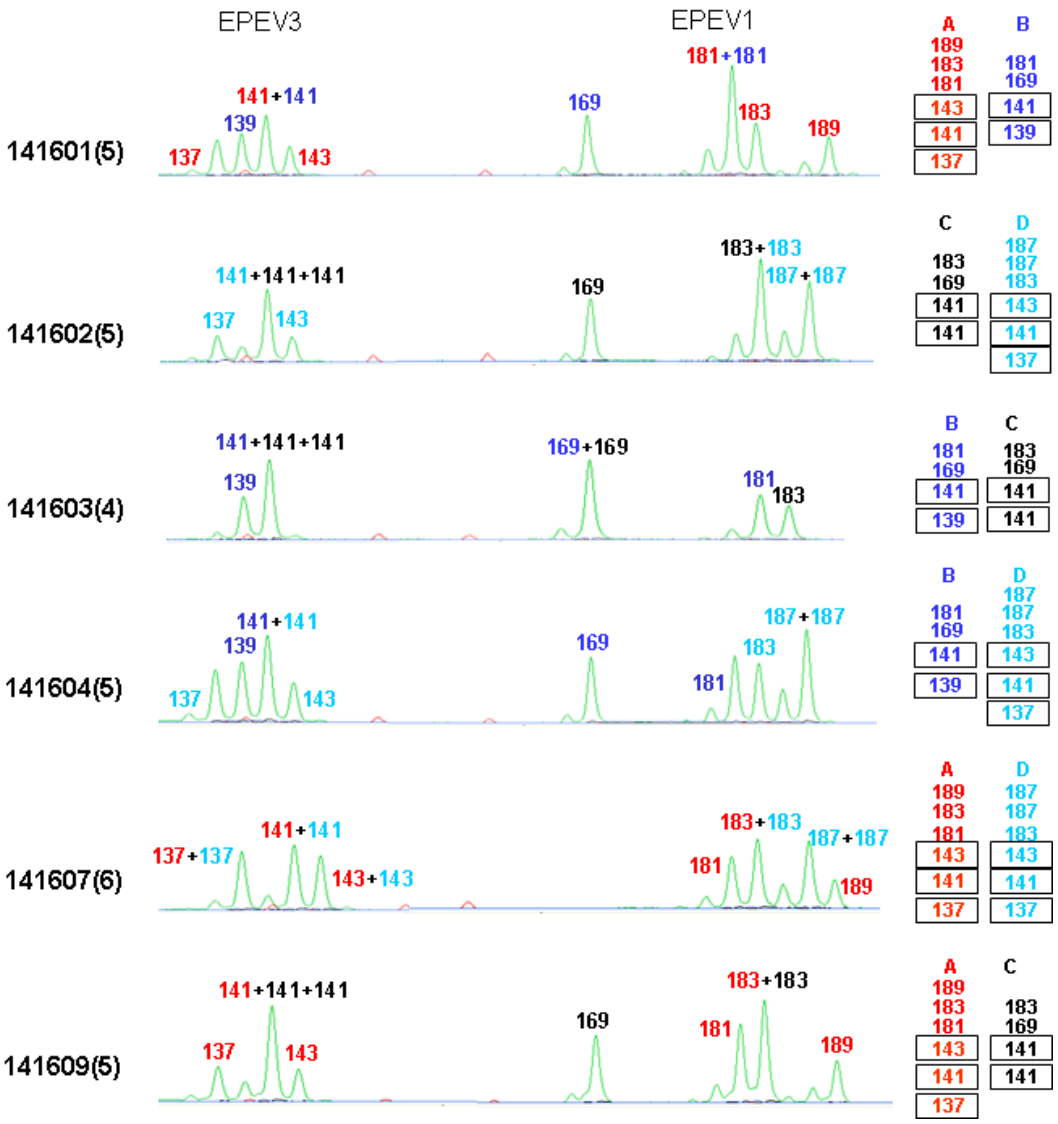


Figure 15: An example of the alleles produced by EPEV1 and EPEV3 in the GeneScan electropherogram for CEPH family 1416 and the four possibilities of haplotype combinations in the children. Beta defensin copy number as measured by the PRT assay is shown in the last column after the individual number in the family, for example 141601(5). The first and second individual numbers are corresponding to the father and mother in this family and followed by children. Haplotype analyses for the whole family are represented in figure 21.

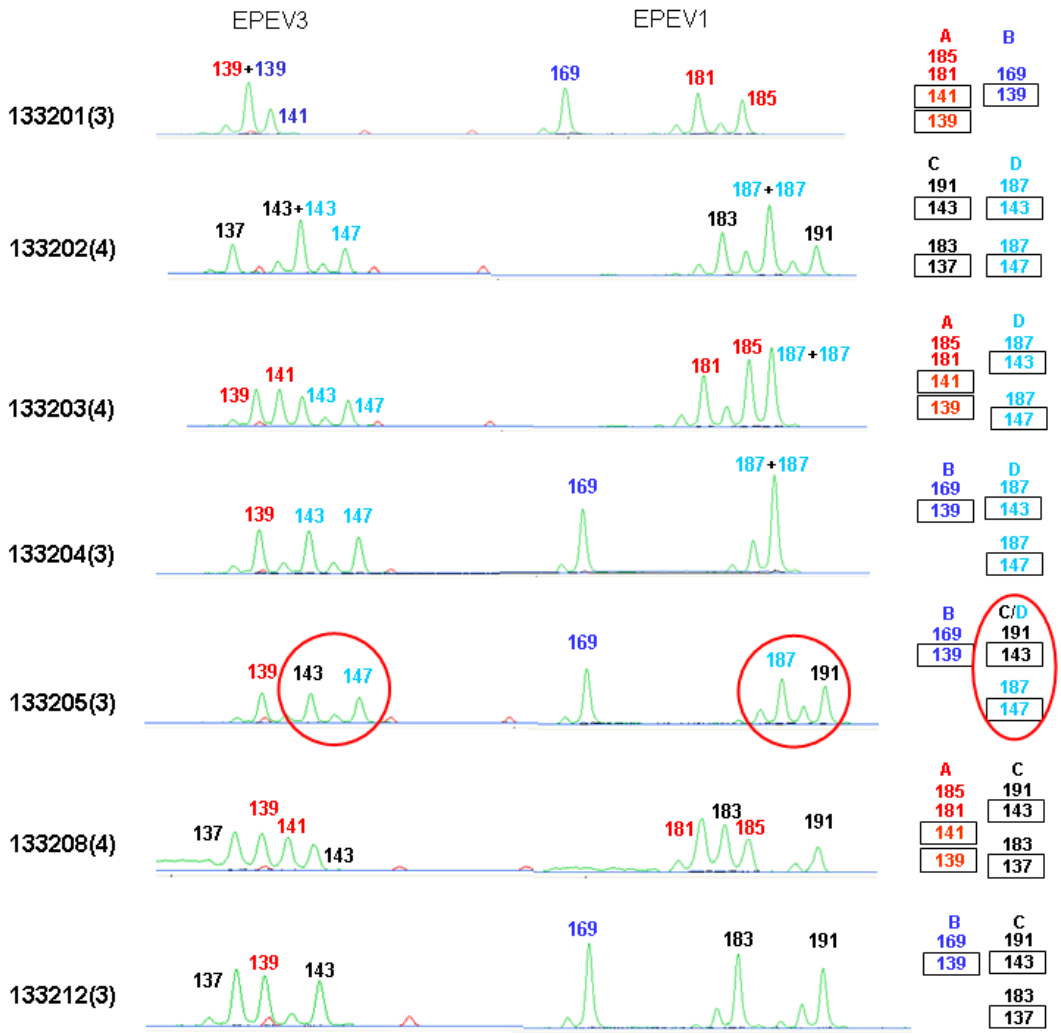


Figure 16: An example of the alleles produced by EPEV1 and EPEV3 in the GeneScan electropherogram for CEPH family 1332 with 5 combinations in the children, showing a recombinant event in child 133205. Although no obvious change in copy number has occurred for child no. 133205, the haplotype has generated diversity by a recombination event. Without a crossing over, this child should match either individual no. 133204 or 133206, and get **C** or **D** from the mother. Haplotype analyses for the whole family are represented in figure 22.

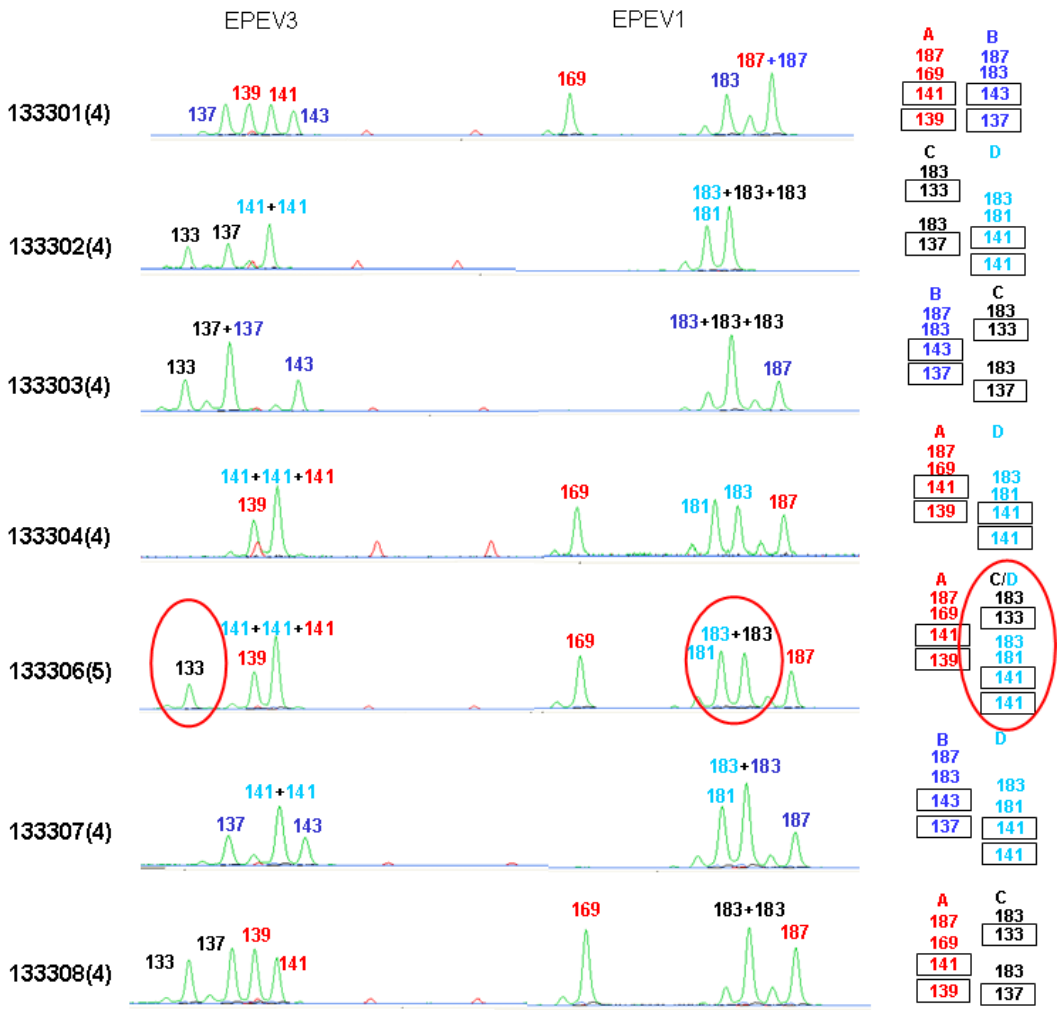


Figure 17: An example of the alleles produced by EPEV1 and EPEV3 in the GeneScan electropherogram for CEPH family 1333 with five combinations in the children, showing a recombinant event in child 133306 which showed five copies instead of four. This event has changed the beta defensin copy number in this child. Once again, this was another example in which the haplotype has been altered by a recombination event. Without a crossing over, this child should match individual no. 133304 or 133308, and get **C** or **D** from the mother. Haplotype analyses for the whole family are represented in figure 23.

3.2.2.2: Indel measurement assays

Short insertion-deletion polymorphisms (indels) have been recognised as an abundant source of other genetic markers that are widely spread across the genome, though not as common as single nucleotide polymorphism. They may be as small as 1 bp or they may involve one or many exons or even the entire genes. Unlike microsatellite analysis, indel polymorphisms can be genotyped with a simple typing procedure and may be easily analyzed, as length differences of PCR products do not encounter the stutter problem as explained in section 2.4.1 and 3.2.2.1. Thus by taking advantage of a 2 bp/5 bp indel at chr8:7,363,859-7,364,008 (about 10 kb downstream of *DEFB107*) in the beta defensin gene cluster, beta defensin copy number variation was verified and segregation of those copy number haplotypes was followed in 26 CEPH pedigrees.

As with microsatellite analysis, to distinguish the haplotypes corresponding to the beta defensin copy number for each member in the CEPH families, alleles produced from the rs5889219 indel in the GeneScan electropherogram were collected and arranged in each CEPH family. There were three distinct alleles. Although the rs5889219 indel is not fully informative (for example in figure 18, the 4 copy sample 141603 has variants in the ratio 1:1, compatible with any multiple of 2), this method provided valuable additional information by distinguishing repeat identities in the segregation analysis. Analysis of variants at the indel in this region for CEPH families (figures 18, 19 and 20) showed one to three alleles per

haplotype, which agreed with allele segregation from microsatellite analyses (Figures 15 - 17). The total numbers of alleles produced from the rs5889219 indel per haplotype in each CEPH family in this study were concordant with microsatellite alleles and also with the copy numbers measured from PRT assays.

In family 1416, as shown in figure 18, father has five copies number of beta defensin from HSPD5.8 PRT assay and displays one copy for allele 119, two copies for each allele 123 and 125, consistent with five copies. Both haplotypes were deduced according to the allele segregation in the children. Therefore, haplotype A contains one copy each of alleles 119,123 and 125, and haplotype B has one copy each of alleles 123 and 125. Even though the mother in this family also has five copies of beta defensin, the three alleles appear in different dosage. Haplotype C has one copy of each allele 123 and 125, and haplotype D has two copies of allele 119 and one copy of allele 125. All the children in this family have inherited non-recombinant haplotypes from their parents.

Meanwhile, in family 1332 (figure 19), both parents only have two alleles but in different dosage. Father has three copies of beta defensin from HSPD5.8 PRT assay, and showed one copy each of alleles 123 and 125 for haplotype A and one copy of allele 125 for haplotype B. Mother, who has four copies of beta defensin from HSPD5.8 PRT assay, showed one copy each of alleles 119 and 125 for haplotype C and two copies of allele

125 for haplotype D. However, child 133205 has received a recombinant haplotype from mother. Even though this recombination event cannot be clarified from these data alone due to the uninformative alleles, the segregation of allele still can be inferred in this family and it agrees with allele segregation in microsatellite analyses.

Additionally, by taking an advantage of 7 bp/8 bp polymorphism of 'A' repeat found near to the 2 bp/5 bp indel location, an extended rs5889219 indel analysis as described in section 2.5 was carried out on some of CEPH families who needed clarification for recombinant events. As an example, in family 1333, father who has four copies of beta defensin from the HSPD5.8 PRT assay, showed one copy each of alleles 169 and 171 for haplotype A and one copy each of alleles 167 and 172 for haplotype B. Meanwhile, mother who has same copy number of beta defensin as father showed two copies of allele 167 for haplotype C and one copy each of alleles 167 and 169 for haplotype D. Child no. 133306 in this family has inherited a recombinant haplotype C/D from mother and this event has changed the beta defensin copy number in this child (Figure 20). Once again, this was another piece of evidence that the haplotype has generated diversity by a recombination event (see also Figure 17).

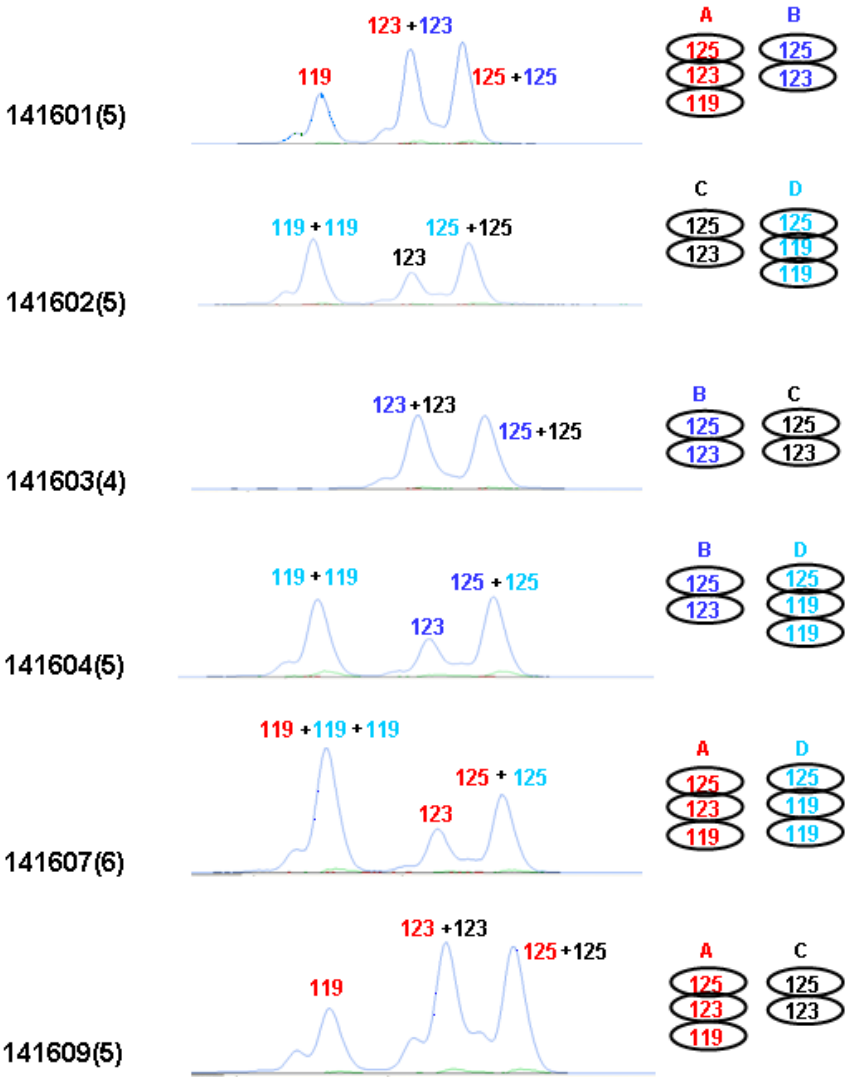


Figure 18: An example of the alleles produced by indel analysis in the GeneScan electropherogram for CEPH family 1416 and the four possibilities of haplotype combinations in the children. Beta defensin copy numbers which are measured by the PRT assay are shown in brackets after the individual number in the family, for example 141601(5).

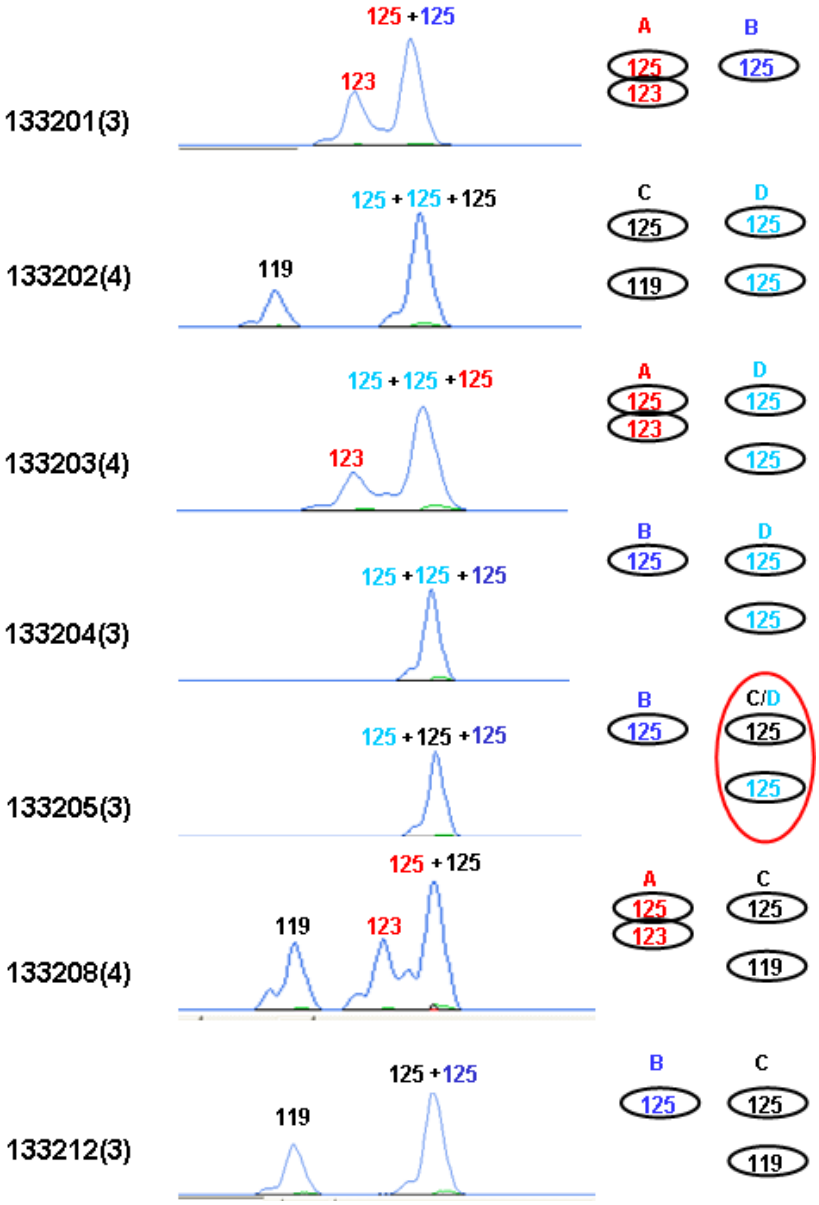


Figure 19: An example of the alleles produced by indel analysis in the GeneScan electropherogram for CEPH family 1332 with 5 combinations in the children, consistent with a recombinant event in child 133205 even though the alleles are uninformative, with the recombinant C/D haplotype indistinguishable in this assay from haplotype D. Compare microsatellite segregation in figure 13.

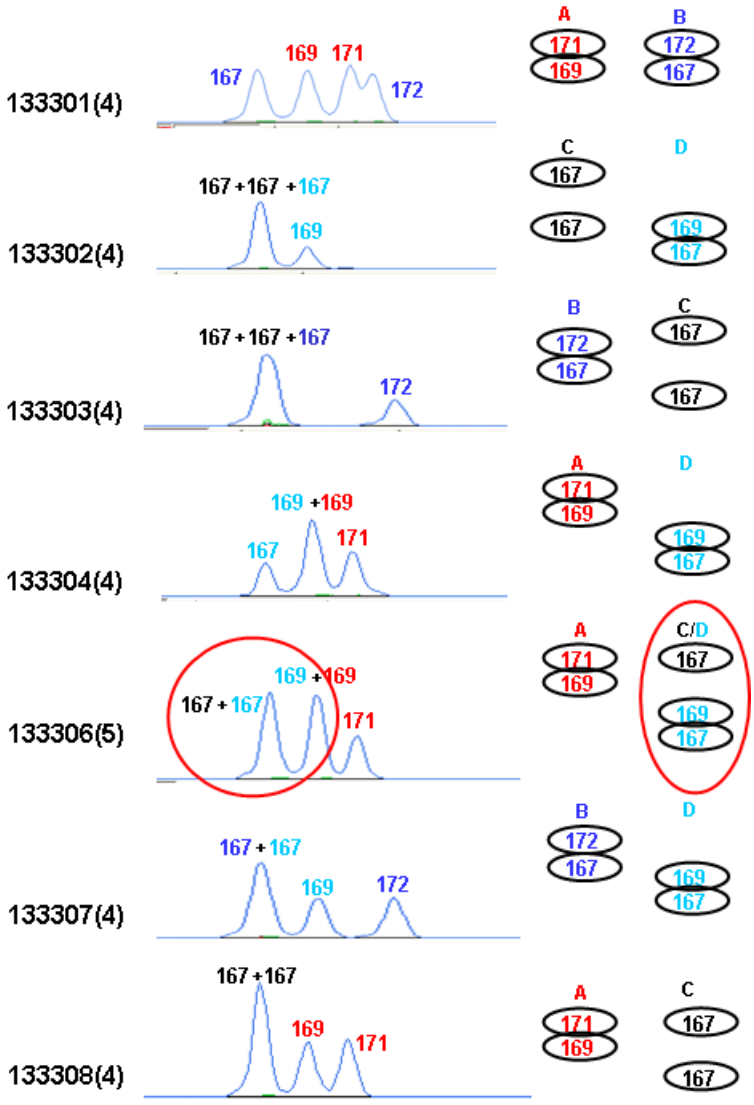


Figure 20: An example of the alleles produced by extended indel analysis in the GeneScan electropherogram for CEPH family 1333 with 5 combinations in the children, showing a recombinant event in child 133306. This event has changed the beta defensin copy number in this child. Once again, this was further evidence that the haplotype has generated diversity by a recombination event. Without a crossing over, this child should match individuals 133304 or 133308, and get either haplotype **C** or **D** from the mother.

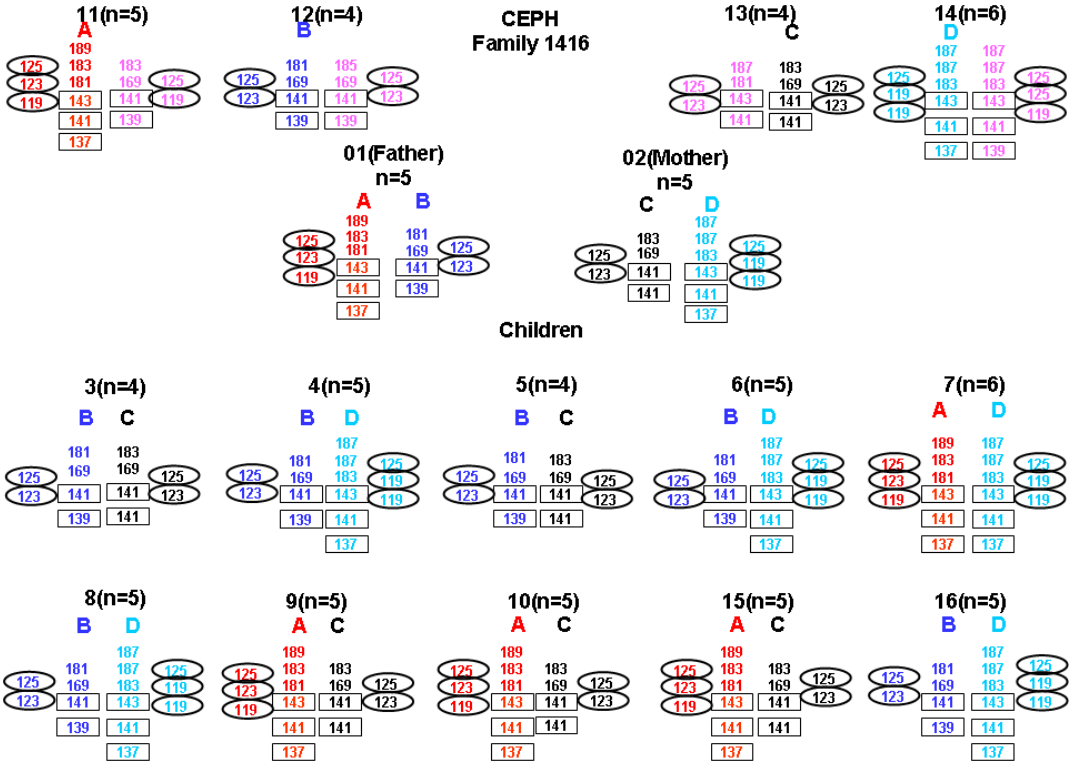


Figure 21: An overview of CEPH family 1416 over three generations. The individuals in this family are numbered from 1 to 14. The last four numbers (11, 12, 13 and 14) are grandparents. Beta defensin copy number measured by HSPD5.8PRT assay is shown in the bracket just after the individual number, for example (n=4). Both haplotypes are shown for each individual, either A, B, C or D. Each allele from microsatellite EPEV1 is shown as unboxed, alleles from microsatellite EPEV3 inside a box and allele from indel analysis inside an oval shape.

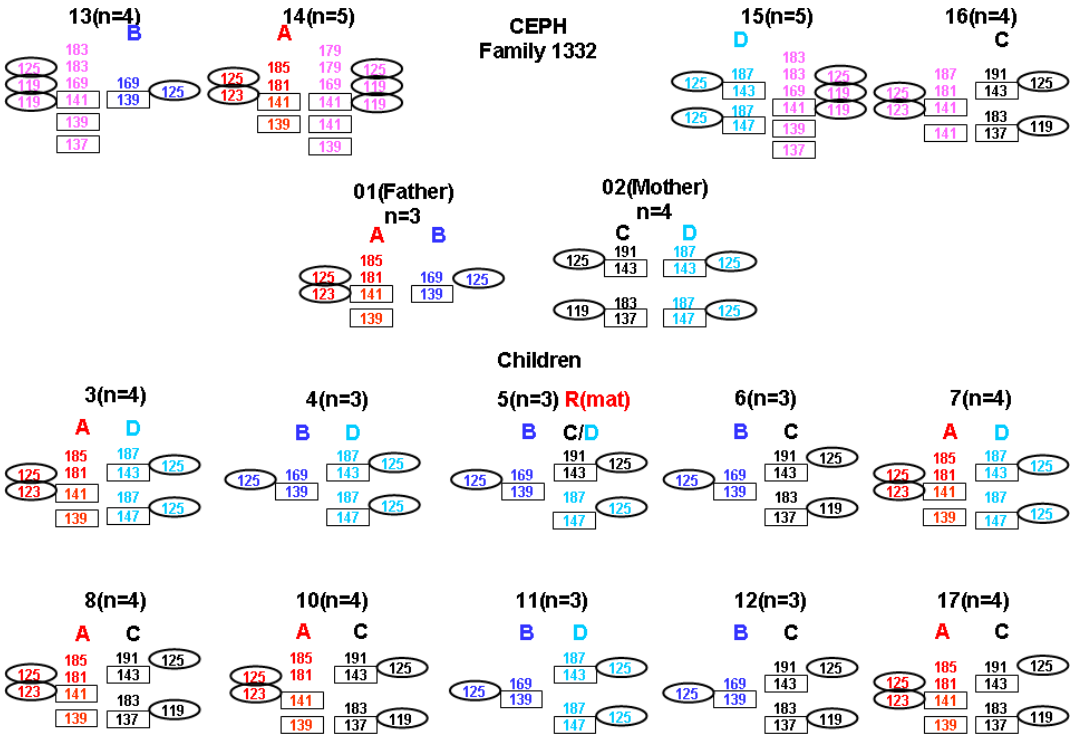


Figure 22: An overview of CEPH family 1332. The individuals in this family are numbered from 1 to 16. The last four numbers (13, 14, 15 and 16) are grandparents (see figure 21 for detailed key). Child 133205 has a recombinant (C/D) maternal haplotype.

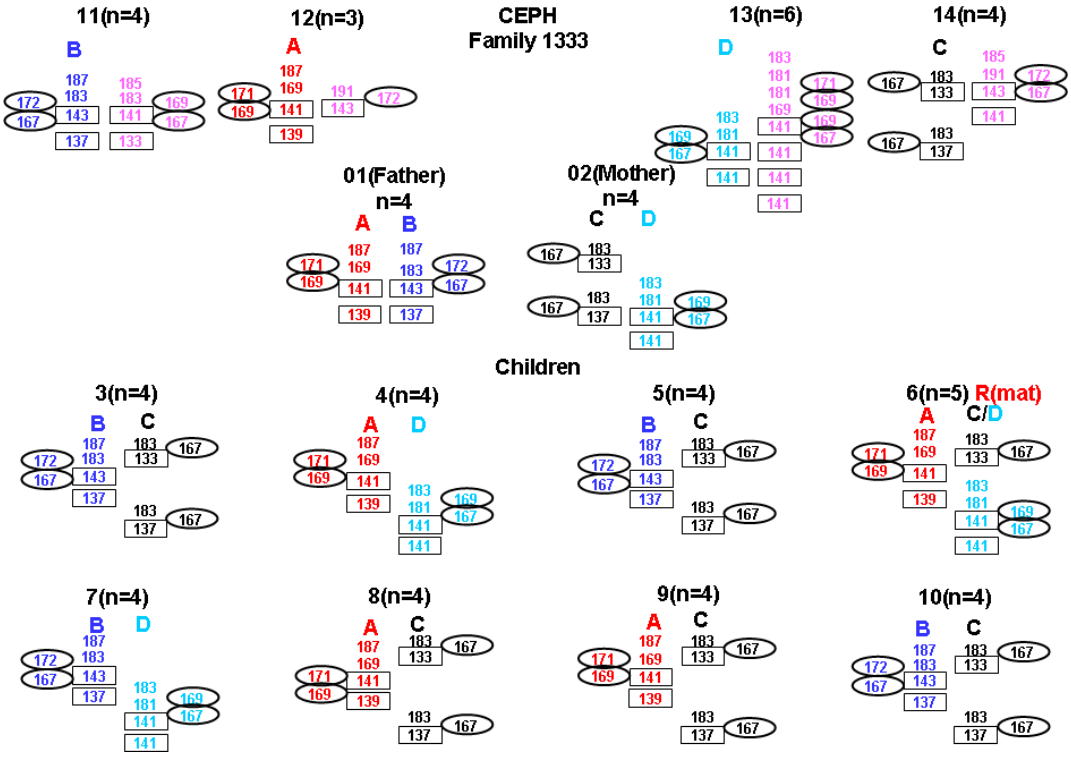


Figure 23: An overview of CEPH family 1333; for details of the symbols, refer to legends for figure 21 and 22. Child 133306 has a recombinant (C/D) maternal haplotype.

3.2.3: Restriction fragment length polymorphism (rs2203837)

The CEPH database genotypes for SNP rs2203837 in family 104 created an anomalous crossing over breakpoint. Therefore the genotypes for individuals 10407 and 10408 were directly verified using a PCR-RFLP assay and were subsequently corrected. After this correction, no anomalous segregation was seen.

3.2.4: Analysis of crossing over and copy number change

In this study, recombination has been observed between the beta defensin repeat units by following the transmission patterns of three microsatellite and one multiallelic length polymorphism (indel) markers in CEPH family pedigrees as shown in the previous section. Each offspring haplotype was derived from one meiotic product from each parent and thus the frequency of crossing over within a particular chromosome interval can be quantified by examining many offspring. This approach has led to the detection of at least 24 out of 416 offspring with inherited recombinant haplotypes on which parental beta defensin repeats have been rearranged. Most of the offspring haplotypes have not changed the copy number, but in three cases a new copy number was found in the recombinant chromosome; for example, child 133306 in figure 17 and 20 receives a recombinant maternal chromosome carrying three repeat units, whereas children inheriting non-recombinant haplotypes in this family show that the mother has two two-copy haplotypes. Data in CEPH family 1341 also shows that children have received different (reciprocal) recombinant maternal haplotypes without changing the haplotype copy number (not shown). The breakpoint of crossing over between the beta defensin repeats has been examined in all CEPH families whose child/children have inherited recombinant haplotypes. From that analysis, all the breakpoints must happen between two beta-defensin repeat which are far apart from each other; the recombinations cannot all simply be occurring within a coherent block of beta defensin tandem repeats, as shown in figure 24. For

example, the breakpoint of the maternal crossover in 133306 is located about 3 Mb proximal to the March 2006 Genome Assembly position of beta defensin repeats. Linkage mapping was introduced in this study to facilitate the location of recombination breakpoints by testing for genetic linkage within the interval for already known markers on chromosome 8p23.1. During chromosome segregation, alleles on the same chromosome can be separated and go to different daughter cells. If the alleles are far apart on the chromosome, there is a greater probability that a crossing over will occur between them.

However, for some recombinant cases, precise mapping of the beta defensin repeats could not be successfully determined because insufficient information was obtained from segregation analysis in the CEPH pedigrees. Therefore, those samples are labelled as unidirectional recombinants, based on the information from linkage analysis (Figure 24). In some transmissions, where parental repeats share many variants of the multiallelic markers typed and were not sufficiently informative to specify crossovers breakpoints. Due to this difficulty, the determined frequencies of these crossover processes are likely to be underestimates.

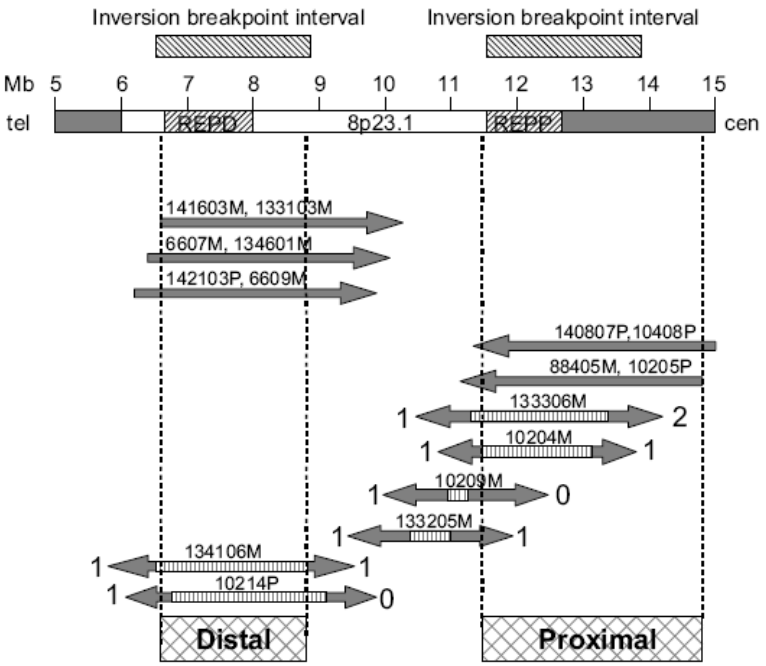


Figure 24: Genetic mapping of beta defensin repeats relative to crossover breakpoints in CEPH pedigrees, shown against genome assembly coordinates. The approximate locations of REPP and REPD are shown, as are the intervals containing the inversion polymorphism endpoints (Giglio *et al.*, 2001; Antonacci *et al.*, 2009). The crossovers indicate, at the top, selected unidirectional recombinants mapping the location of all repeat units on a haplotype, identified by family and individual name and parent of origin (e.g.,"6607M" is the maternal haplotype of child 07 in family 66). Below are shown selected crossovers which map repeats to the proximal and/or distal sites; the numbers of repeats mapping in each direction are indicated, and the hatched segments indicate uncertainty in the placement of breakpoints because of the lack of informative markers. Based on these data, beta defensin repeats can be mapped to the distal and proximal

intervals shown at the bottom (Figure adapted from (Abu Bakar *et al.*, 2009)).

Although the segregation of some repeat units was consistent with the accepted location for the beta defensin genes, some of the beta defensin repeats segregating in these CEPH pedigrees could not be located at the established interval in the genome assembly within REPD at approximately chr8:6,900,000-8,150,000, but instead mapped to another site more proximally located on chromosome 8p (Abu Bakar *et al.*, 2009). Analysis of crossover breakpoints in CEPH segregation data (table 7) showed that this second site must be bounded by AFMb294yg5 (chr8:11,515,050) distally and AFM205tb10 (chr8:14,724,000) proximally. There is no example of segregation inconsistent with location of beta defensin repeat units either at one or both of these sites on chromosome 8p. Abu Bakar and co workers (2009) also found that one crossover in child 135009 involved a breakpoint between two beta defensin repeats already mapped to the proximal site by crossovers in other children (135005, 135006 and 135008); the placement of the breakpoint in 135009 is consistent with the results of other mapping data and allows further refinement of the proximal site to a smaller interval by placing a proximal bound at chr8:12,880,230.

Table 7: Properties of 24 crossovers in CEPH pedigrees leading to reassortment of beta defensin repeat units. The numbers of copies of beta-defensin at the proximal and distal sites are shown for the recombinant haplotype. The final column indicates events leading to a recombinant haplotype with a copy number not found on either parental haplotype. Maternal haplotypes in family 1362 are in the inverted orientation, and distal and proximal markers are shown in the corresponding order.

Crossover child	Parental origin	Beta-defensin copy number		Distal flanking genetic marker	Assembly coordinate	Proximal flanking genetic marker	Assembly coordinate	Copy number Change?
		distal	proximal					
10204	M	1	1	AFMb294yg5/(AC)n	11515050	AFM320yf1/(AC)n	12786453	no
10206	M	1	0	afm304ze9/(AC)n	10918926	rs2898254	10929940	no
10207	P	2	1	rs751009	6741958	rs895252	8997910	yes
10209	M	1	0	afm304ze9/(AC)n	10918926	rs2898254	10929940	no
10214	P	1	2	rs751009	6741958	rs895252	8997910	yes
10215	M	1	1	rs2898254	10929940	AFM144zb2/(AC)n	11317121	no
10216	M	1	1	AFMa216zb1/(AC)n	10177958	afm304ze9/(AC)n	10918926	no
10406	M	1	1	cos140B11/pcr	9781807	rs2898254	10929940	no
10407	M	1	0	rs2001329	11024269	rs2203837	12864210	no
133205	M	1	1	AFMa216zb1/(AC)n	10177958	afm304ze9/(AC)n	10918926	no
133306	M	1	2	rs2001329	11024269	rs2203837	12864210	yes
133404	P	1	1	cos140B11/pcr	9781807	afm304ze9/(AC)n	10918926	no
134106	M	1	1	afm198wd2/(AC)n	6504084	rs433960	8805729	no
134109	M	1	1	afm198wd2/(AC)n	6504084	rs433960	8805729	no
135005	M	1	2	cos140B11/pcr	9781807	afm304ze9/(AC)n	10918926	no
135006	M	1	1	rs895252	8997910	cos140B11/pcr	9781807	no
135008	M	1	1	rs895252	8997910	cos140B11/pcr	9781807	no
135009	M	1	2	rs2001329	11024269	GATA23D06/pcr	12880230	no
136210	M	1	0	afm304ze9/(AC)n	10918926	AFMa216zb1/(AC)n	10177958	no
136211	M	1	0	afm304ze9/(AC)n	10918926	AFMa216zb1/(AC)n	10177958	no
136217	M	1	0	AFMa216zb1/(AC)n	10177958	rs433960	8805729	no
1329206	M	(ambiguous)		rs2898254	10929940	rs939557	12814009	no
1329404	M	1	1	rs751009	6741958	rs2898254	10929940	no
1329407	P	1	1	rs881362	6686304	rs433960	8805729	no

3.2.5: Targeted sequencing of variant repeats

To further clarify the evidence for a dual location of the beta defensin genes from analysis of crossover breakpoints in CEPH families, sequence analysis was undertaken in this study to compare sequences of beta defensin at different locations. In this work, DNA sequencing analysis has been used to determine all the sequence polymorphisms in the targeted sequencing blocks (SBs) within CEPH pedigrees, whether common or rare in the population. By following the segregation of all the sequence variants found in the SBs in these families, we hoped to distinguish the sequences of beta defensin at two different locations on chromosome 8p23.1, and at the same time to verify the finding from the previous work on using segregation of other markers (refer section 3.2.2). Haplotypes for both locations of the beta-defensins have been constructed from sequence variants found in order to identify the similarity of nucleotide sequence. Therefore, two genes, which are DEFB103 and DEFB4 in the beta defensin cluster, were chosen as target loci for sequence variant analysis. There were 77 CEPH individuals, selected from the ten (out of twenty-six) CEPH families who showed recombinant events, as the subjects for this analysis. We chose the parents, all the recombinant children and some non-recombinant children that inherited different haplotypes for each of the ten CEPH families (table 8). As reference and sequence quality control, we included our four standard reference samples, C11, C18, C62 and C66, in this analysis.

Table 8: List of ten CEPH families and numbers of members that were selected for sequencing analysis. Selected individuals in the second column are according to the individual number in each family. Numbers coloured in blue indicate the parents, numbers in black the non-recombinant children and numbers in red the recombinant children.

CEPH families (no.of members undergo sequencing analysis)	Individuals number
102(12)	01,02,03,04,05,06,07,09,11,14,15,16
104(8)	01,02,03,05,06,07,11,12
1332(7)	01,02,04,05,06,07,08
1333(6)	01,02,03,04,06,08
1334(5)	01,02,03,04,06
1341(8)	01,02,03,05,06,08,09,10
1350(9)	01,02,03,04,05,06,07,08,09
1362(9)	01,02,03,04,05,06,10,11,17
13292(7)	01,02,03,05,06,07,08
13294(6)	01,02,03,04,05,07

This work aimed to look at the sequence variants for both *DEFB103* and *DEFB4* for the above subjects (see table 8). Analysis of ratios of sequence chromatogram peaks might not accurately reflect the correct copy number, but with the data from the PRT assay for diploid copy number and haploid copy number obtained from microsatellites and indel (5DELI) analyses, there is sufficient information to predict the copy number from sequence variant ratios as shown in figure 25. For example, the genotype of 133201 (father) for the C/T variant has been deduced as two Cs and one T, while 133202 (mother) only has T, predicted to be four Ts according to the copy number measured by the PRT assay. Inferring the sequence variants on each of haplotypes A, B, C and D was carried out by looking at the

children's alleles. For example children 133207 and 08 have inherited both alleles (C and T) compared to the other children who only inherited the T allele, however, child 133207 inherited haplotype D which is different from child 08 (haplotype C). Since mother only has T alleles, and children 133204, 05 and 06 have only inherited T alleles from both parents, the two Cs have been traced to paternal haplotype A. Even though there is no T variant found in child 133205 who is known to inherit recombinant haplotype C/D from mother, it is still possible to map the variant into two locations on haplotypes according to the previous segregation and from other sequence variants in other locations as described later.

Figure 25: Patch of sequence traces collected from *DEFB103* shows segregation of one variant in CEPH family 1332 with five segregations in the children. Haplotype A in father has been deduced from the appearances of two Cs in children 133207 and 08. Segregation of this variant is consistent with a recombinant event in child 133205 even though the sequence alleles at this position are uninformative. Combination of more than one informative sequence variant in this analysis is able to distinguish the sequence for distal and proximal locations.

In order to illustrate clearly the use of sequencing analysis on the two loci within beta defensin repeats at distal and proximal sites, the identities of haplotypes in recombinant child 133205 have been drawn in figure 26. Three informative sequence variants have been deduced from each of the *DEFB103* and *DEFB4* loci for this family as shown in the tables (figure 26). Total number (unsegregated) of each variant in the second column corresponds to the beta-defensin copy number measured previously from other methods. As shown in figure 22, each copy of each sequence variant has been mapped to a haplotype, haplotype B or C/D, by consideration of the other children's haplotypes. Each of the alleles found in two microsatellite analyses and one multiallelic length polymorphism (indel) has been placed separately in a repeat unit according to their location in the genome assembly. For haplotype B, the exact location of the beta defensin is not known because there is no paternal recombinant; for simplicity, this copy is shown here at the established interval in the genome assembly (near to REPD). Moreover, allele 125 (in oval shape) from indel rs5889219 (5DELI), 139 (boxed) from EPEV3 and 169 from EPEV1 have been placed following the position in the genome browser. In haplotype C/D which has two copies, analysis of linkage showed that one copy is located at the distal and other copy at the proximal location. One copy at the distal location consists of allele 125 (oval) from5DELI, 143 (boxed) from EPEV3 and 191 from EPEV1, but allele 125 (oval) from 5DELI, 147 (boxed) from EPEV3 and 187 from EPEV1 at the proximal location. This was supplemented first with the sequence variants collected from

DEFB103 locus and then with the *DEFB4* locus for each repeat in each haplotype as shown in figure 26. This individual, for example, has illustrated different DNA sequence blocks between two sites which are TCTTGT at distal and TCACAC at proximal for haplotype C/D. Meanwhile haplotype B contained TCTTAC.

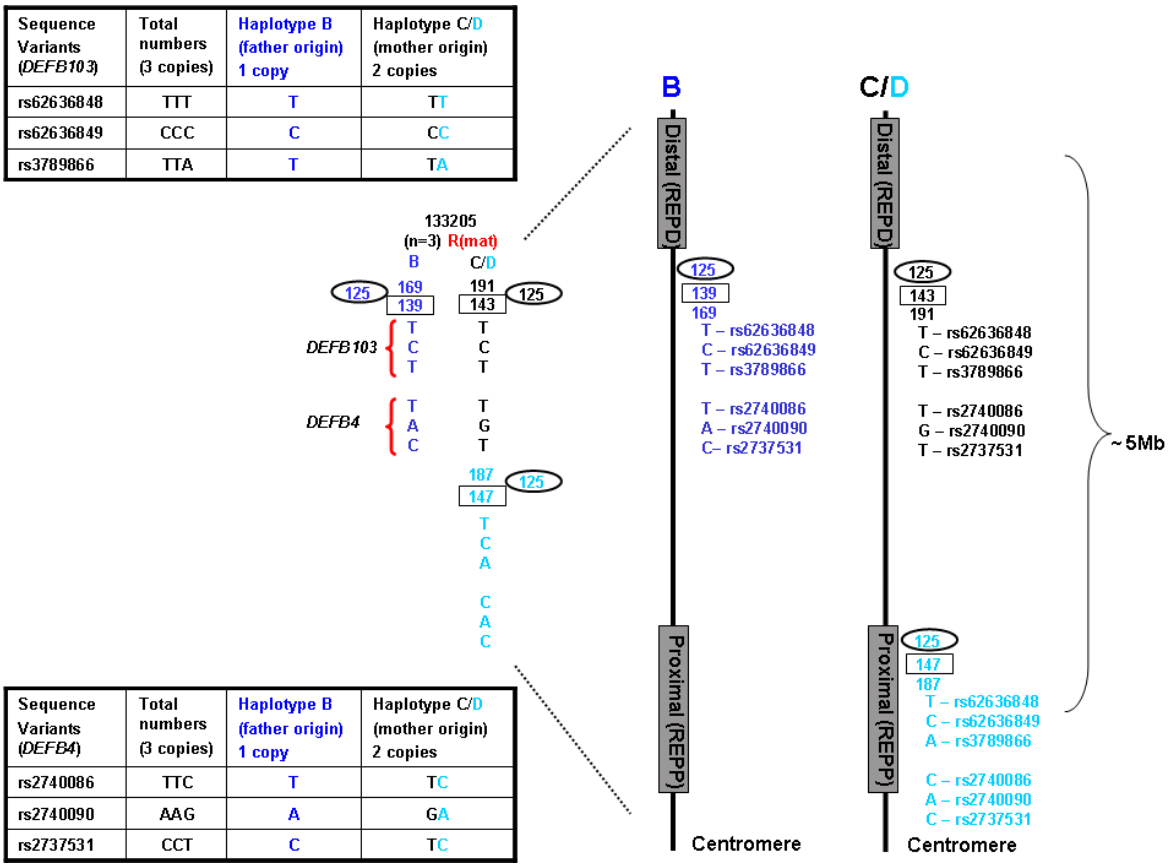


Figure 26: Figure of child 133205 (3 copies) with the two haplotypes, B and recombinant C/D together with the allele patterns collected from two microsatellite and one indel (5DELI). Each allele from microsatellite EPEV1 is shown as unboxed, alleles from microsatellite EPEV3 inside a box and allele from indel analysis inside an oval shape. Two tables on the top and

left corner describe the three informative sequence variants collected from *DEFB103* and *DEFB4* loci. Haplotypes B and C/D have been enlarged to show one copy of beta defensin at the distal location for haplotype B whereas for haplotype C/D, one copy at the distal and other copy at the proximal location. The location of each allele inside the repeat has been arranged according to the amplicon position in the genome assembly. The six letters drawn just after the third allele indicates the example of sequence variants found from sequencing analysis on *DEFB103* and *DEFB4* loci. All of the sequence variants have been mapped to the above locations by following the segregation in the children.

Table 9: List of sequence variants annotated from the March 2006 Genome Assembly for *DEFB103* and *DEFB4* loci (dbSNP release 130) used in sequencing analysis.

DEFB103	**chr8:7,273,588-7,275,448**	***DEFB4***	**chr8:7,789,266-7,791,843**
Sequence variants	**Position**	**Sequence variants**	**Position**
rs4143139	chr8:7273823	rs2740085	chr8:7789348
rs3988931	chr8:7273925	rs2698827	chr8:7789376
rs62636848	chr8:7274151	rs2740086	chr8:7789537
rs2740724	chr8:7274236	rs2740087	chr8:7789596
rs2740723	chr8:7274290	rs71509106	chr8:7789686
rs62636849	chr8:7274449	rs2740088	chr8:7789731
rs71513062	chr8:7274566-7274576	rs2740089	chr8:7789993
rs3789866	chr8:7274566	rs71511245	chr8:7790060
rs3789865	chr8:7274576	rs2698815	chr8:7790410
rs3789864	chr8:7274717	rs2740734	chr8:7790667
rs3789863	chr8:7274776	rs67055119	chr8:7790740
rs3789862	chr8:7274787	rs2737534	chr8:7790752
rs2740080	chr8:7274840	rs5008437	chr8:7790773
rs3866477	chr8:7275338	rs2737533	chr8:7790781
rs2740722	chr8:7275391	rs2698816	chr8:7790809
		rs71537820	chr8:7790957
		rs2737532	chr8:7790975
		rs71251804	chr8:7791113
		rs2740090	chr8:7791431
		rs72137464	chr8:7791511
		rs72437751	chr8:7791549
		rs72494252	chr8:7791586
		rs2740091	chr8:7791586
		rs2737912	chr8:7791646
		rs2737531	chr8:7791696

3.2.5.1: *DEFB103* locus

For *DEFB103,* the full sequence of a 1.8kb product (chr8:7273588-7275448, including the entire coding sequence) was successfully sequenced for 77 CEPH members as shown in table 8 by using two forward and three reverse primers (table 5). First, amplification from genomic DNA was carried out using DEFB103F and DEFB103R primers. Due to poly A and T repeats within this locus, three other primers were used as the internal sequence primers in order to obtain the full DNA sequence of *DEFB103* locus. The product covered about 15 reported sequence variants in the March 2006 Genome Assembly for this locus (dbSNP release 130) as shown in table 9. Only three of these sequence variants were actually observed and segregated in the CEPH families for this analysis. In addition, this work detected one new variant which is not found in dbSNP.

3.2.5.1.1: Segregation of the sequence variants found

To determine the sequence variants at the two locations of beta defensin, segregation of observed sequence variants in selected CEPH families was analyzed by comparison with information from segregation of alleles in microsatellite and multiallelic length polymorphisms (indel) (see section 3.2.2). The ratio of each sequence variant for each individual was estimated according to the copy numbers measured from PRT, microsatellite and indel assays. However, the non-linear signal at mixed positions in sequencing traces makes estimation of the ratio rather difficult.

As shown in figure 27 and 28, three and four sequence variants were successfully detected in families 1332 and 1333 respectively, and their haplotypes inferred in each individual. For family 1332, father (3 copies) showed two similar sets of sequence variants CCT for haplotype A and a set of sequence variants TCT for haplotype B. The first position of sequence variation, which is C (haplotype A) and T (haplotype B), has a ratio of 2:1. However, the second position (C) in this family appeared invariant, and differs from other CEPH families that show variants in this position. The third position T also did not show polymorphism in the father but was variable in the mother. Therefore, the recombinant pattern could be followed in this family because the mother has a polymorphism in the third position.

Moreover in family 1333 (figure 28), four variants including one which is not in dbSNP were collected to build the DNA sequences for each individual. Father 133301 has two two-copy haplotypes, but there is no recombinant event to determine the location of these two repeats even though two different sets of DNA sequences are shown in each haplotype. The first three positions are found to be polymorphic in father with the ratio of 3:1. The fourth position T did not show polymorphism in both paternal haplotypes. Meanwhile, mother 133302 carries polymorphism only at the first position, with a ratio of 2:2. Furthermore, two different DNA sequences for distal and proximal copies in haplotype C could be determined based on the recombinant pattern in child 133305, but haplotype D, appeared to

have two copies with the same alleles at this position. Furthermore, four forms of sequence haplotypes for three common variants have been observed in this work, TCA, TAT, TCT and CCT for *DEFB103* locus (Table 10). Three of them have been observed at both distal and proximal locations, namely TCA, TCT and CCT.

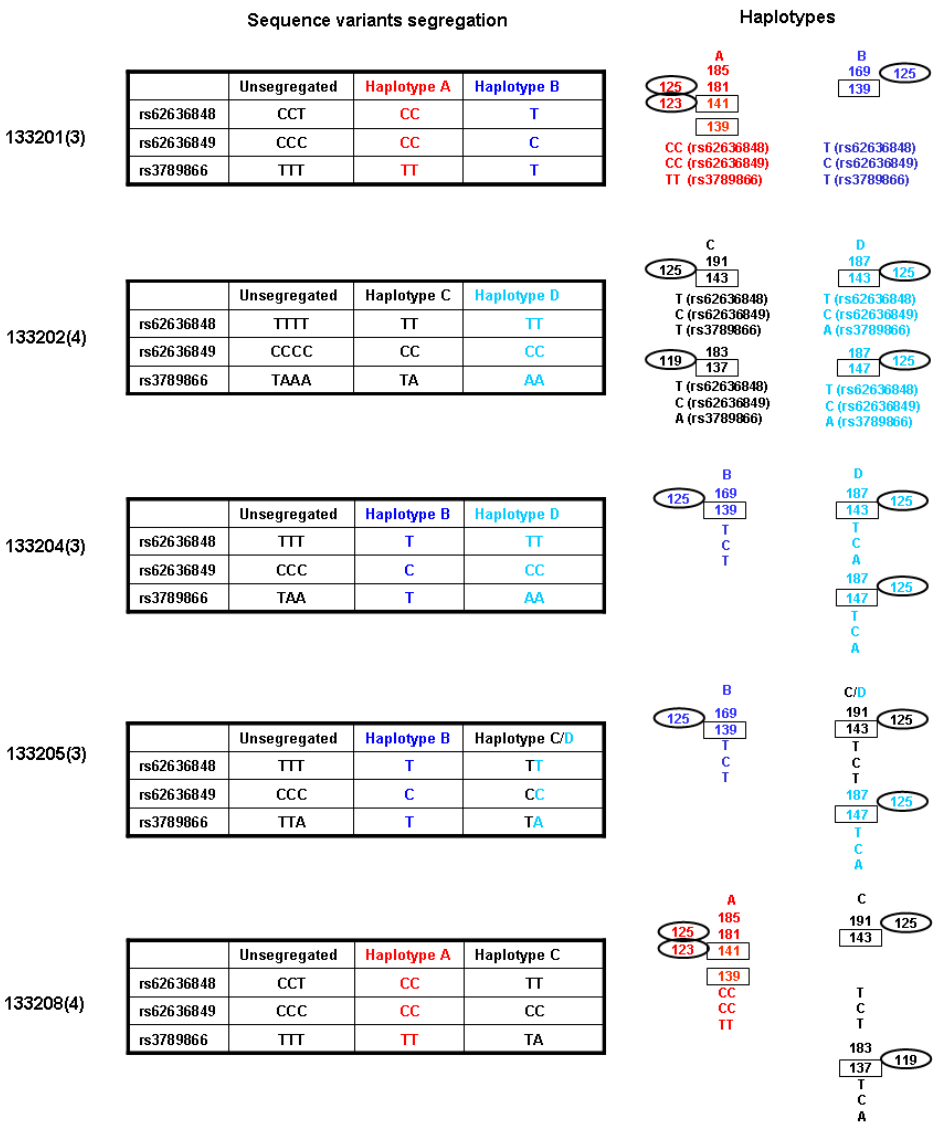


Figure 27: This is an overview of five members in CEPH family 1332 for *DEFB103* locus. For details of the symbols, refer to legends for figure 18 and 19. Each sequence variant can be assigned to haplotypes and locations using segregation. For example, father's T at rs62636848 is assigned to haplotype B because of its appearance in children 133204 and 05. Similarly, mother's T at rs3789866 is assigned to haplotype C from its appearance in 133208, and to a distal location from its appearance in 133205.

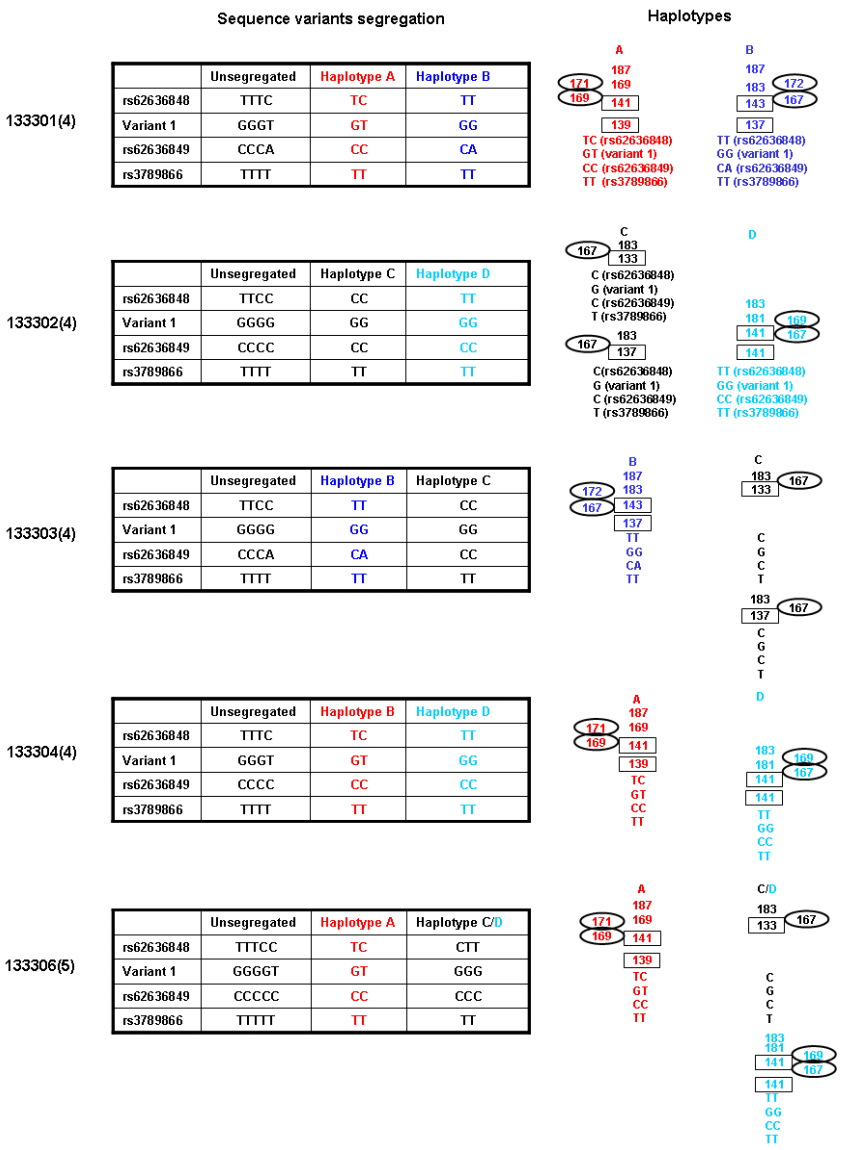


Figure 28: An overview of five members in CEPH family 1333 for *DEFB103* locus. For details of the symbols, refer to legends for figure 18 and 19. Each sequence variant can be assigned to haplotypes and locations using segregation. For example, father's C at rs62636848 is assigned to haplotype A because of its appearance in child 133304 and 06, and A at rs62636849 is assigned to haplotype B because of its appearance in children 133303. Meanwhile, mother's C at rs62636848 is

assigned to haplotype D from its appearance in 133304, and to a proximal location from its appearance in 133306.

Table 10: Sequence variants annotated from the March 2006 Genome Assembly for *DEFB103* (dbSNP release 130). Haplotype blocks constructed from 3 sequence variants (marked with red), are observed at distal and proximal sites; three similar sequence variants are found at both sites (TCA, TCT and CCT).

DEFB103	chr8:7,273,588-7,275,448	Observed							
Sequence variants	Position	Distal				Proximal			
rs62636848	chr8:7274151	T	T	T	C	T		T	C
rs62636849	chr8:7274449	C	A	C	C	C		C	C
rs3789866	chr8:7274566	A	T	T	T	A		T	T

3.2.5.2: *DEFB4* locus

Inspection of the sequence variants at more loci is needed to have strong evidence of variance in haplotype blocks at the two sites of beta-defensin. Consequently, *DEFB4* was chosen as another locus to be sequenced. In this study, a 2.5kb product of *DEFB4* (chr8:7789378-7791843) was sequenced in 77 CEPH members as shown in the table 8. As explained in section 2.8.2, amplification of the *DEFB4* locus is not as simple as the *DEFB103* locus, due to unequal distribution of GC-rich and AT-rich regions, and a block of AGGG repeats, which is known as microsatellite (EPEV2) in our experimental study. Therefore, two overlapping nested PCRs were performed to give primary amplicons (about 1.6 to 1.8 kb), and

several internal primers were designed as shown in table 6 in order to obtain the complete sequence at this locus.

Moreover, the product covered 25 reported single nucleotide polymorphisms, given in the March 2006 Genome Assembly for this locus (dbSNP build 130) as shown in table 9. A 9 base insertion/deletion (indel) was found from DEFB4Fb and DEFB4R2 sequencing primers (chr8:7789958-7789967) within this locus which could be used as the basis of allele-specific PCR to define association of variants within each repeat copy. In addition, ratio analysis on the 9 indel has been carried out as described in section 2.5.2 to identify the 9 indel genotypes for the CEPH members. Consequently, all of the collected single sequence variant and 9 indel ratio analysis data could be inferred together for each haplotype in the CEPH members as shown later. In the meantime, four primers, DEFB4InsF, DEFB4IndR, DEFB4InsR and DEFB4Fb were designed to specifically amplify only either insertion or deletion allele within the *DEFB4* locus. Furthermore, examination of all the sequence variants found from other primers (DEFB4Fb, DEFB4R2, DEFB4F2 and DEFB4R) has been performed and the DNA sequence for each haplotype has also been deduced in CEPH members.

3.2.5.2.1: Segregation of the sequence variants found

Deducing complete haplotypes from *DEFB4* sequence variants found in CEPH families was technically quite challenging, due to failed sequence of

some family members, and time constraint in order to complete the *DEFB4* analysis. Nevertheless, segregation of the 9 indel polymorphism based on ratio analysis has successfully been carried out in 10 CEPH families (data not shown). Both insertion and deletion forms have been observed at both sites (distally and proximally). Their transmission patterns also agree with the previous segregation from other markers. In order to determine the haplotype block at distal and proximal sites from this locus, the 9 indel segregation has been joined with the sequence variants. However, the haplotype blocks could only be deduced in CEPH family 1332 when taking into consideration the information sequence variants and the 9 indel from the *DEFB4* locus. The father (3 copies) in family 1332 (figure 29) showed three similar sets of sequence variants TTG; two for haplotype A and a set of TTG for haplotype B. However, 9 indel discriminated the two repeats for haplotype A. Meanwhile, one copy of the insertion has segregated together with TTG sequences in haplotype B, and is transferred to children 133204 and 05. The mother's haplotypes are far more useful as all of the three sequence variants are informative. Mother showed two different sets of sequence variants, CCA distally and TTG proximally for haplotype C (because of its appearance in child 133205). The indel discriminated the two similar sequence sets TTG for haplotype D because of the recombinant pattern in this family. There were three haplotypes when both 9 indel and other variants are considered in the analysis which are TTC with deletion, CCA with deletion and TTG with insertion (Table 11 or Figure 29). Even though three different haplotype blocks were observed as shown

in table 11, due to the relatively sparse analysis obtained for *DEFB4* locus, there are no examples of identical haplotypes at both sites (distal and proximal) as seen in *DEFB103* analysis; thus, combining haplotypes from both loci does not provide additional information.

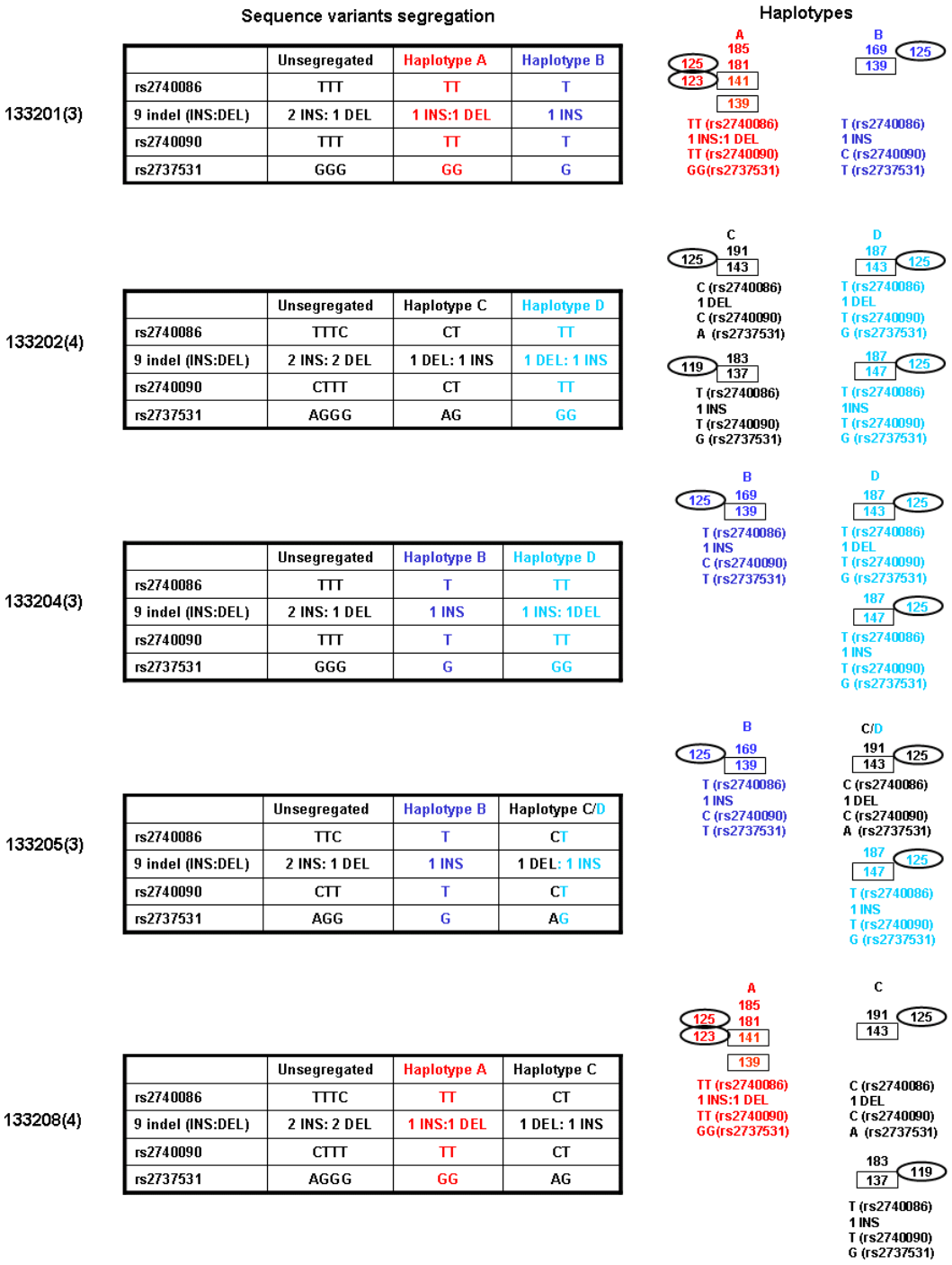


Figure 29: This is an overview of five members in CEPH family 1332 including results from sequencing analysis on *DEFB4* locus. For details of the symbols, refer to legends for figure 18 and 19. Each sequence variant

can be assigned to haplotypes and locations using segregation. For example, one copy of insertion from father's 9 indel is assigned to haplotype B because of its appearance in children 133204 and 05. Meanwhile, each of mother's haplotypes (C and D) have one copy of insertion and deletion, because of its appearance in children 133204, 05 and 08.

Table 11: Sequence variants annotated from the March 2006 Genome Assembly for *DEFB4* (dbSNP release 130). Haplotype blocks constructed from 3 sequence variants and 9 indel (marked with red), observed at distal and proximal sites. Two different haplotypes are observed for the two distal copies and one haplotype at the single example of a proximal copy.

DEFB103	chr8:7,273,588-7,275,448	Observed		
Sequence variants	Position	Distal		Proximal
rs2740086	chr8:7789537	T	C	T
9 indel	chr8:7789958-7789967	DEL	DEL	INS
rs2740090	chr8:7791431	T	C	T
rs2737531	chr8:7791696	G	A	G

3.3: Discussion

In this study, by utilising highly polymorphic markers (microsatellites and a multiallelic indel) CEPH pedigree analysis has provided valuable information on the segregation of the copy number variants. The various other CNV measurement methods available, with the exception of Southern blot hybridization of pulsed-field gels (Yang *et al.*, 2007) and of some fibre-FISH analyses (Perry *et al.*, 2007), determine diploid copy number for multiallelic loci but do not report the real co-dominant genotypes (i.e., in terms of component haplotypes) in an individual. There is still less information on the copy number location and the parental origin of copy number. Furthermore, for the human beta-defensin repeat, in which the variation involves a large physical distance, analysis of the copy number variants descending through a pedigree provides a means to

dissect the copy number variation into its constituent haplotypes. Surprisingly, this approach not only verified the diploid copy number of beta defensin measured by PRT, but also determined that diploid copy number was the sum of contributions from variants at each of two distinct locations. Even though some of the variants might be not informative to distinguish copy number identities, amplification from different locations of the variants in the region of interest have clarified the individual genotype for the copy number variants and also to helped to determine whether copy number inferred was the product of meiotic recombination. Transmission of alleles within the beta-defensin region in CEPH families demonstrated that recombinant events help in mapping the location of the alleles contributing to the copy number.

By using a genetic approach in the CEPH pedigrees and subsequent linkage analysis, this work has shown that homologous recombination is involved in beta-defensin genome exchange patterns in the individual meioses. From the recombinant chromosomes, the polymorphic beta-defensin region has been identified and mapped explicitly into two loci separated by several Mb of single-copy sequence to locations consistent with the REPP and REPD repeat regions. This study has shown that beta-defensin copy number is diversified by this recombination mechanism. Nevertheless, this 8p23.1 region that contains the beta-defensin cluster has been shown to be involved with another structural polymorphism, a large inversion (Giglio *et al.*, 2001; Antonacci *et al.*, 2009). Giglio *et al.*

(2001) demonstrated not only the involvement of OR gene clusters in mediating common chromosome rearrangements, but also some recurrent rearrangements, such as the recurrent inversion, indicating the possibility of developing a profile of the individual risk of having progeny with chromosome rearrangements by looking at the 8p23.1 region. In this work the inversion polymorphism could be a mechanism for 'transporting' the beta-defensin repeats between the distal location and the proximal location at REPP, which could arise by frequent 'flipping' of this sequence between different inversion states to sometimes include, and sometimes exclude, beta-defensins at each location. There were examples of zero, one or two copies of beta defensins at the proximal location, and zero, one and three at the distal location observed in this study (see Figure 19 and 20).

In addition, sequence analysis on *DEFB4* and *DEFB103* in these CEPH pedigrees, using segregation in recombinant offspring to map sequence variants to the distal or proximal sites, has clearly verified the evidence from microsatellite and indel alleles that sometimes the same alleles and sequence variants can be found at both proximal and distal sites. In figure 24, sequence polymorphism at 3 positions found in the mother for family 1332 has shown that T, C and A mapped at both proximal and distal sites from their appearance in child 133205. Insertion and deletion forms of 9 indel could also each be found at both sites from their appearance in children 133205 (Figure 29).

3.4: Conclusion

In conclusion, human beta defensin copy number in 26 reference CEPH families was measured accurately by using two PRT assays, HSPD5.8 and PRT107A, and confirmed by microsatellite and indel analyses. Measurement of variation in copy numbers of beta defensin clarified that there are commonly between two and eight copies per diploid genome, and one copy to five copies per haploid genome, with two and three copies being the most common haploid copy number. The additional methods used in this study, microsatellite analysis and multiallelic length polymorphism “indel”, assisted in looking at the transmission of alleles and segregation patterns for the whole CEPH families. Recombinant events between homologous chromosomes were discovered as a simple mechanism for exchanging the beta defensin material between homologous chromosomes while inferring a child’s haplotype and allowed the location of beta defensin repeats to be mapped relative to the position of crossovers inferred from the CEPH segregation database. Notably, some of those recombinant events have produced a new haplotype that results in changes in the copy number of beta defensin, at a frequency about 0.7% per gamete (3 out of 416), whereas most of the recombinants do not change the copy number. Finally, this study concludes that beta-defensin repeats can be in two positions in the human genome. One of these sites is consistent with the location in distal 8p23.1 indicated on the genome assembly, but there is also a separate location about 5Mb

proximally. The existence of two distinct loci for variation in beta-defensins is unusual, and may be the only such example in the genome, but it nevertheless illustrates that mechanisms of copy number variation may be more varied and complex than previously suspected.

Chapter 4: Development of a Multiplex PRT Measurement System for Beta-Defensin Copy Number Variants

4.1: Background information

Standard methods of molecular genetic analysis at individual genetic loci, including Southern blotting, cytogenetic methods, and PCR-based approaches are the best known methods for mapping structural variants at different resolution. Screening of single nucleotide polymorphisms in certain individual genomes has been extensively utilised in order to analyze the genetic determinants of complex traits and common disorders. This approach was mainly based on their abundance in the genome, and their use was boosted by the technological development of tools for high-throughput analysis of SNP variants. Until recently, the roles of SNPs in human disease have been focused exclusively in gene mapping studies. Subsequently, development of the latest generation of array comparative genomic hybridization, next generation sequencing and SNP genotyping arrays have accelerated the discovery of structural variations. Numerous studies have noted the greater extent of various types of structural genomic variation, ranging from 100 bp to 100 Mb, than previously expected (Iafrate *et al.*, 2004; Feuk *et al.*, 2006). Furthermore, recognition of the critical role of structural genetic variation in gene expression and

disease phenotype is growing rapidly, including the effect of copy number variants (CNVs) (McCarroll and Altshuler, 2007).

Early work on determination of beta-defensin copy number variants has begun by using a combination of different methods independently, Multiplex Amplifiable Probe Hybridization (MAPH), semiquantitative fluorescence *in situ* hybridization (SQ-FISH), segregation and microsatellite analysis. These were able to detect the distribution of beta-defensin copy number per diploid genome as mentioned before and deduced that individual chromosomes had between one and seven copies of this repeat unit (Hollox *et al.*, 2003). Later, in another work which was the first study on the association of beta-defensin and a clinical phenotype using (MAPH) and microsatellite analysis, Hollox *et al*. (2005) examined copy number in a cohort of patients with cystic fibrosis (CF), with no association found. Subsequently, Fellermann *et al.* (2006) analysed the human beta-defensin 2 (*DEFB4*) gene in healthy individuals and patients with Crohn's disease and ulcerative colitis using genome wide DNA copy number profiling by array-based comparative genomic hybridization and quantitative polymerase-chain-reaction approach. They proposed that a lower *DEFB4 g*ene copy number predisposes to colonic Crohn's disease. Even though the results fit with a model that presents susceptibility alleles as reducing the antimicrobial barriers in the gut, this work was only reported from a non-replicated study based on small sample numbers.

Hollox *et al.* (2008) used three combined methods to diplotype beta defensin copy number: paralogue ratio tests (PRT), restriction enzyme digest variant ratios (REVDR) and MAPH in investigation of the relationship between beta-defensin gene copy number and risk of psoriasis. The authors have similar PRT methods (Armour *et al.*, 2007) as in the previous study (refer sections 2.3.1 and 3.2.1), and compared these data to the typing from MAPH and ratios of multisite variants (REVDR) mapping around the *DEFB4* gene to determine copy number of beta defensin repeat per diploid genome. Copy number data from *DEFB4* are consistent with variation of integer copy numbers, so typing by more than one method or repeat typing can often improve results and produce clear clusters of samples falling into each integer copy number diplotype (Hollox *et al.*, 2008).

To fill the gap arising from the absence of an accurate typing method for measurement of beta-defensin copy number variants, a triplex system has been developed in this study. This system consist of two paralogue ratio tests (PRTs); PRT107A (section 2.3.2) and HSPD21 (section 2.3.3), and 5DELR4 which is an extended rs5889219 indel analysis as described in section 2.5. Similar to the other two PRT assays described previously, HSPD21 primers are designed precisely to amplify from copies of a diverged (low copy number) repetitive sequence at the copy-variable locus 3kb upstream of *DEFB4* on chromosome 8, and at only one other (reference) location (from chromosome 21) as shown in figure 30. As

described in chapter two, sections 2.3.1, 2.3.2 and 2.3.4, both products were detected without other detectable fragments under the conditions used.

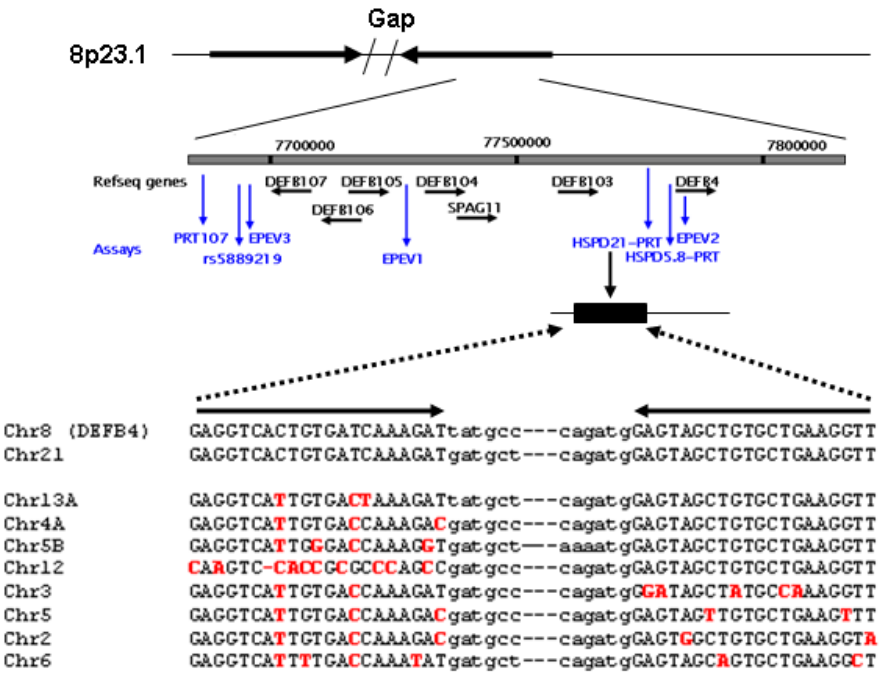


Figure 30: Principle of the HSPD21 assay at *DEFB4*. The top line shows the general structure of the repeat unit containing beta defensin genes (which has two inverted rather than tandem repeats in the March 2006 assembly). The middle panel shows the location of genes *SPAG11*, *DEFB4* and *DEFB103-07* (coloured in black), together with the locations of assays (coloured in blue) used in this study including HSPD21-PRT. In the detailed display at the bottom, the primers amplify products at 3kb upstream of *DEFB4* on chromosome 8, and from a reference copy on chromosome 21, but have multiple mismatches with other copies of the element. In this way a single primer pair can be used to amplify two very similar products, one from near *DEFB4*, the other from chromosome 21.

In each sample which underwent this triplex system, a collection of traces as shown in figure 31 was obtained that consisted of two PRT assays (PRT107A and HSPD21) and one indel assay (5DELR4). Due to the size differences, products from each assay could be analyzed in the same capillary. In the first row, this person has an allele ratio of 2:1 for 5 DELR4, and is measured as 3 copies from both PRTs. In the second row, this person has allele ratio 1:1 from 5DELR4, and 4 copies from both PRTs. In the last row, this person with allele ratio 2:2:1 from 5DELR4 has 5 copies from both PRTs. These traces showed consistency between three assays that could be used to determine the copy number by a single assay. Likelihood analysis as described in section 2.10 was carried out to allow evaluation of the joint probability of all the data (two PRTs and 5DELR4) observed for a given sample for a likely integer copy number together with evaluation of the statistical confidence. This PRT-based triplex assay has been practically applied on large case-control association studies of Crohn's disease samples in this study.

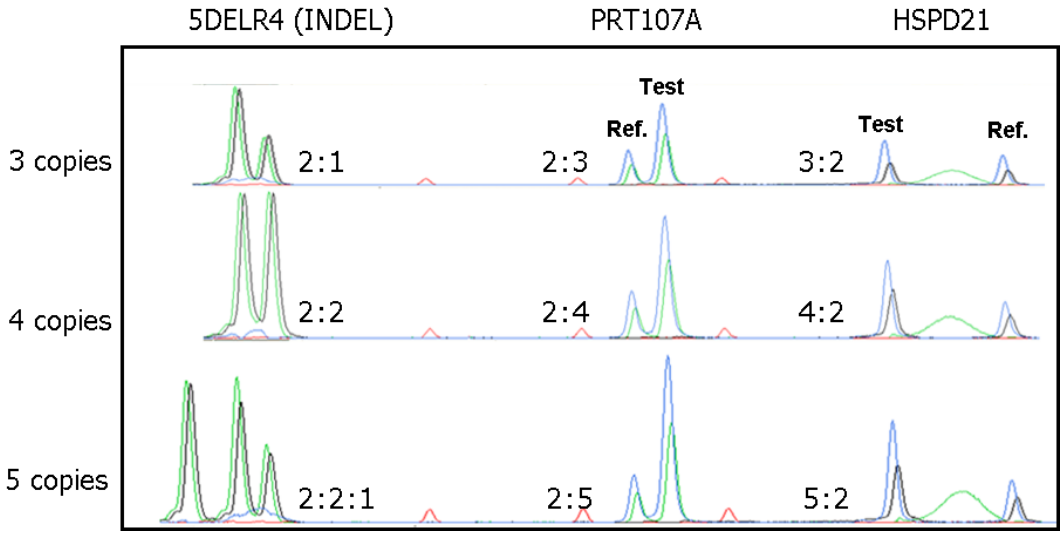


Figure 31: This figure shows the traces obtained from the triplex system, (two PRT assays (PRT107A and HSPD21) and one indel assay (5DELR4). Predicted copy number for each individual in these traces from this system is shown to the left of the box. Inside the box, there are two peaks visible for each PRT; one from the reference locus and one from the variable (test) locus. The genomic ratio of test:reference copies for each PRT is shown just before the peaks. For the indel, there are up to three distinct alleles (123, 125 and 127bp) detected in this experiment. The genomic ratio from each peak to the smallest peak is also shown here just after the peaks.

Crohn's disease is an inflammatory disease of the intestine that may affect any part of the gastrointestinal tract from mouth to anus, causing a wide range of symptoms. It usually causes abdominal pain, diarrhoea (which may be bloody but less common), vomiting, or weight loss that results from malabsorption but may also cause complications outside of the

gastrointestinal tract to create other chronic inflammatory features such as skin rashes, arthritis and inflammation of the eye (Baumgart and Sandborn, 2007). Crohn's disease is also recognized as an autoimmune disease, in which the body's immune system attacks the gastrointestinal tract, causing inflammation; it is classified as a type of inflammatory bowel disease. The IBDs are common causes of chronic gastrointestinal disease in developed countries, with reported prevalence rates of 1 in 250 in Northern European populations (Rubin *et al.*, 2000). They comprise two main forms; Crohn's disease (CD) which is often localized with a discontinuous distribution along the intestine, and ulcerative colitis (UC) which always involves the rectum and is continuous in distribution. In adults, CD is often limited to the ileum (the last portion of the small bowel) and/or colon whereas CD in children is typically more extensive. CD is classified into two groups of phenotype; colonic and ileal Crohn's disease.

The etiology of CD and UC is complex and remains elusive, but epidemiological data conclusively point to an impaired immune response against the bacterial flora in a genetically susceptible host (Loftus, 2004). Investigation of familial clustering of inflammatory bowel disease, ethnic variability and twin studies have provided strong evidence to support a role for inherited factors, which are found to be much stronger for CD than UC (Russell *et al.*, 2004; Van Limbergen *et al.*, 2009). Both techniques of candidate gene analysis and genome wide scanning (GWAS) have been used to identify genes that influence disease susceptibility and progression

in CD. Candidate gene studies have involved linkage analysis, case-control studies and family-based transmission disequilibrium testing (Russell *et al.*, 2004; Van Limbergen *et al.*, 2009).

Hugot *et al.* demonstrated the first IBD susceptibility locus, *IBD1* on chromosome band 16q12 (Hugot *et al.*, 1996). Closer examination of the *IBD1* region allowed identification of frameshift mutations in the *NOD2* (nucleotide-binding oligomerization domain 2) gene which was later renamed *CARD15* (caspase activation recruitment domain 15) as the first susceptibility gene for CD (Ogura *et al.*, 2001). This gene is a member of the innate immunity system which belongs to the family of pattern-recognition receptors (PRRs) and comprises a central nucleotide binding domain (NBD), N-terminal caspase recruitment domains, and a leucine-rich repeat (LRR) region that recognize bacterial peptidoglycans; they are among the first lines of defence against microbial flora in the gastrointestinal tract (Inohara *et al.*, 2002).

Since the discovery of NOD2 as a CD susceptibility gene, the role of the innate immune system in IBD has been studied extensively, and mutations in a number of genes of the innate immunity pathway have been found to be associated with CD susceptibility. In addition to *IBD1*, *IBD2* at chromosome band 12p13, *IBD3* (chromosome 6p21), *IBD4* at 14q11, *IBD5* (15q13), *IBD6* (19p13), *IBD7* (1p36), *IBD8* (16p) and *IBD9* (3p26) were identified in genome wide linkage screens involving 1200 family pedigrees

affected by IBDs (Steven and Yin Yao, 2004). A SNP in the *TNFSF15* gene at chromosome 9q32 has also been reported to be associated with CD (Yamazaki *et al.*, 2005). The *TNFSF15* gene encodes tumor necrosis factor (TNF) superfamily member 15. This gene is upregulated in mucosal cells lining the intestine of patients with CD.

In 2009, Van Limbergen and colleagues reviewed more than 30 distinct genomic loci which are involved in genetic susceptibility to Crohn's disease (CD) from epidemiological data, based on concordance information in family studies, via linkage analysis to genome wide association studies (Van Limbergen *et al.*, 2009). They have described the loci which have most influenced the understanding of CD pathophysiology and/or provide strong therapeutic potential. These loci including innate pattern recognition receptors (NOD2/CARD15, TLR4, CARD9), and genes involved in the differentiation of Th17-lymphocytes (IL-23R, JAK2, STAT3, CCR6, ICOSLG), autophagy (ATG16L1, IRGM, LRRK2), maintenance of epithelial barrier integrity (IBD5, DLG5, PTGER4, ITLN1, DMBT1, and XBP1) and the orchestration of the secondary immune response (HLA-region, TNFSF15/TL1A, IRF5, PTPN2, PTPN22, NKX2-3, IL-12B, IL-18RAP, MST1).

Nevertheless, Wehkamp and colleagues (2005) have proposed that the imbalance between the commensal flora and the host epithelium may be crucial to Crohn's disease pathogenesis. Crohn's mucosa may often be the

target of various infections in which there may be a primary defect in the peptide barrier of intestinal antibiotic defensins, which protect the normal mucosa extremely efficiently against adherent or invasive microbes as illustrated in figure 32 (Wehkamp *et al.*, 2005). Thus, a comprehensive understanding of these functionally relevant peptides is essential for the understanding of mucosal biology, and an impaired balance of these effector molecules might explain various aspects of pathogenesis in IBDs (Wehkamp *et al.*, 2005). Defensins are endogenous antibiotic peptides that produce a chemical barrier at the epithelial surface, and their respective insufficiency may lead to bacterial adherence to the mucosa, slow invasion, and secondary mucosal inflammation (Wehkamp *et al.*, 2005). These authors suggested that the main clinical phenotypes of IBDs, ileal and colonic Crohn's disease, could be linked to a characteristic lowered defensin profile. They found ileal disease is associated with NOD2 mutations that have been shown to reduce the expression of the main ileal alpha-defensin molecules (HD5 and HD6) in Paneth cells, and conversely, Crohn's colitis is associated with induction of the epithelial beta-defensins HBD-2 (*DEFB4*) , HBD-3 (*DEFB103*), and HBD-4 (*DEFB104*).

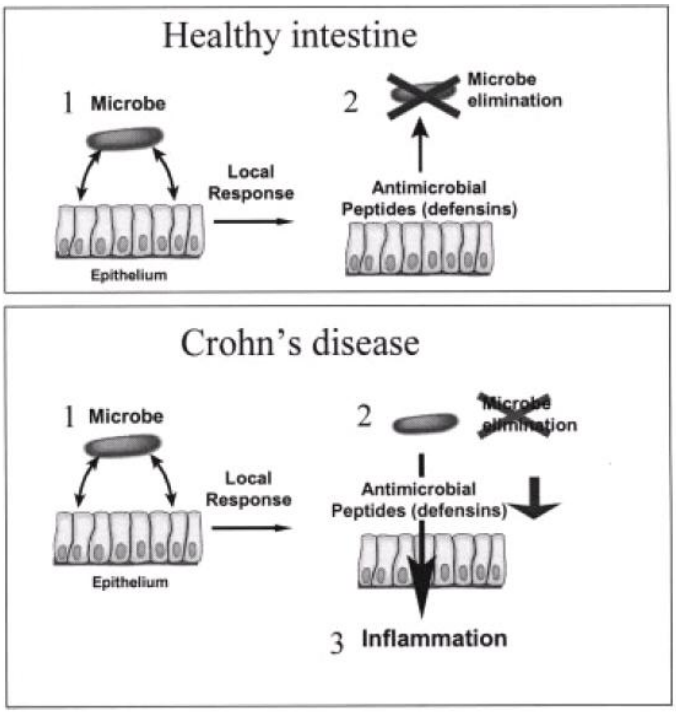


Figure 32: Simplified model of the normal reaction (upper) in the healthy intestine as well as the defective antimicrobial barrier in the intestine of Crohn's disease patients. In the healthy gut, the microbes cannot invade the mucosa because of an effective antimicrobial barrier (defensins). In Crohn's disease (lower), especially in patients with ileal disease, this antimicrobial barrier is disturbed, and bacteria can invade the mucosa. According to this hypothesis, a bacterial invasion as a result of a defensin deficiency is the primary reason for the secondary inflammation (taken from Wehkamp *et al.* (2005)).

Investigations on the relationship between human defensin genes and CD phenotypes have continuously been studied when all these defensins are found to be coordinately underexpressed in colonic CD. Furthermore there was a study on the human beta-defensin copy number changes that are positively correlated with the expression of HBD-2 in leucocytes (Hollox *et al.*, 2003). Fellermann *et al*. (2006) therefore measured DNA copy number in the beta-defensin cluster in two independent cohorts with inflammatory bowel diseases and related it to mucosal HBD-2 gene expression. They concluded from a small number of samples that a lower HBD-2 gene copy number in the beta-defensin locus predisposes to colonic CD, most likely through diminished beta-defensin expression (Fellermann *et al.*, 2006). Meanwhile, Bentley and co workers have challenged those data with their contrasting results which indicate that elevated *DEFB4* copy number is a risk factor for CD (irrespective of intestinal location) (Bentley *et al.*, 2009). To overcome these issues, we therefore proposed our triplex assay as the applicable technique to be applied on large case-control association studies of Crohn's disease to be compared with the previous studies. Understanding the beta-defensin copy number roles in this disease would also drastically change the understanding of therapy, and new strategies would try to strengthen the barrier function and protective innate immunity.

4.2: Results

4.2.1: Testing and quality control

For each single experiment of typing copy number at the beta-defensin region in this study, standard samples from ECACC panel 1, C11 (n = 4), C18 (n = 3), C62 (n = 5) and C66 (n = 6) which always gave reproducible results from several PRT assays, and for which the beta defensin copy number has been clarified with other methods and other laboratories, were included as internal controls and used for calibration of the beta-defensin copy number from PRT assays as shown in figures 4a, 4b and 4c (section 2.3.4). However, due to the appearance of a spurious 'dye peak' close to one of the HSPD21 peaks, as shown in figure 33, and after comparing the quality of data produced from peak area and peak height (data not shown), all the data produced from this system in this study were collected by using peak height. All the three dye peaks are shifted from the true peak sizes as shown in figure 33(a) and 33(c). Hence, one of the HSPD21 PCRs used NED labelling and produced a range of mobilities close to the NED dye peak, and if there is a shift in the NED dye peak to overlap with the real peak of HSPD21, data from peak height seems to be most reliable for further analysis.

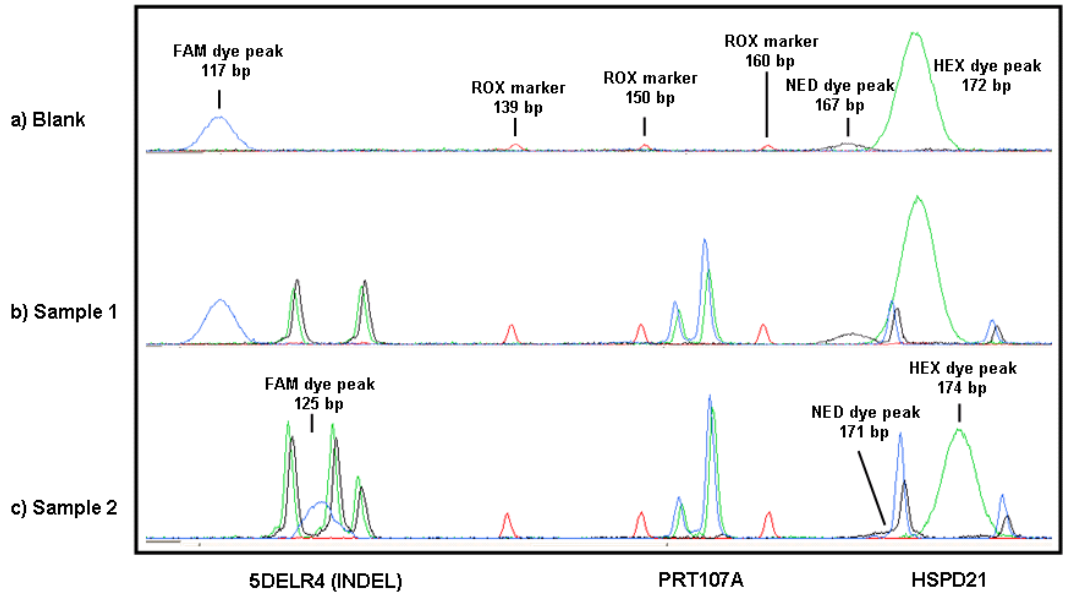
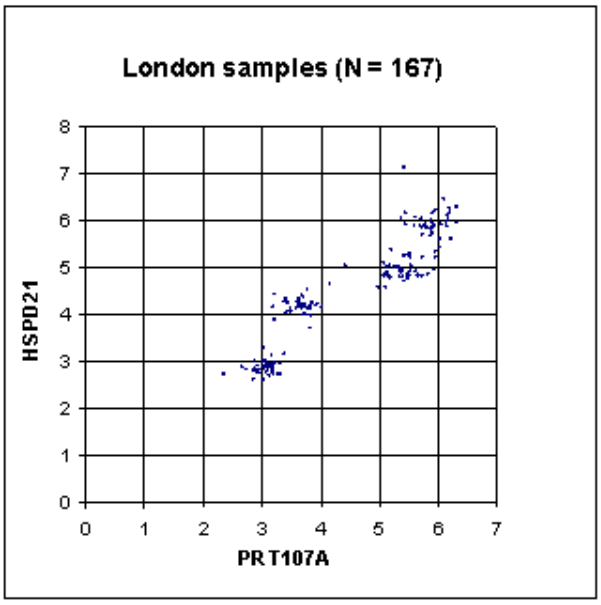


Figure 33: The appearance of dye peaks produced from three different labels on the multiplex PRT-based system, FAM, HEX and NED with their apparent sizes. a) Location of dye peaks from FAM, NED and HEX in a blank control with mobilities corresponding to apparent sizes of 117 bp, 167 bp and 172 bp respectively. b) The appearance of dye peaks is similar for sample 1 compared to blank, but in c) sample 2 showed relatively shifted dye peaks from the original location in the blank.

Beta defensin typing for these four standard samples with the triplex system was investigated by looking at the clustering of PRT107A and HSPD21 ratios as shown in figure 34. As expected, both typings from PRT assays in this system produced clear clusters of results corresponding to copy number diplotypes of 3, 4, 5 and 6, either in the analysis of London samples or Edinburgh samples for Crohn's disease. Perhaps due to the different methods of preparation for standard DNA samples, very tight

clusters are observed for Edinburgh data compared to London data for 3 and 4 copies, but the clustering is similar for 5 and 6 copies. For C11 samples (4 copies) the cluster does not fall exactly at n = 4.0 for PRT107A, and we assume the cause was some variation in the degree of amplification from a third locus to yield products indistinguishable from the reference locus product (chromosome 2); this has only one mismatch, and variable amplification from this source might make the yield of product from the reference locus appear more variable for PRT107A. When we used that ratio to predict the copy number for this sample in PRT107A, the unrounded copy number appeared less than 4.0 but still close to that integer. Nevertheless, there were 326 tests of the above reference samples typed from the triplex system for two cohorts and an incorrect copy number was called on only 11 occasions in all cases involving 5- and 6- copy samples. This suggested that the per-test error rate is of the order of 3-4%. The triplex system then was also applied to the collection of reference samples from ECACC panel 1 and panel 2.

a)



b)

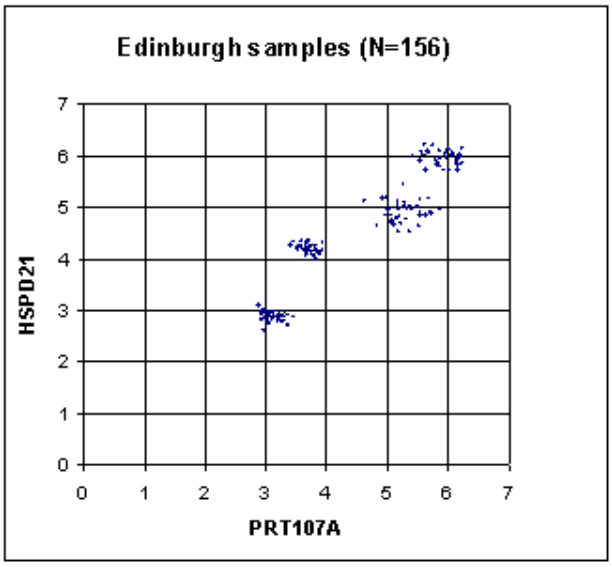


Figure 34: Results produced by two PRT assays (PRT107A and HSPD21) in the triplex system for multiple typing of four standard samples. **(a)** Scatterplot of triplex results that agree between PRT107A and HSPD21 for London data showing very clear clustering and **(b)** Scatterplot of PRT results for Edinburgh data showing very clear clustering.

4.2.2: Application to case-control association studies of Crohn's samples

4.2.2.1: Genotyping completion and concordance

Beta defensin genomic copy number was analyzed in 732 samples of Crohn's disease patients from Professor Christopher Mathew (section 2.1.1.3) and 185 controls (from ECACC panel 1 and 2). Furthermore, a total of 350 controls and 384 colonic disease patients received from Edinburgh collaborators (section 2.1.1.3) were also genotyped by using the same triplex system. For the Edinburgh group, 36 controls and 26 CD samples failed to give any genotype, and after removal of the duplicated samples in the London group only 666 samples were taken into account, from which 18 samples failed to give results. Independent analysis was carried out for acceptable results obtained from both groups; 648 cases and 185 controls from London samples and 358 cases and 314 samples provided by Edinburgh samples.

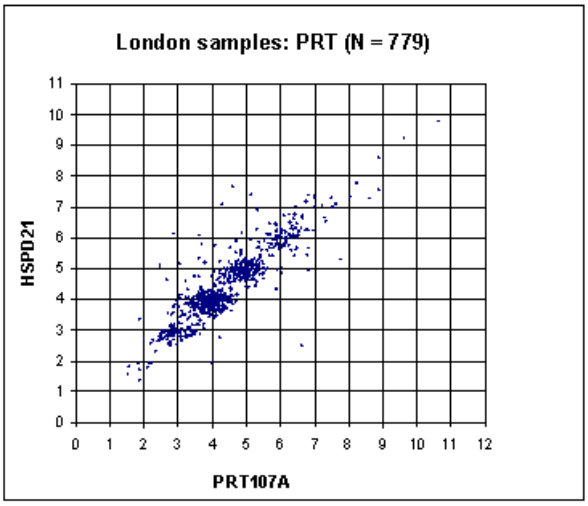
Before proceeding to further analysis, several procedures were applied to the acceptable results to achieve a high confidence in the results produced by triplex system. This began with an inspection of the discordances of copy number measurement between different components of the assay in the triplex system and/or where the support for the maximum likelihood copy number (ML CN) was relatively low. This applied to 48 samples for the London group and 51 samples for the Edinburgh cohort. To continue with this procedure, the mean of the two PRT results from the triplex assay

was calculated and the peak information from 5DELR4 (such as peak ratio and number of the peaks) was included in this analysis. Then, ML CN was compared to the rounded mean of the two PRTs. If both of them agreed between each other then, the ML CN was not changed. Whenever they did not agree, investigation of the information from 5DELR4 peak was carried out. If the 5DELR4 information had very strong concordance evident either with the rounded mean of the two PRTs or with the ML CN, the final copy number followed the one that agreed with 5DELR4. Whenever neither of them (rounded mean of two PRT nor ML CN) was consistent with the 5DELR4 information, analysis of two microsatellites, (EPEV1 and EPEV3) (section 2.4) was carried out on 35 of 99 samples in order to clarify the correct copy number. Allele ratios from microsatellite analyses were used as the final criterion in this procedure. Strong evidence from microsatellites to support any of the results from ML CN, rounded mean of the two PRTs or 5DELR4 was taken into account to alter the ML CN. If none of the above gave strong evidence to support changing the ML CN, then the ML CN was unchanged. As a result, there were 20 cases of altered copy number in this study.

Then analysis of the accuracy of beta defensin copy number measurement by this system was made by looking at the clustering of two PRT assays as shown in figure 35. As before, clustering around integer copy numbers of 2, 3, 4, 5, 6 and 7 in the two cohort studies has been clearly produced. In this study, the initially calibrated distributions for PRT107A values in cases

from both cohorts, and controls from the Edinburgh collection was not clustered exactly on the integer values (Figure 36); by contrast, the ECACC samples used as controls for the London-based cases showed no such shift in the centre of PRT107A clusters (Figure 37). This implies the sensitivity of copy number values measured by PRT107A to differences in DNA preparation methods. Therefore, to avoid propagating errors into further analyses, a linear transformation was applied to all PRT107A values from non-ECACC samples to improve the consistency with integer clustering.

a)



b)

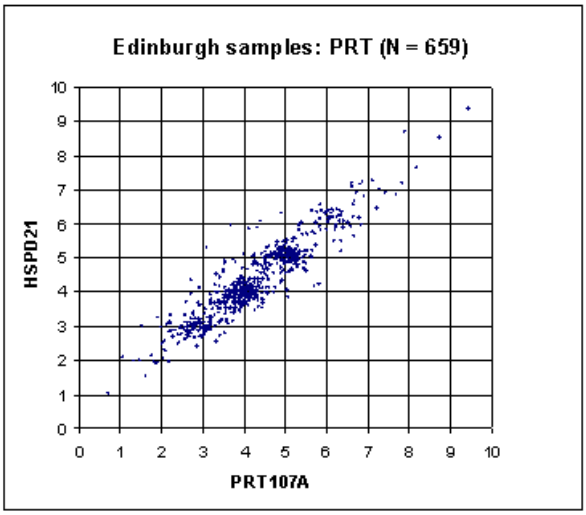


Figure 35: Results produced by two PRT assays (PRT107A and HSPD21) in the triplex system for two different cohorts; London and Edinburgh. (a) Scatterplot of triplex results that agree between PRT107A and HSPD21 for samples received from London showing very clear clustering and **(b)** Scatterplot of PRT results for samples received from Edinburgh showing very clear clustering.

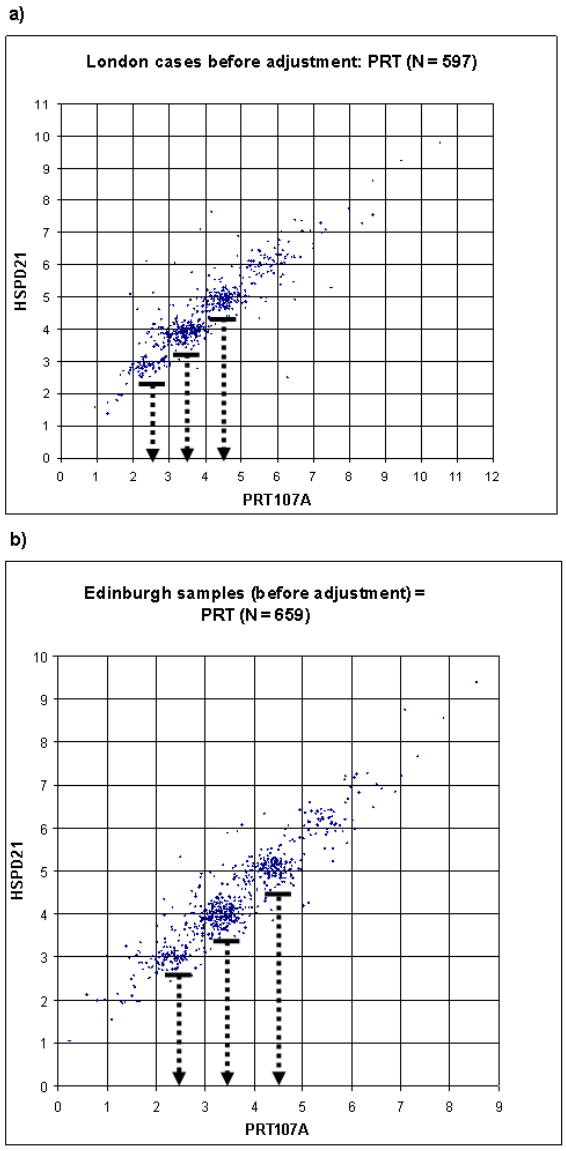


Figure 36: Initial results produced by two PRT assays (PRT107A and HSPD21) in the triplex system for two different cohorts; Edinburgh and London before a linear adjustment. (a) Scatterplot of triplex results that showing PRT107A values shifted from the integer copy number compared to HSPD21 values for samples received from London and **(b)** Scatterplot of PRT107A results for samples received from Edinburgh showing a similar shifted compared to HSPD21.

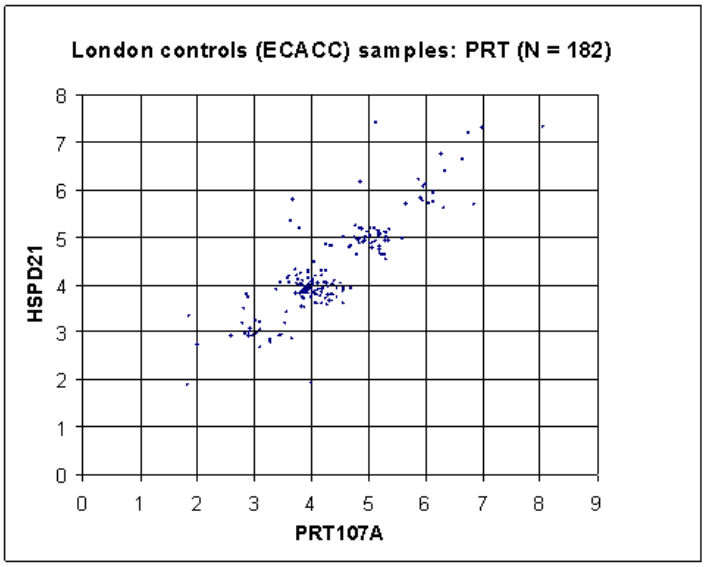
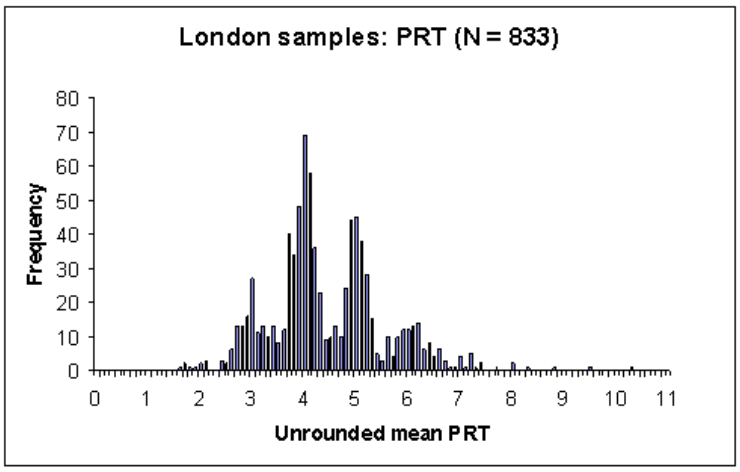


Figure 37: Initial results produced by two PRT assays (PRT107A and HSPD21) in the triplex system for ECACC panel 1 and 2 used as controls for Crohn's disease analyses from London samples. No shift of PRT107A values compared to HSPD21 is seen for the ECACC controls.

Another way to show the accuracy of the copy number measurement, namely examination of the frequency distribution of copy number measured by triplex system, was carried out; a histogram of the unrounded mean PRT, which was calculated from both PRT (PRT107A and HSPD21) was plotted for both cohorts as shown in figure 38. In both cases, it clearly showed the grouping of data around each integer copy number even though there was no clear gap between each copy number division; rather, there was an overlap with each division at the end of the left and right sides for each copy number group.

The PRT107A values appeared to show quite a spread at each end of the distribution for certain samples in both cohorts (Figure 39(a) and 40(a)) in nearly all the groups of copy numbers compared to the HSPD21 values which produced tighter distributions for each copy number (Figure 39(b) and 40(b)). As explained in section 4.2.1, this is perhaps due to differences in the batches of sample preparation, some small degree of amplification from another chromosome (chromosome 2 which only has one mismatch) or not enough amplification from the copy number variable locus. As a result, the PRT107A values for certain samples did not achieve the correct copy number. Thus, distributions of copy number from unrounded mean PRTs overlapped between clusters for copy number of three, four and five (Figures 39(c) and 40(c)).

a)



b)

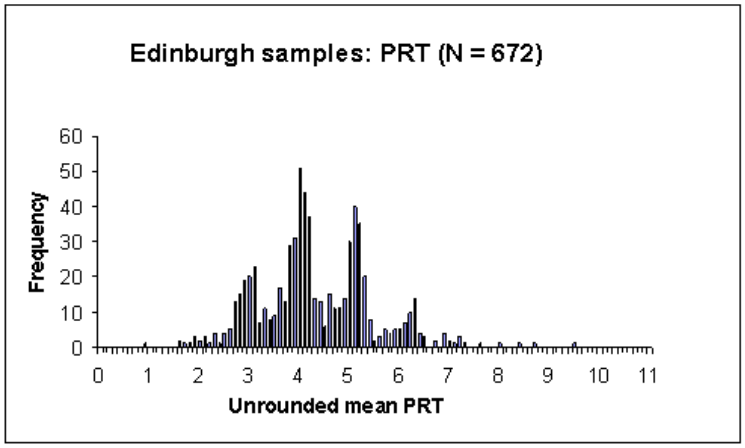


Figure 38: Histogram of unrounded mean PRT calculated from both PRT107A and HSPD21 from two different cohorts. (a) Distribution of mean PRT for each copy number showing very clear clustering in the London group. (b) Distribution of mean PRT for each copy number also showing very clear clustering in the Edinburgh group.

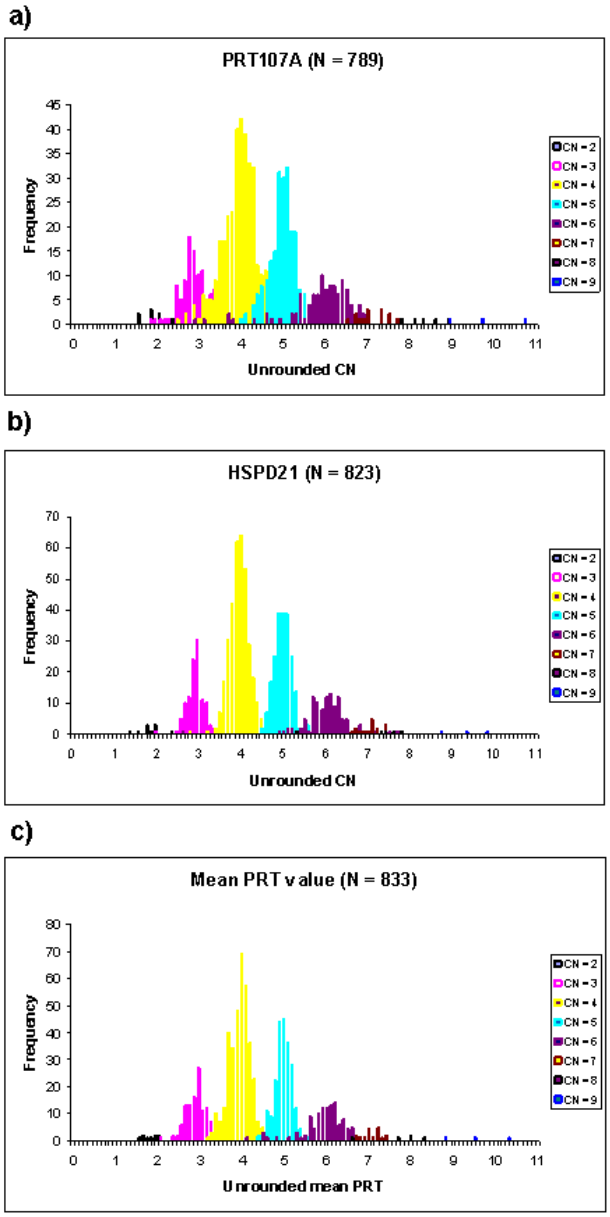


Figure 39: Distribution of each copy number from London samples coloured according to final ML CN value. (a) Distributions of PRT107A values for each copy number. (b) Distributions of HSPD21 for each copy number. (c) Distributions of each copy number from unrounded mean PRT.

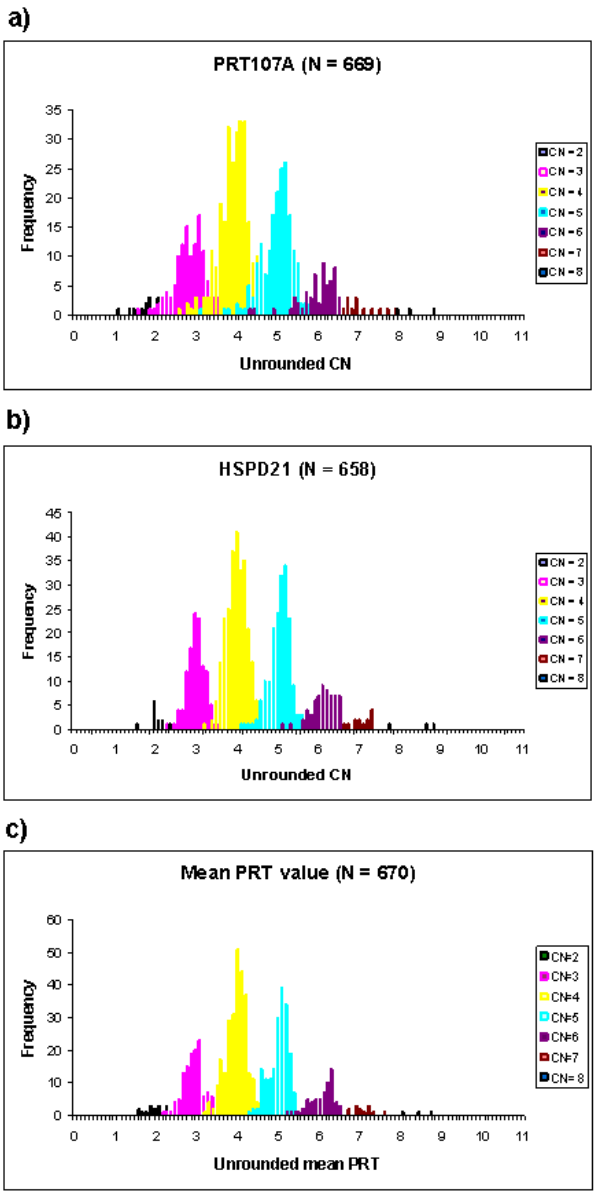


Figure 40: Distribution of each copy number from Edinburgh samples coloured according to final ML CN. (a) Distributions of PRT107A values for each copy number. (b) Distributions of HSPD21 for each copy number. (c) Distributions of each copy number from unrounded mean PRT excluded two samples which correspond to the copy number of 1 and 9.

In this study, 821 Edinburgh samples were additionally typed using real-time PCR methods by Marian C. Aldhous (Edinburgh collaborator). Analysis of clustering around integer values from real-time PCR gave a strong correlation between the two methods. However, it also produced a wide range of overlap between results from neighbouring integer copy number (Figure 41), in contrast to the results from PRT shown in figure 35(b) with a distinct clustering around integer values. Distribution of unrounded mean copy number from both methods of real-time PCR was also plotted in order to examine the clustering of each copy number (Figure 42). Neither Def:ΔΔct nor Def:Alb showed clear grouping for each copy number.

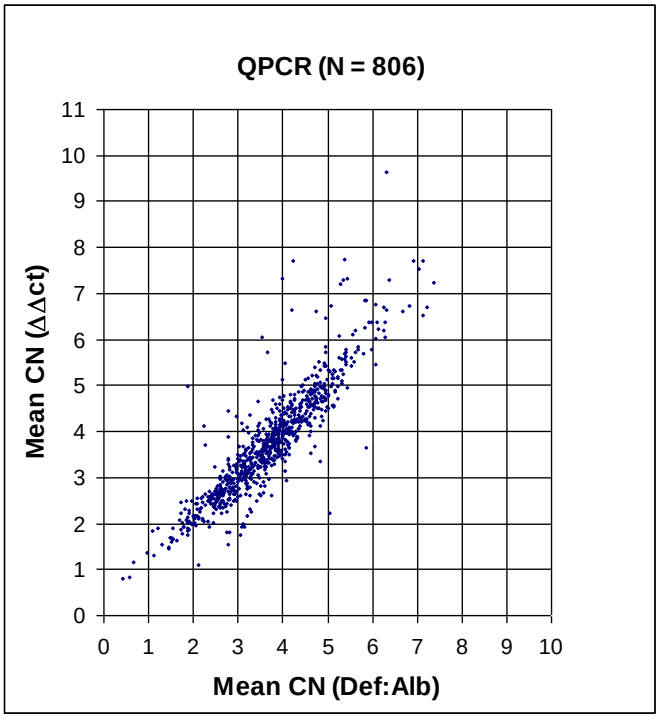
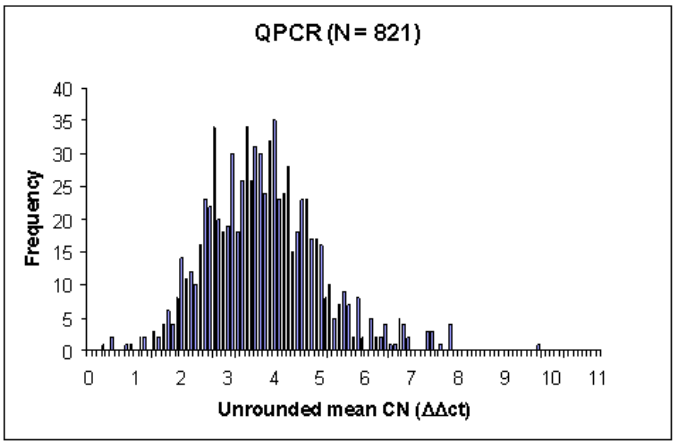


Figure 41: Scatterplot produced by two methods of real-time PCR assay (Def:Alb and Def:ΔΔct) for Edinburgh samples.

a)



b)

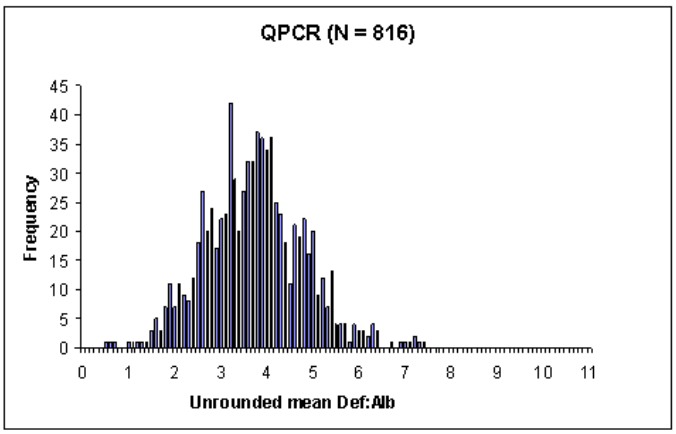
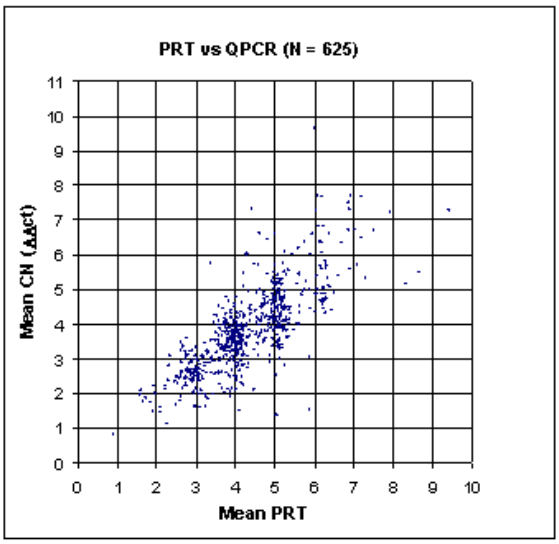


Figure 42: Histogram of unrounded mean copy number of duplicated typing calculated from both methods of real-time PCR; standard-curve Albumin and ΔΔct. (a) Distribution for each copy number from ΔΔct (b) Distribution for each copy number from standard-curve controls.

Of 821 samples, there were 625 (322 case and 302 control samples) for which CNV measurement data were obtained on the same DNA samples both from real-time PCR and the PRT-based triplex assay. Either using the values from ΔΔct or standard curve methods of real-time PCR, clustering

around integer values from those samples showed a wide spread, in which some results from real-time PCR fell outside the integer range of each copy number (Figure 43). Concordances between two assays (PRT-based triplex assay and real-time PCR (both methods)) for those 625 samples were also examined, with only 44% of integer copy number values from ΔΔct and 47% from standard curve analyses from real-time PCR measurements agreeing with corresponding integers deduced from the triplex assay.

a)



b)

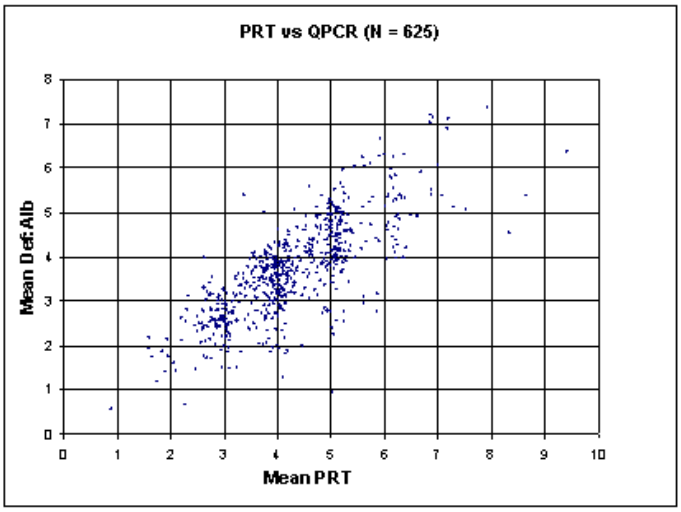


Figure 43: Comparison of results between two PRT assays (PRT107A and HSPD21) in the triplex system with real-time PCR data obtained from Edinburgh collaborators. (a) Scatterplot of results comparing mean PRT and standard-curve method of real-time PCR showing very broad range which some results from real-time PCR is falling outside the cluster for each copy number and **(b)** Scatterplot of mean PRT and ΔΔct of real-time PCR results showing small improvement in clustering of raw real-time PCR

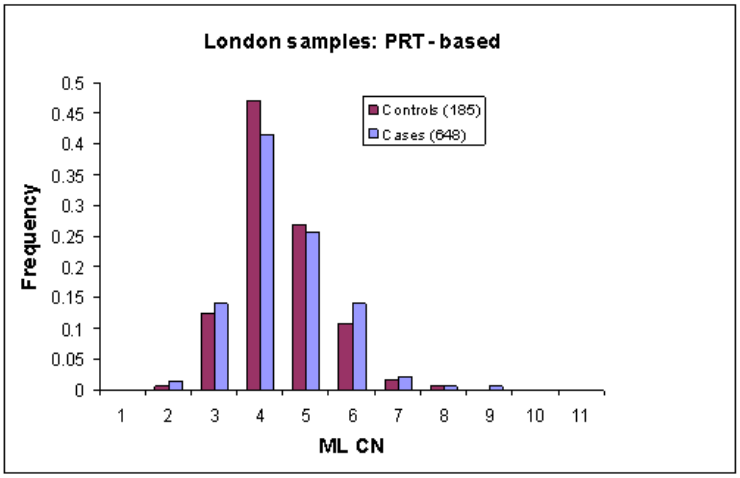
data but still with much of the data from real-time PCR falling outside the clusters.

4.2.2.2: Association results

By using the triplex PRT-based assay, this study found that the beta-defensin copy number varied between two to eight copies per diploid genome in controls and two to nine copies between colonic disease (CD) cohort (figure 44(a)) for the London group. For the Edinburgh group, between one to nine copies for controls and two to eight copies for CD cases (Figure 44(b)). There was no significant association determined between beta-defensin copy number and Crohn's disease for the London samples with the T-test P value calculated as 0.489 (Figure 44(a)). However, due to an increased frequency of 5-copy samples among Crohn's patients in Edinburgh samples, the mean copy number between controls (4.236) and cases (4.397) appeared to be slightly elevated, with the T-test P value given as 0.049. A similar increase of mean copy number relative to controls was detected using $\Delta\Delta$ct analysis of real-time PCR data which was 3.64 for controls and 3.82 for cases (Figure 45(a)) with a T-test P value of 0.025. However, there was no evidence of increasing mean copy number relative to the controls in standard-curve analysis of the same samples, (mean copy number for controls was 3.65 and 3.75 for cases (T-test P value = 0.248) (Figure 45(b)). Additionally, the same 303 control DNA samples from both assays (PRT-based triplex and real-time PCR)

were compared, and the PRT-based triplex data had a mean copy number of 4.23, but the same samples have a mean of 3.79 with real-time PCR/$\Delta\Delta$ct (T-test P value = 1.26×10^{-5}). This analysis suggests that real-time PCR methods could introduce an error in copy number typing sufficient to produce a false impression of a significant association.

a)



b)

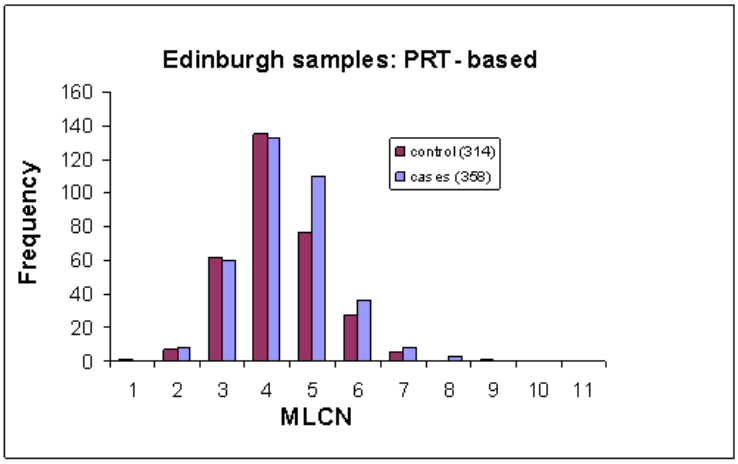
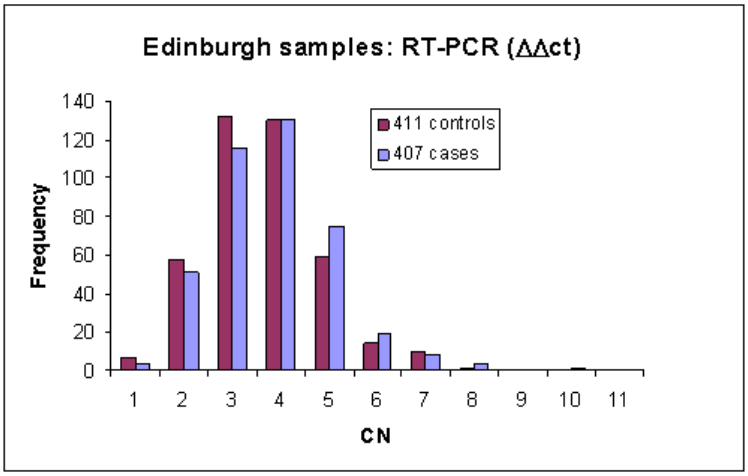


Figure 44: Distribution of copy number values from the PRT-based triplex assay. (a) Comparison of copy number frequency between ECACC controls and cases from London samples (the P-value is 0.489) and (b) comparison of copy number frequency between controls and cases from Edinburgh samples (the P-value is 0.049).

a)



b)

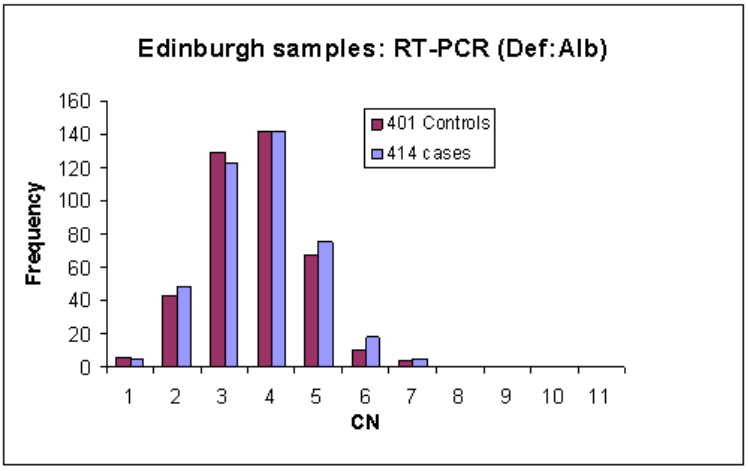
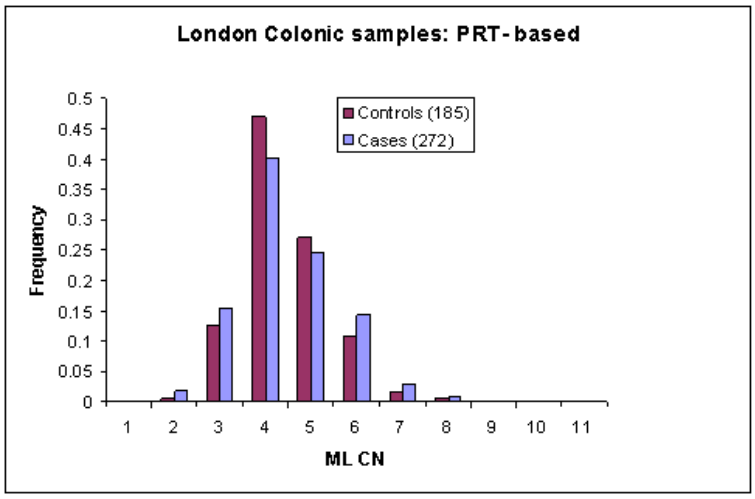


Figure 45: Distribution of copy number values from the real-time PCR methods for Edinburgh samples. (a) Comparison of copy number frequency between controls and cases obtained from ΔΔct method (P = 0.025) and (b) comparison of copy number frequency between controls and cases from standard-curve method (P = 0.248).

Furthermore, there were no significant associations determined between beta-defensin copy number either in colonic or ileal samples for the London samples with the T-test P value calculated as 0.704 (Figure 46 (a)) and 0.403 (Figure 46 (b)) respectively. A similar finding of no significant association for beta-defensin copy number on colonic and ileal samples from Edinburgh with the T-test P value were 0.445 (Figure 47 (a)) and 0.439 (Figure 47 (b)) respectively. Nevertheless, due to an increased frequency of 5-copy samples among Crohn's patients in pooled ileal samples, the mean copy number between controls (4.305) and ileal samples (4.467) appeared to be slightly elevated, with the T-test P value given as 0.015 (Figure 48 (b)).

a)



b)

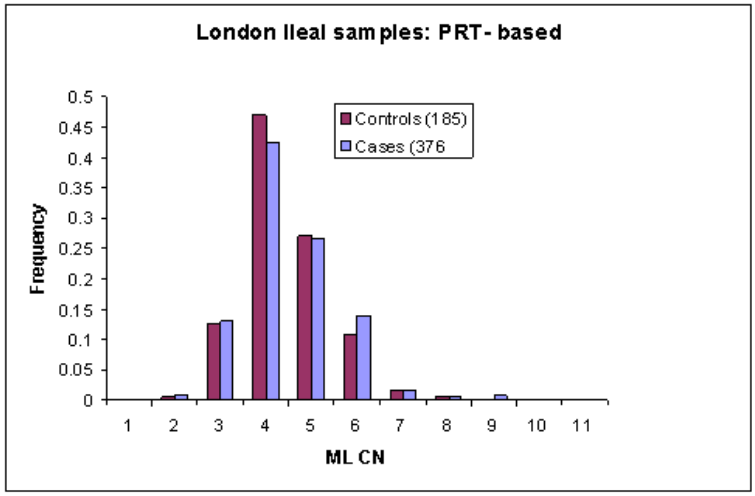
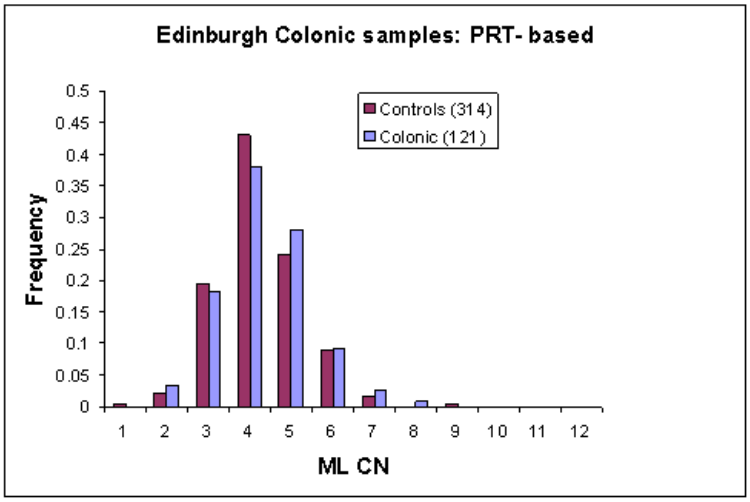


Figure 46: Distribution of copy number values from the PRT-based triplex assay in London group. (a) Comparison of copy number frequency between ECACC controls and colonic samples with a P-value of 0.704 and (b) comparison of copy number frequency between controls and ileal samples with the P-value of 0.403.

a)



b)

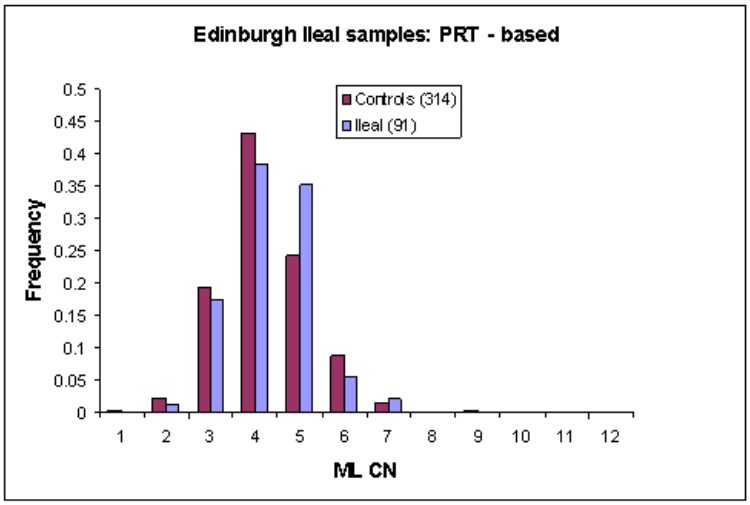
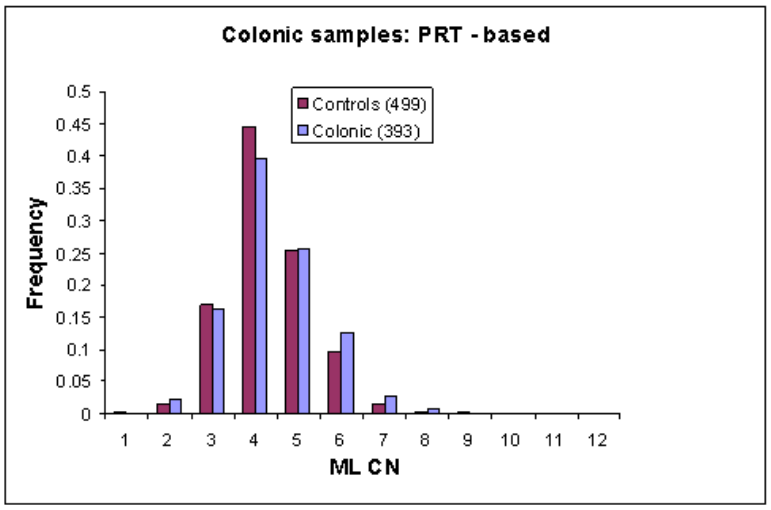


Figure 47: Distribution of copy number values from the PRT-based triplex assay in Edinburgh group. (a) Comparison of copy number frequency between controls and colonic samples (P = 0.445) and (b) comparison of copy number frequency between controls and ileal samples with a P-value of 0.439.

a)



b)

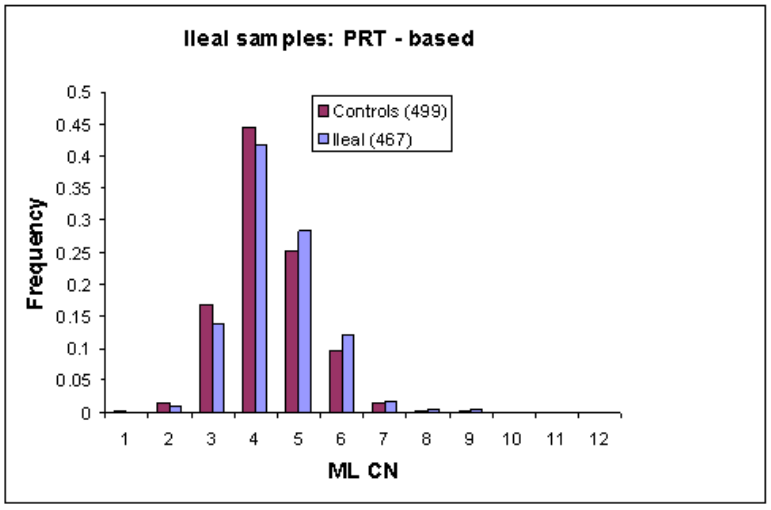


Figure 48: Distribution of copy number values from the PRT-based triplex assay for both types of Crohn's disease. (a) Comparison of copy number frequency between pooled controls and pooled colonic samples from London and Edinburgh groups (P = 0.113) and (b) comparison of copy number frequency between pooled controls and pooled ileal samples from both groups (P = 0.015).

4.3: Discussion

This study aimed to develop an efficient and powerful method to accurately measure beta-defensin copy number variation suitable for large case-control association studies. All the analyses revealed that the PRT-based triplex assay which has been used to measure beta defensin copy number in Crohn's disease in this study is an accurate method that meets the above criteria. Reproducibility of the results was investigated by repeated testing of internal reference samples in which the beta-defensin copy number was known from other typing methods and showed a clustering around integer values (section 4.2.1, figure 34), and suggested a high standard of precision for this triplex assay with a per-test error rate of the order of 3-4%. Therefore, calibrating each copy number produced from the Crohn's disease samples to the centres of these clusters (Figure 35) overcame potential accuracy differences between experiments. This clustering enabled the final copy number identification and resulted in six distinct clusters between two and seven gene copies per diploid genome.

Furthermore, this study has also provided evidence for the power of copy number measurements that comes from combining methods and repeat typing (Hollox *et al.*, 2008), even though the reference locus of the PRT107A assay appeared to include small degree of amplification from another chromosome (chromosome 2) which only has one mismatch, and which would result in a reduced copy number value in the affected system,

but would not affect the other two assays (HSPD21 and 5DELR4). Conversely, there is also no significant evidence of 'drop-out' of one or more copies of the reference locus that would lead to apparent doubling of the copy number of the test locus (Armour *et al.*, 2007), and neither of two PRT assay results showed a discrepancy of exactly double the consensus value. Multiplex PRT was firstly described by Walker *et al.* (2009), who used three combinations of PRT systems to measure copy number at *CCL3L1* and *CCL4L1* genes. They demonstrated that this multiplex PRT produces accurate and reproducible data in the copy number range common in Europe but that a further development is necessary for a robust typing system for samples of African origin, in which copy numbers of *CCL3L1* appear to range as high as 14 (Walker *et al.*, 2009).

Comparison of integer clustering between ECACC (panel 1 and 2) and non-ECACC samples (London cases and Edinburgh controls and cases) (Figure 36 and 37) on the initial results (before linear adjustment has been carried out effectively) showed that different methods of DNA preparation could make a difference to the typing copy number. This experience has reinforced the idea that great care should be taken throughout the investigation in case-control association studies using multiallelic copy number variants (McCarroll, 2008). Moreover, side-by-side typing of beta-defensin copy number variation by two methods on the same samples that produced different results might also be evidence for influence from

different methods of DNA preparation and perhaps, the storage and handling of DNA samples (Clayton *et al.*, 2005).

The comparison between our PRT-based triplex assay and real-time PCR results from the same samples (Figure 43) strongly suggests that real-time PCR measurements are insufficiently accurate for reliable copy number variable typing. There is a high correlation between the two real-time PCR results despite the low accuracy of copy number typing which is actually subject to measurement error of sufficient magnitude to obscure the resolution of results into integer-centred copy number classes (Figure 41), as is also apparent from the general failure of real-time PCR methods to resolve copy number cluster clearly for N > 2 at the *CCL3L1* copy number variant (Field *et al.*, 2009).

Obviously, our results from the PRT-based triplex assay do not provide strong support for the finding of Bentley *et al.* (2009) that there is a significantly increased beta-defensin copy number among Crohn's disease patients. The investigators found a mean increase in copy number of 0.41 repeats among affected individuals of European origin from New Zealand with a mean copy number of 4.36 in cases and 3.95 in controls (Bentley *et al.*, 2009). Their findings contrast with this study that found no clear increase in mean copy number in either cohorts, London and Edinburgh (4.48 for cases and 4.42 for controls) and (4.40 for cases and 4.24 for controls) respectively. Certainly, this work is in addition unable to provide

any support for the conclusions of Fellermann *et al.* (2006), that the copy number of beta-defensins in Crohn's disease of the colon is significantly reduced (Fellermann *et al.*, 2006). Both of the previous studies Bentley *et al.* and Fellermann *et al.* depended on real-time PCR methods for analysing beta-defensin copy number (Fellermann *et al.*, 2006; Bentley *et al.*, 2009). Analysis between different phenotypes of Crohn's disease in this study also do not illustrate any genuine association of the beta-defensin copy number either with colonic or ileal Crohn's disease.

4.4: Conclusion

In conclusion, the PRT-based triplex assay, which has been successfully applied in this study, is an accurate and robust method that would be sufficiently accurate for measuring copy number variation in large case-control association studies. Even though this assay contains three components - two PRTs (PRT107A and HPD21) and a multi-allelic indel (rs5889219) measure, it uses a single capillary electrophoresis column per sample, and so this PRT-based triplex assay remains simple and inexpensive in practice for investigation of a single copy number variable locus in a large sample set. Without any doubt the copy number values produced from these three components has increased the power of copy number measurement of the beta-defensin variation. This assay has illustrated strong reliability in the quantification of low copy number to high

copy number (two to seven copies) by clear clustering and its ability to discriminate between each integer class of copy number.

Chapter 5: General Discussion and Conclusion

5.1: Origins of diversity at human beta defensin copy number

In chapter three, copy numbers of the human beta-defensin repeat between two and eight copies per diploid genome were identified in 26 CEPH families originating from European populations; there were one to five copies of the locus per haploid genome, with two and three copies being the most common. Pedigree analyses within CEPH kindreds in this study allowed us to trace the origin of parental haplotypes of these beta-defensin copy numbers. Moreover, the origin of this copy number variation has been discovered by investigation of recombinant haplotypes inherited in children and is due to simple allelic crossing-over events between homologous chromosomes during meiosis. As discussed in section 1.1.4 and illustrated in figure 3, the chromosome 8p23.1 region is flanked by two highly homologous sequences of low copy repeats (LCRs) or segmental duplications (SDs) named REPD distally and REPP proximally. The beta-defensin gene clusters located on chromosome 8p23.1 that are annotated on the genome assembly are found to be flanked with low copy repeats (LCRs) distally (REPD). SD blocks are known to act as mediators for the exchange of DNA sequence through non-allelic homologous recombination (NAHR) from many examples involved with duplications and deletions associated with genomic disorders, for example, the alpha-globin locus

(Higgs *et al.*, 1989), *PMP22* genes (Reiter *et al.*, 1998), the *RHD* genes (Wagner and Flegel, 2000) and genes which encode for colour vision variation (Deeb, 2004) and even deletion and reciprocal duplication of 8p23.1 (Giglio *et al.*, 2001).

Subsequently in this study, allelic recombination events discovered from segregation of copy number, microsatellite and multiallelic length polymorphism (indels) in the pedigree analyses, together with linkage analysis of other markers in this region has revealed the beta-defensin gene clusters in two different locations including a new placement near REPP, about 5 Mb away from the distal location, REPD. The precise crossing-over points between the two parental haplotypes in 24 individual's recombinant haplotypes are placed within the interval corresponding to the 8p23.1 region, but all the beta-defensin genes were consistently mapped either onto the accepted location at REPD or to another site close to REPP, the proximal set of SDs on 8p23.1.

Construction of the haplotype blocks from the sequence variants is essential to distinguish the beta-defensin repeats between two sites at REPD and REPP. Comparisons of sequence reads from recombinant CEPH families to the reference sequence genome have shown that inversion events might contribute to the exchange of the beta-defensin locations since the identical sequence haplotypes have been observed at both sites (table 10). The inversion polymorphism between REPD and

REPP in the chromosome band 8p23.1 region has been well documented from previous studies. Nevertheless, none of the previous studies has considered information about the beta-defensin clusters at the REPP sites. Investigation of the exact breakpoint of the inversion polymorphism, which may include the beta-defensin region, is a challenging problem because these repeats are situated in a complex segmental duplication (REPD). These duplications have been separated by a gap in the assembled sequence, and only with evidence from recombinant haplotypes could this copy number variable region be placed at REPP. The other problem is there could be several different types of inverted alleles with different precise breakpoints.

Furthermore, the beta-defensin copy number seems to be more variable proximally (REPP) than distally (REPD) as demonstrated in table 7. Thus, investigation of copy number variation by using a single nucleotide polymorphism close to the accepted REPD location might not show an association with copy number variation. A proper analysis of the sequence variants for the whole repeat at both sites may provide important information on the possibility of using linkage disequilibrium (LD) with SNPs to predict copy number at the beta-defensin region. Such structural genomic variation involving the beta-defensin gene clusters has provided us insight into the origin for diversity of copy number. In general, this study has contributed valuable information on the beta-defensin mapping in the human genome and human chromosome evolution, based on family

analysis. Moreover, this finding could lead to the development of more efficient strategies to identify copy number variants within this complex region that are suspected to have clinical consequences.

5.2: Development of a multiplex PRT measurement system for beta-defensin copy number variants

In chapter four, a multiplex PRT measurement system was applied to accurately measure the multiallelic copy number polymorphism of the beta-defensin region. The process of defining and mapping CNVs has been facilitated by many methods. These include array-based CGH using oligonucleotides and bacterial artificial chromosome clones, comparison of clone paired-end sequence to the reference human assembly and detection of deletions and duplications based on single nucleotides polymorphisms (Iafrate *et al.*, 2004; Tuzun *et al.*, 2005; Redon *et al.*, 2006; Kidd *et al.*, 2008). However, these methods provide very little information in complex regions of copy number variation. Nowadays, investigating association between the possessions of common variants, for example, copy number changes and/or SNPs, and particular traits by using such high throughput methods has been the source of many discoveries about disease causation (McCarroll *et al.*, 2008). Nevertheless, simple, inexpensive, and accurate analysis measurements are still in demand particularly for the most challenging regions such as the beta-defensin genes. Results shown in chapter four have proved that this new system is

highly capable in classifying samples into a distinct (integer) copy number (figure 35), ranging from two to seven copies and possibly up to nine copies (figures 39 and 40). In addition, the beta-defensin copy numbers given by all the three methods in this system, are generally in concordance as described in section 4.2.2.1. Thus, in this study, assessment of our new measurement system based on PRT assays, with indel ratio measurement as a complementary method, have demonstrated highly accurate prediction of the beta-defensin copy number in an association study, especially when directly compared to RT-PCR. Perne *et al*. (2009) found in their study that two quantification methods, real-time PCR and MLPA were affected similarly by decreased reliability in the discrimination of higher copy numbers. Other studies on *CCL3L1* copy number determination highlighted the technical difficulty of obtaining convincingly accurate copy number measurements using methods based on real-time PCR, even in the relatively low copy number range of those CNV genes (Field *et al.*, 2009). Field *et al*. (2009) found a difference for an apparently significant association between the data from the same samples (several thousand cases and controls in a study of type I diabetes), using either real-time PCR or assays using the Paralogue Ratio Test (PRT). There was an apparently highly significant association between *CCL3L1* copy number and type I diabetes using real-time PCR data but no association with type I diabetes was replicated by PRT-based measurement of copy number (Field *et al.*, 2009).

Association studies based on a candidate gene is one very well known approach to examine the relationship between variation at a gene and predisposition to a multifactorial disease, for example Crohn's disease. Crohn's disease has provided a remarkable example in defining the genetic factors responsible for the observed clinical phenotype (Van Limbergen *et al.*, 2009). However, there are still significant issues when attempts to replicate association with specific genes and variants have failed in different kind of observations; for example, different measurement methods may have different accuracy and resolution, and there may be variation in the source and phenotype identity of the samples. Thus, the illustration of the beta-defensin copy number variable region may provide important understanding of the Crohn's disease pathogenesis and individual susceptibility in particular, and more generally application of multiallelic copy-number variants in case-control association studies. In order to examine association of copy number variants with the disease, an accurate assay was needed in order to measure precisely what is the correct copy number of the genes involved and to report any real relationship with the exact range of the copy number of the gene.

In this study, we have demonstrated our new system is a robust technique to determine the association of the beta-defensin genes in Crohn's disease. Small sample sizes are a common shortcoming of association studies. For example, our finding using more than 1000 samples did not provide any strong support for association reported from previous studies

that have highlighted an association of lower beta-defensin copy number with a small number of 71 colonic Crohn's patients and 169 controls (Fellermann *et al.*, 2006), and only weak support, for the association with higher copy number reported by Bentley *et al*. (2009). Minimizing heterogeneity between studies for example, variation within a phenotype related to aetiology, is important to maximize the chances of success. Investigations using the large samples size and clearly identified phenotypes of Crohn's disease in individual samples might overcome such problems.

REFERENCES

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." Proceedings of the National Academy of Sciences of the United States of America **106**(3): 853-858.

Aitman, T. J., Dong, R., Vyse, T. J., Norsworthy, P. J., Johnson, M. D., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A. J. and Petretto, E. (2006). "Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans." Nature **439**(7078): 851 - 855.

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." Human Molecular Genetics **14**(14): 2045-2052.

Antonacci, F., Kidd, J. M., Marques-Bonet, T., Ventura, M., Siswara, P., Jiang, Z. and Eichler, E. E. (2009). "Characterization of six human disease-associated inversion polymorphisms." Human Molecular Genetics **18**: 2555 - 2566.

Armour, J. A., Sismani, C., Patsalis, P. C. and Cross, G. (2000). "Measurement of locus copy number by hybridisation with amplifiable probes." Nucleic Acids Research **28**(2): 605-9.

Armour, J. A. L. (2006). "Tandemly repeated DNA: Why should anyone care?" Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **598**(1-2): 6-14.

Armour, J. A. L., Barton, D. E., Cockburn, D. J. and Taylor, G. R. (2002). "The detection of large deletions or duplications in genomic DNA." Human Mutation **20**(5): 325-337.

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 Research **35**(3): e19.

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W. and Eichler, E. E. (2002). "Recent segmental duplications in the human genome." Science **297**(5583): 1003-7.

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. and Eichler, E. E. (2001). "Segmental duplications: organization and impact within the current human genome project assembly." Genome Research **11**(6): 1005-17.

Barber, J. C., Joyce, C. A., Collinson, M. N., Nicholson, J. C., Willatt, L. R., Dyson, H. M., Bateman, M. S., Green, A. J., Yates, J. R. and Dennis, N. R. (1998). "Duplication of 8p23.1: a cytogenetic anomaly with no established clinical significance." Journal of Medical Genetics **35**(6): 491-6.

Barber, J. C. K., Maloney, V., Hollox, E. J., Stuke-Sontheimer, A., du Bois, G., Daumiller, E., Klein-Vogler, U., Dufke, A., Armour, J. A. L. and Liehr, T. (2005). "Duplications and copy number variants of 8p23.1 are cytogenetically indistinguishable but distinct at the molecular level." European Journal of Human Genetics **13**(10): 1131-1136.

Barrois, M., Bieche, I., Mazoyer, S., Champeme, M. H., Bressac-de Paillerets, B. and Lidereau, R. (2004). "Real-time PCR-based gene dosage assay for detecting BRCA1 rearrangements in breast-ovarian cancer families." Clinical Genetics **65**(2): 131-136.

Baumgart, D. C. and Sandborn, W. J. (2007). "Inflammatory bowel disease: clinical aspects and established and evolving therapies." The Lancet **369**(9573): 1641-1657.

Bennett, S. T., Lucassen, A. M., Gough, S. C. L., Powell, E. E., Undlien, D. E., Pritchard, L. E., Merriman, M. E., Kawaguchi, Y., Dronsfield, M. J., Pociot, F., Nerup, J., Bouzekri, N., Cambon-Thomsen, A., Ronningen, K. S., Barnett, A. H., Bain, S. C. and Todd, J. A. (1995). "Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus." Nature Genetics **9**(3): 284-292.

Bensch, K. W., Raida, M., Mägert, H.-J., Schulz-Knappe, P. and Forssmann, W.-G. (1995). "hBD-1: a novel [beta]-defensin from human plasma." FEBS Letters **368**(2): 331-335.

Bentley, R. W., Pearson, J., Gearry, R. B., Barclay, M. L., McKinney, C., Merriman, T. R. and Roberts, R. L. (2009). "Association of Higher DEFB4 Genomic Copy Number With Crohn's Disease." American Journal of Gastroenterology **105**(2): 354-359.

Bosch, N., Escaramís, G., Mercader, J. M., Armengol, L. and Estivill, X. (2008). "Analysis of the multi-copy gene family FAM90A as a copy number variant in different ethnic backgrounds." Gene **420**(2): 113-117.

Bosch, N., Morell, M., Ponsa, I., Mercader, J. M., Armengol, L. s. and Estivill, X. (2009). "Nucleotide, Cytogenetic and Expression Impact of the Human Chromosome 8p23.1 Inversion Polymorphism." PLoS ONE **4**(12): e8269.

Brice, A., Ravise, N., Stevanin, G., Gugenheim, M., Bouche, P., Penet, C. and Agid, Y. (1992). "Duplication within chromosome 17p11.2 in 12 families of French ancestry with Charcot-Marie-Tooth disease type 1a. The French CMT Research Group." Journal of Medical Genetics **29**(11): 807-812.

Chen, G. K., Slaten, E., Ophoff, R. A. and Lange, K. (2006). "Accommodating Chromosome Inversions in Linkage Analysis." American Journal of Human Genetics **79**(2): 238-251.

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D. and Todd, J. A. (2005). "Population structure, differential bias and genomic control

in a large-scale, case-control association study." Nature Genetics **37**(11): 1243-1246.

Cole, A. M., Hong, T., Boo, L. M., Nguyen, T., Zhao, C., Bristol, G., Zack, J. A., Waring, A. J., Yang, O. O. and Lehrer, R. I. (2002). "Retrocyclin: a primate peptide that protects cells from infection by T- and M-tropic strains of HIV-1." Proceeding of the National Academy of Sciences of the United States of America **99**(4): 1813-8.

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

Deeb, S. S. (2004). "Molecular genetics of color-vision deficiencies." Visual Neuroscience **21**(03): 191-196.

Diamond, G., Kaiser, V., Rhodes, J., Russell, J. P. and Bevins, C. L. (2000). "Transcriptional regulation of beta-defensin gene expression in tracheal epithelial cells." Infection and Immunity **68**(1): 113-9.

Edelmann, L., Pandita, R. K., Spiteri, E., Funke, B., Goldberg, R., Palanisamy, N., Chaganti, R. S., Magenis, E., Shprintzen, R. J. and Morrow, B. E. (1999). "A common molecular basis for rearrangement disorders on chromosome 22q11." Human Molecular Genetics **8**: 1157 - 1167.

Edwards, J. H., Harnden, D. G., Cameron, A. H., Crosse, V. M. and Wolf, O. H. (1960). "A new trisomic syndrome." The Lancet **275**(7128): 787-790.

Eichler, E. E. (2001). "Recent duplication, domain accretion and the dynamic mutation of the human genome." Trends in Genetics **17**(11): 661-669.

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

Fellermann, K., Stange, D. E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C. L., Reinisch, W., Teml, A., Schwab, M., Lichter, P., Radlwimmer, B. and Stange, E. F. (2006). "A Chromosome 8 Gene-Cluster Polymorphism with Low Human Beta-Defensin 2 Gene Copy Number Predisposes to Crohn Disease of the Colon." American Journal of Human Genetics **79**: 439-448.

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

Feuk, L., Marshall, C. R., Wintle, R. F. and Scherer, S. W. (2006). "Structural variants: changing the landscape of chromosomes and design of disease studies." Human Molecular Genetics **15**: R57-R66.

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." Nature Medicine **15**(10): 1115-1117.

Ganz, T. (1999). "Defensins and host defense." Science **286**(5439): 420-1.

Ganz, T. (2003). "Defensins: antimicrobial peptides of innate immunity." Nature Review Immunology **3**(9): 710-20.

Ganz, T. and Lehrer, R. I. (1995). "Defensins." Pharmacology & Therapeutics **66**(2): 191-205.

Ghosh, D., Porter, E., Shen, B., Lee, S. K., Wilk, D., Drazba, J., Yadav, S. P., Crabb, J. W., Ganz, T. and Bevins, C. L. (2002). "Paneth cell trypsin is the processing enzyme for human defensin-5." Nature Immunology **3**(6): 583-90.

Giglio, S., Broman, K. W., Matsumoto, N., Calvari, V., Gimelli, G., Neumann, T., Ohashi, H., Voullaire, L., Larizza, D., Giorda, R., Weber, J. L., Ledbetter, D. H. and Zuffardi, O. (2001). "Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements." American Journal of Human Genetics **68**: 874 - 883.

Giglio, S., Calvari, V., Gregato, G., Gimelli, G., Camanini, S., Giorda, R., Ragusa, A., Guerneri, S., Selicorni, A., Stumm, M., Tonnies, H., Ventura, M., Zollino, M., Neri, G., Barber, J., Wieczorek, D., Rocchi, M. and Zuffardi, O. (2002). "Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation." American Journal of Human Genetics **71**(2): 276-85.

Glenn, C. C., Driscoll, D. J., Yang, T. P. and Nicholls, R. D. (1997). "Genomic imprinting: potential function and mechanisms revealed by the Prader-Willi and Angelman syndromes." Molecular Human Reproduction **3**(4): 321-332.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J., Murthy, K. K., Rovin, B. H., Bradley, W., Clark, R. A., Anderson, S. A., O'Connell, R. J., Agan, B. K., Ahuja, S. S., Bologna, R., Sen, L., Dolan, M. J. and Ahuja, S. K. (2005). "The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility." Science **307**(5714): 1434-1440.

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

Groth, M., Karol, S., Stefan, T., Klaus, H., Oliver, M., Philip, R., Anders, O. H. N., Stefan, S., Gerd, B. and Matthias, P. (2008). "High-resolution mapping of the 8p23.1 beta-defensin cluster reveals strictly concordant copy number variation of all genes." Human Mutation **29**(10): 1247-1254.

Harder, J., Bartels, J., Christophers, E. and Schroder, J. M. (1997). "A peptide antibiotic from human skin." Nature **387**(6636): 861.

Harder, J., Bartels, J., Christophers, E. and Schroder, J. M. (2001). "Isolation and characterization of human beta -defensin-3, a novel human inducible peptide antibiotic." Journal of Biological Chemistry **276**(8): 5707-13.

Harwig, S. S., Park, A. S. and Lehrer, R. I. (1992). "Characterization of defensin precursors in mature human neutrophils." Blood **79**(6): 1532-1537.

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

Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. Y. and Frazer, K. A. (2006). "Common deletions and SNPs are in linkage disequilibrium in the human genome." Nature Genetics **38**(1): 82-85.

Holloway, K., Lawson, V. E. and Jeffreys, A. J. (2006). "Allelic recombination and de novo deletions in sperm in the human {beta}-globin gene region." Human Molecular Genetics: ddl025.

Hollox, E. and Armour, J. (2008). "Directional and balancing selection in human beta-defensins." BMC Evolutionary Biology **8**(1): 113.

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

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

Hollox, E. J., Davies, J., Griesenbach, U., Burgess, J., Alton, E. W. F. W. and Armour, J. A. L. (2005). "Beta-defensin genomic copy number is not a modifier locus for cystic fibrosis." Journal of Negative Results in BioMedicine **4**: 9.

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

Horsten, H. H. v., Schäfer, B. and Kirchhoff, C. (2004). "SPAG11/isoform HE2C, an atypical anionic [beta]-defensin-like peptide." Peptides **25**(8): 1223-1233.

Hugot, J.-P., Laurent-Puig, P., Gower-Rousseau, C., Olson, J. M., Lee, J. C., Beaugerie, L., Naom, I., Dupas, J.-L., Van Gossum, A., Orholm, M., Bonaiti-Pellie, C., Weissenbach, J., Mathew, C. G., Lennard-Jones, J. E., Cortot, A., Colombel, J.-F. and Thomas, G. (1996). "Mapping of a susceptibility locus for Crohn's disease on chromosome 16." Nature **379**(6568): 821-823.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. and Lee, C. (2004). "Detection of large-scale variation in the human genome." Nature Genetics **36**(9): 949-51.

Ingelman-Sundberg, M., Sim, S. C., Gomez, A. and Rodriguez-Antona, C. (2007). "Influence of cytochrome P450 polymorphisms on drug therapies: Pharmacogenetic, pharmacoepigenetic and clinical aspects." Pharmacology & Therapeutics **116**(3): 496-526.

Inohara, N., Ogura, Y. and Nuñez, G. (2002). "Nods: a family of cytosolic proteins that regulate the host response to pathogens." Current Opinion in Microbiology **5**(1): 76-80.

Jack, J. P. (1999). An Introduction to human molecular genetics: mechanism of inherited diseases, FITZGERALD SCIENCE PRESS.

Jacobs, A. P. and Strong, J. A. (1959). "A case of human intersexuality having a possible XXY sex-determining mechanism." Nature **183**(302-303).

Jacobs, P., Dalton, P., James, R., Mosse, K., Power, M., Robinson, D. and Skuse, D. (1997). "Turner syndrome: a cytogenetic and molecular study." Annals of Human Genetics **61**(6): 471-483.

Johansson, I., Lundqvist, E., Bertilsson, L., Dahl, M. L., SjÃ¶qvist, F. and Ingelman-Sundberg, M. (1993). "Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine." Proceedings of the National Academy of Sciences of the United States of America **90**(24): 11825-11829.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F. and Pinkel, D. (1992). "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors." Science **258**(5083): 818-821.

Khaja, R., Zhang, J. J., MacDonald, J. R., He, Y. S., Joseph-George, A. M., Wei, J., Rafiq, M. A., Qian, C., Shago, M., Pantano, L., Aburatani, H., Jones, K., Redon, R., Hurles, M., Armengol, L., Estivill, X., Mural, R. J., Lee, C., Scherer, S. W. and Feuk, L. (2006). "Genome assembly comparison identifies structural variants in the human genome." Nature Genetics **38**(12): 1413-1418.

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

Klotman, M. E. and Chang, T. L. (2006). "Defensins in innate antiviral immunity." Nature Reviews Immunology **6**(6): 447-456.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C. E., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M. and Snyder, M. (2007). "Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome." Science **318**(5849): 420-426.

Korenberg, J. R., Chen, X. N., Schipper, R., Sun, Z., Gonsky, R., Gerwehr, S., Carpenter, N., Daumer, C., Dignan, P. and Disteche, C. (1994). "Down syndrome phenotypes: the consequences of chromosomal imbalance." Proceedings of the National Academy of Sciences of the United States of America **91**(11): 4997-5001.

Lander, E. S. and Green, P. (1987). "Construction of multilocus genetic linkage maps in humans." Proceedings of the National Academy of Sciences of the United States of America **84**(8): 2363-2367.

Lee, J. A., Carvalho, C. M. B. and Lupski, J. R. (2007). "A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders." Cell **131**(7): 1235-1247.

Lejeune, J., Gautier, M. and Turpin, R. (1959). "Study of somatic chromosomes from 9 mongoloid children." Comptes Rendus Hebdomadaires des Seances de l Academie des Sciences () **248**: 1721-1722.

Libin, D., Yuezheng, Z., Jian, K., Tao, L., Hongbin, Z., Yang, G., Chaohua, L., Hao, P., Xiaoli, T., Dunmei, W., Tianhua, N., Huanming, Y. and Changqing, Z. (2008). "An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism." Human Mutation **29**(10): 1209-1216.

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**(1-2): 205-11.

Linzmeier, R. M. and Ganz, T. (2005). "Human defensin gene copy number polymorphisms: comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23." Genomics **86**(4): 423-30.

Liu, L., Zhao, C., Heng, H. H. and Ganz, T. (1997). "The human beta-defensin-1 and alpha-defensins are encoded by adjacent genes: two peptide families with differing disulfide topology share a common ancestry." Genomics **43**(3): 316-20.

Loftus, E. V. (2004). "Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences." Gastroenterology **126**(6): 1504-1517.

Lupski, J. (2009). "Genomic disorders ten years on." Genome Medicine **1**(4): 42.

Lupski, J. R. (1998). "Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits." Trends in Genetics **14**(10): 417-422.

Macdonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S. A., James, M., Groot, N., Macfarlane, H., Jenkins, B., Anderson, M. A., Wexler, N. S., Gusella, J. F., Bates, G. P., Baxendale, S., Hummerich, H., Kirby, S., North, M., Youngman, S., Mott, R., Zehetner, G., Sedlacek, Z., Poustka, A., Frischauf, A. M., Lehrach, H., Buckler, A. J., Church, D., Doucettestamm, L., Odonovan, M. C., Ribaramirez, L., Shah, M., Stanton, V. P., Strobel, S. A., Draths, K. M., Wales, J. L., Dervan, P., Housman, D. E., Altherr, M., Shiang, R., Thompson, L., Fielder, T., Wasmuth, J. J., Tagle, D., Valdes, J., Elmer, L., Allard, M., Castilla, L., Swaroop, M., Blanchard, K., Collins, F. S., Snell, R., Holloway, T., Gillespie, K., Datson, N., Shaw, D. and Harper, P. S. (1993). "A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntingtons-Disease Chromosomes." Cell **72**(6): 971-983.

Machuca-Tzili, L., Brook, D. and Hilton-Jones, D. (2005). "Clinical and molecular aspects of the myotonic dystrophies: A review." Muscle & Nerve **32**(1): 1-18.

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." Journal of Biological Chemistry **270**(51): 30371-6.

McCarroll, S. A. (2008). "Copy-number analysis goes more than skin deep." Nature Genetics **40**(1): 5-6.

McCarroll, S. A. and Altshuler, D. M. (2007). "Copy-number variation and association studies of human disease." Nature Genetics.

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., Dallaire, S., Gabriel, S. B., Lee, C., Daly, M. J. and Altshuler, D. M. (2006). "Common deletion polymorphisms in the human genome." Nature Genetics **38**: 86 - 92.

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

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 Research **16**(9): 1182-1190.

Nathans, J., Piantanida, T. P., Eddy, R. L., Shows, T. B. and Hogness, D. S. (1986). "Molecular genetics of inherited variation in human color vision." Science **232**: 203 - 210.

Nicholls, R. D. (1993). "Genomic imprinting and candidate genes in the Prader-Willi and Angelman syndromes." Current Opinion in Genetics & Development **3**(3): 445-456.

Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R. H., Achkar, J.-P., Brant, S. R., Bayless, T. M., Kirschner, B. S., Hanauer, S. B., Nunez, G. and Cho, J. H. (2001). "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease." Nature **411**(6837): 603-606.

Patau, K., Smith, D., Therman, E., Inhorn, S. and Wagner, H. (1960). "Multiple congenital anomaly caused by an extra autosome." The Lancet **275**(7128): 790-793.

Perne, A., XiangHong, Z., Lutz, E. L., Marco, G., Frank, S. and Malte, B. (2009). "Comparison of multiplex ligation-dependent probe amplification and real-time PCR accuracy for gene copy number quantification using the β-defensin locus." Biotechniques **47**(6): 1023-1028.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L. and Misra, R. (2007). "Diet and the evolution of human amylase gene copy number variation." Nature Genetics **39**(10): 1256 - 1260.

Peter, D. T. and Ellard, S. (2007). Emery's Element of medical genetics, Churchill Livingstone Elsevier.

Pieretti, M., Zhang, F., Fu, Y.-H., Warren, S. T., Oostra, B. A., Caskey, C. T. and Nelson, D. L. (1991). "Absence of expression of the FMR-1 gene in fragile X syndrome." Cell **66**(4): 817-822.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B.-m., Gray, J. W. and Albertson, D. G. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays." Nature Genetics **20**(2): 207-211.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R. and Chen, W. (2006). "Global variation in copy number in the human genome." Nature **444**(7118): 444 - 454.

Reiter, L. T., Hastings, P. J., Nelis, E., De Jonghe, P., Van Broeckhoven, C. and Lupski, J. R. (1998). "Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients." American Journal of Human Genetics **62**: 1023 - 1033.

Reiter, L. T., Murakami, T., Koeuth, T., Pentao, L., Muzny, D. M., Gibbs, R. A. and Lupski, J. R. (1996). "A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element." Nature Genetics **12**: 288 - 297.

Rodriguez-Jimenez, F. J., Krause, A., Schulz, S., Forssmann, W. G., Conejo-Garcia, J. R., Schreeb, R. and Motzkus, D. (2003). "Distribution of new human beta-defensin genes clustered on

chromosome 20 in functionally different segments of epididymis." Genomics **81**(2): 175-83.

Rubin, G. P., Hungin, A. P. S., Kelly, P. J. and Ling, J. (2000). "Inflammatory bowel disease: epidemiology and management in an English general practice population." Alimentary Pharmacology & Therapeutics **14**(12): 1553-1559.

Russell, R. K., Nimmo, E. R. and Satsangi, J. (2004). "Molecular genetics of Crohn's disease." Current Opinion in Genetics & Development **14**(3): 264-270.

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., Hurles, M. E. and Feuk, L. (2007). "Challenges and standards in integrating surveys of structural variation." Nature Genetics **39**(7 Suppl): S7 - 15.

Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F. and Pals, G. (2002). "Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification." Nucleic Acids Research **30**(12): e57.

Schrock, E., du Manoir, S., Veldman, T., Schoell, B., Wienberg, J., Ferguson-Smith, M. A., Ning, Y., Ledbetter, D. H., Bar-Am, I., Soenksen, D., Garini, Y. and Ried, T. (1996). "Multicolor Spectral Karyotyping of Human Chromosomes." Science **273**(5274): 494-497.

Schutte, B. C., Mitros, J. P., Bartlett, J. A., Walters, J. D., Jia, H. P., Welsh, M. J., Casavant, T. L. and McCray, P. B., Jr. (2002). "Discovery of five conserved beta -defensin gene clusters using a computational search strategy." Proceeding of the National Academy of Sciences of the United States of America **99**(4): 2129-33.

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

Sharp, A. J., Cheng, Z. and Eichler, E. E. (2006). "Structural variation of the human genome." Annual Review of Genomics and Human Genetics **7**: 407-442.

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D. and Eichler, E. E. (2005). "Segmental duplications and copy-number variation in the human genome." American Journal of Human Genetics **77**: 78 - 88.

Shaw, C. J. and Lupski, J. R. (2004). "Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease10.1093/hmg/ddh073." Human Molecular Genetics **13**(suppl_1): R57-64.

Small, K. and Warren, S. (1998). "Emerin deletions occurring on both Xq28 inversion backgrounds." Human Molecular Genetics **7**(1): 135-139.

Speicher, M. R. and Carter, N. P. (2005). "The new cytogenetics: blurring the boundaries with molecular biology." Nature Review Genetics **6**(10): 782-792.

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., Desnica, N., Hicks, A., Gylfason, A., Gudbjartsson, D. F., Jonsdottir, G. M., Sainz, J., Agnarsson, K., Birgisdottir, B., Ghosh, S., Olafsdottir, A., Cazier, J. B., Kristjansson, K., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., Kong, A. and Stefansson, K. (2005). "A common inversion under selection in Europeans." Nature Genetics **37**(2): 129-137.

Steven, R. B. and Yin Yao, S. (2004). "Inflammatory bowel disease gene hunting by linkage analysis. Rationale, methodology, and present status of the field." Inflammatory Bowel Diseases **10**(3): 300-311.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A. and Lee, C. (2007). "Relative impact of nucleotide and copy number variation on gene expression phenotypes." Science **315**(5813): 848 - 853.

Sugawara, H., Harada, N., Ida, T., Ishida, T., Ledbetter, D. H., Yoshiura, K., Ohta, T., Kishino, T., Niikawa, N. and Matsumoto, N. (2003). "Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23." Genomics **82**: 238 - 244.

Sun, C. Q., Arnold, R., Fernandez-Golarz, C., Parrish, A. B., Almekinder, T., He, J., Ho, S.-m., Svoboda, P., Pohl, J., Marshall, F. F. and Petros, J. A. (2006). "Human {beta}-Defensin-1, a Potential Chromosome 8p Tumor Suppressor: Control of Transcription and Induction of Apoptosis in Renal Cell Carcinoma." Cancer Research **66**(17): 8542-8549.

Sybert, V. P. and McCauley, E. (2004). "Turner's Syndrome." The New England Journal of Medicine **351**(12): 1227-1238.

Tang, Y. Q., Yuan, J., Osapay, G., Osapay, K., Tran, D., Miller, C. J., Ouellette, A. J. and Selsted, M. E. (1999). "A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated alpha-defensins." Science **286**(5439): 498-502.

Taudien, S., Galgoczy, P., Huse, K., Reichwald, K., Schilhabel, M., Szafranski, K., Shimizu, A., Asakawa, S., Frankish, A., Loncarevic, I., Shimizu, N., Siddiqui, R. and Platzer, M. (2004). "Polymorphic segmental duplications at 8p23.1 challenge the determination of individual defensin gene repertoires and the assembly of a contiguous human reference sequence." BMC Genomics **5**(1): 92.

Townson, J. R., Barcellos, L. F. and Nibbs, R. J. B. (2002). "Gene copy number regulates the production of the human chemokine CCL3-L1." European Journal of Immunology **32**(10): 3016-3026.

Turner, D. J., Miretti, M., Rajan, D., Fiegler, H., Carter, N. P., Blayney, M. L., Beck, S. and Hurles, M. E. (2008). "Germline rates of de novo meiotic deletions and duplications causing several genomic disorders." Nature Genetics **40**(1): 90-95.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D. and Pinkel, D. (2005). "Fine-scale structural variation of the human genome." Nature Genetics **37**(7): 727 - 732.

Van Limbergen, J., Wilson, D. C. and Satsangi, J. (2009). "The Genetics of Crohn's Disease." Annual Review of Genomics and Human Genetics **10**(1): 89-116.

Vollrath, D., Nathans, J. and Davis, R. W. (1988). "Tandem array of human visual pigment genes at Xq28." Science **240**(4859): 1669-1672.

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

Wain, L. V., Armour, J. A. L. and Tobin, M. D. (2009). "Genomic copy number variation, human health, and disease." The Lancet **374**(9686): 340-350.

Walker, S., Janyakhantikul, S. and Armour, J. A. L. (2009). "Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs." Genomics **93**(1): 98-103.

Wehkamp, J., Schmid, M., Fellermann, K. and Stange, E. F. (2005). "Defensin deficiency, intestinal microbes, and the clinical phenotypes of Crohn's disease." Journal of Leukocyte Biology **77**(4): 460-465.

Yamaguchi, Y., Nagase, T., Makita, R., Fukuhara, S., Tomita, T., Tominaga, T., Kurihara, H. and Ouchi, Y. (2002). "Identification of Multiple Novel Epididymis-Specific {beta}-Defensin Isoforms in Humans and Mice." Journal of Immunology **169**(5): 2516-2523.

Yamazaki, K., McGovern, D., Ragoussis, J., Paolucci, M., Butler, H., Jewell, D., Cardon, L., Takazoe, M., Tanaka, T., Ichimori, T., Saito, S., Sekine, A., Iida, A., Takahashi, A., Tsunoda, T., Lathrop, M. and Nakamura, Y. (2005). "Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease." Human Molecular Genetics **14**(22): 3499-3506.

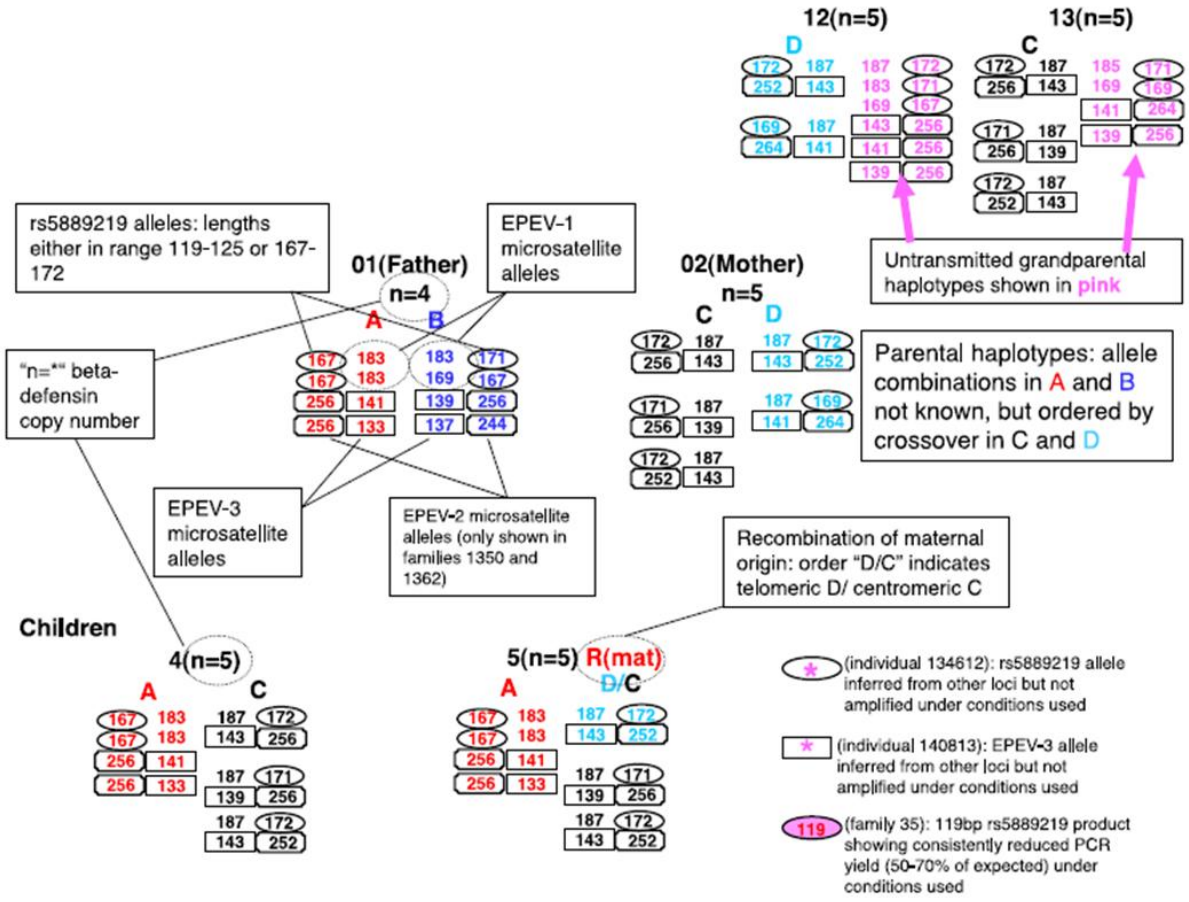
Yang, D., Chertov, O. and Oppenheim, J. J. (2001). "The role of mammalian antimicrobial peptides and proteins in awakening of innate host defenses and adaptive immunity." Cellular and Molecular Life Sciences **58**(7): 978-89.

Yang, Y., Chung, E. K., Wu, Y. L., Savelli, S. L., Nagaraja, H. N., Zhou, B., Hebert, M., Jones, K. N., Shu, Y. and Kitzmiller, K. (2007). "Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans." American Journal of Human Genetics **80**(6): 1037 - 1054.

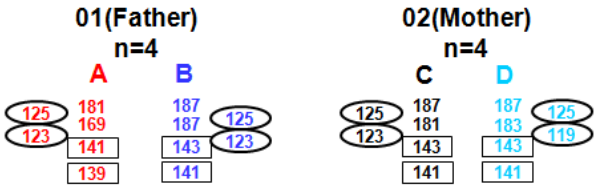
Yunis, J. J. and Sanchez, O. (1973). "G-banding and chromosome structure." Chromosoma **44**(1): 15-23.

APPENDICES

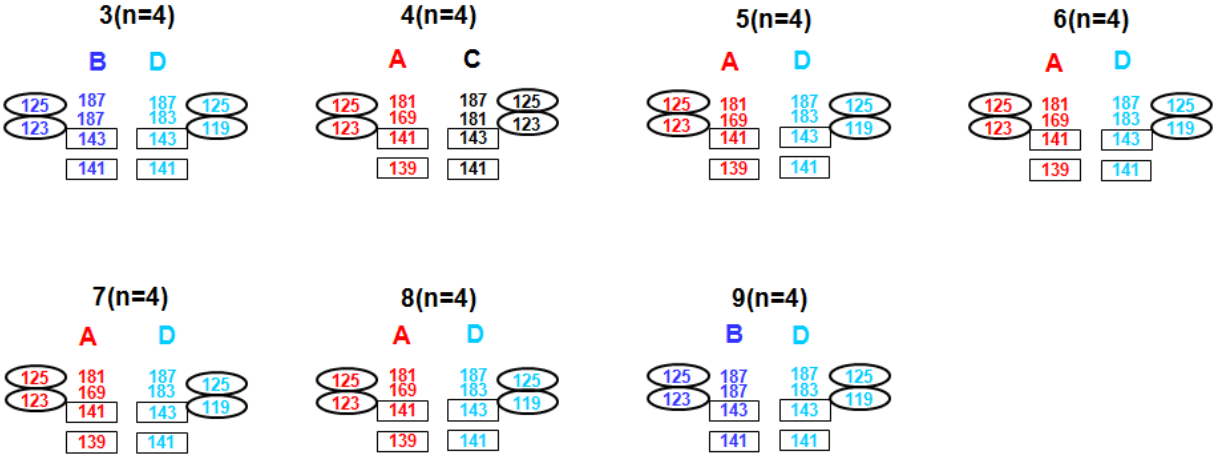
Page 184 show the explanatory key to the symbols used on the other page.
Page 185 – 210 show the segregation of all markers typed in 26 CEPH pedigrees.

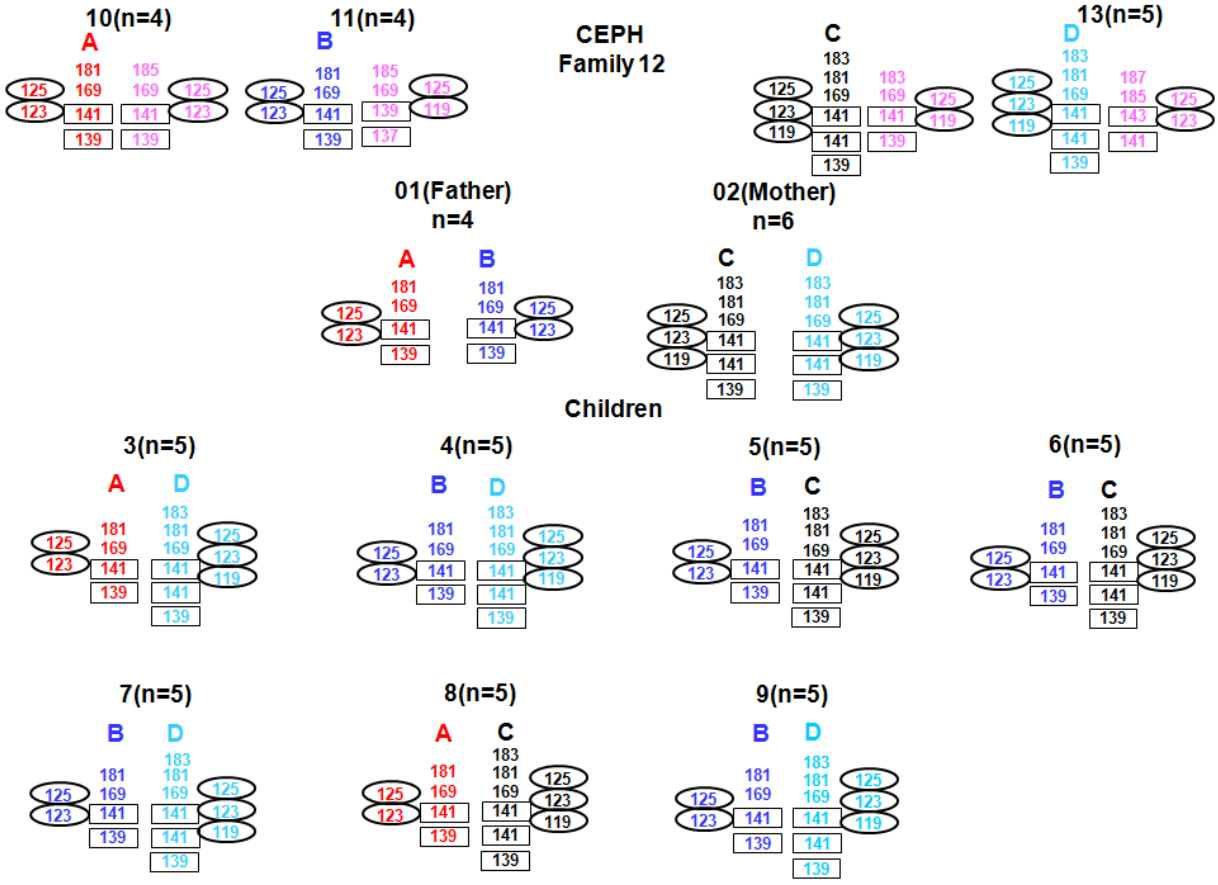


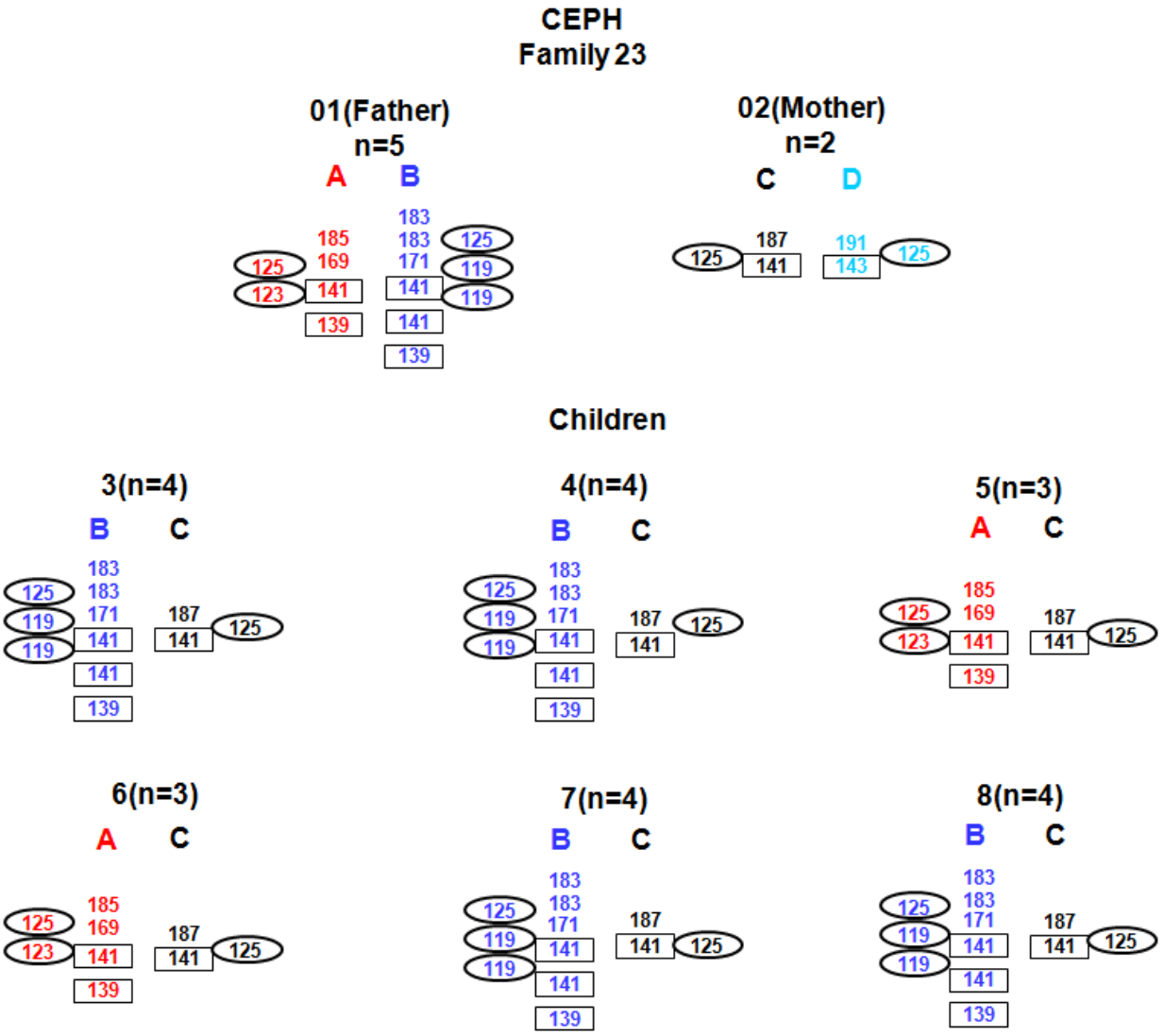
CEPH
Family 2

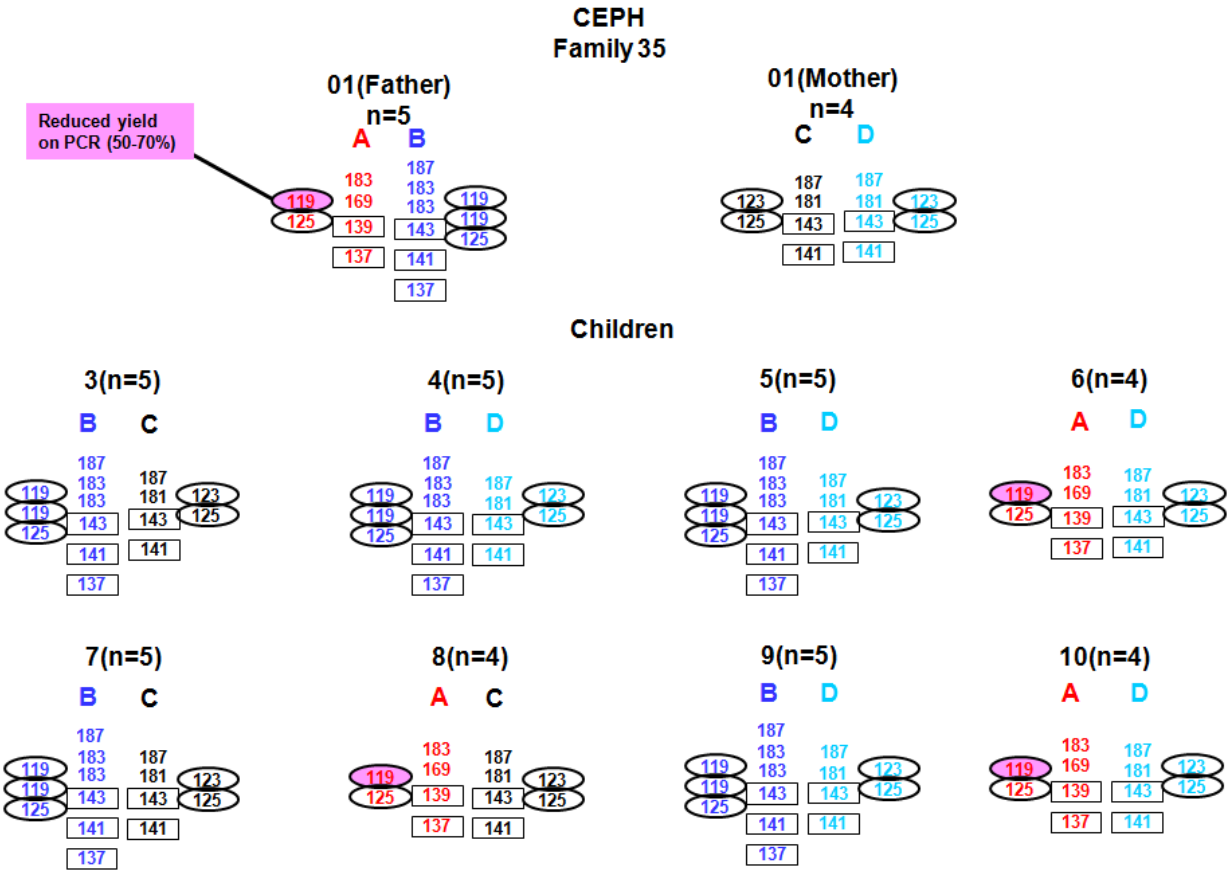


Children

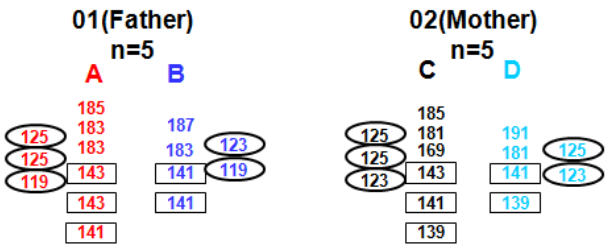



10(n=4)
A
181 185
169 169
125 141 141 125
123 141 123
139 139
11(n=4)
B
181 185
169 169
125 141 139 125
123 141 139 119
139 137
CEPH
Family 12
C
183
181
169
125 141 183 169 125
123 141 141 141 119
119 141 139
139
13(n=5)
D
183
181
125 169 187
123 141 185 125
119 141 143 123
141 141
139
01(Father)
n=4
A B
181 181
169 169 125
125 141 141 123
123 139 139
02(Mother)
n=6
C D
183 183
181 181
169 169
125 125
123 141 141 123
119 141 141 119
139 139
Children
3(n=5)
A D
181 183
169 181
125 169 125
123 141 141 123
139 141 119
139
4(n=5)
B D
181 183
169 181
125 169 125
123 141 141 123
139 141 119
139
5(n=5)
B C
181 183
169 181
125 169 125
123 141 141 123
139 141 119
139
6(n=5)
B C
181 183
169 181
125 169 125
123 141 141 123
139 141 119
139
7(n=5)
B D
181 183
169 181
125 169 125
123 141 141 123
139 141 119
139
8(n=5)
A C
183
181
169
125 125
123 141 123
139 141 119
139
9(n=5)
B D
181 183
169 181
125 169 125
123 141 141 123
139 141 119
139

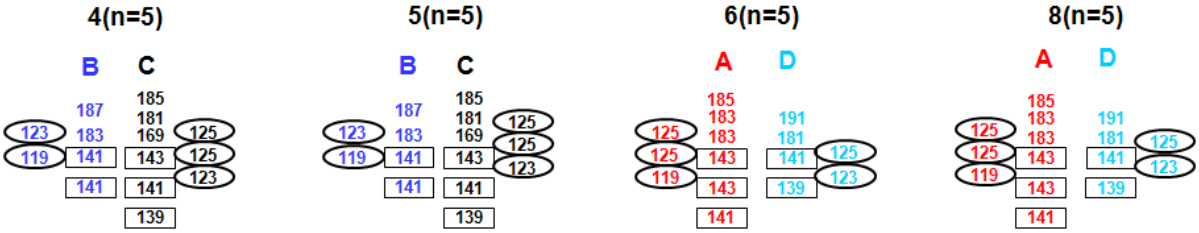

CEPH
Family 23
01(Father)
n=5
A
B
185
169
125
123
141
139
183
183
171
125
119
141
119
141
139
02(Mother)
n=2
C
D
187
141
191
143
125
125
Children
3(n=4)
B
C
183
183
171
125
119
141
119
141
139
187
141
125
4(n=4)
B
C
183
183
171
125
119
141
119
141
139
187
141
125
5(n=3)
A
C
185
169
125
123
141
139
187
141
125
6(n=3)
A
C
185
169
125
123
141
139
187
141
125
7(n=4)
B
C
183
183
171
125
119
141
119
141
139
187
141
125
8(n=4)
B
C
183
183
171
125
119
141
119
141
139
187
141
125

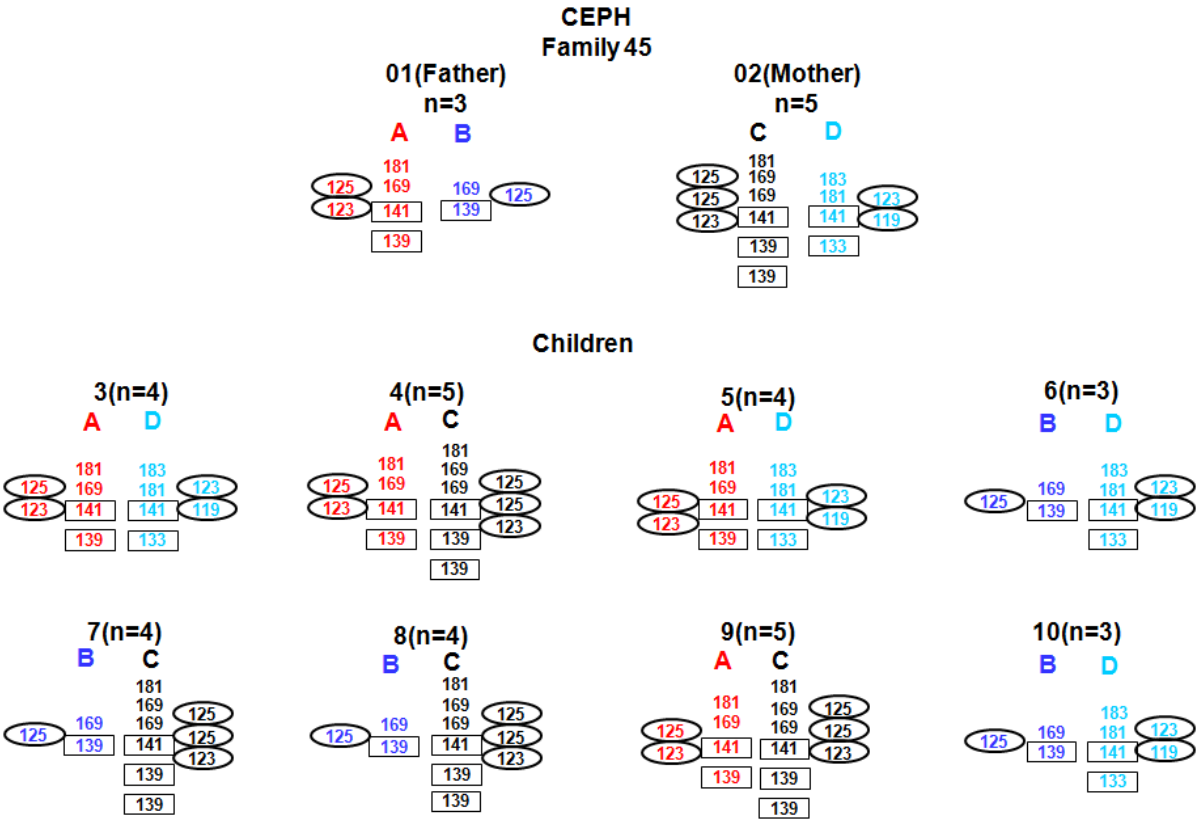

CEPH
Family 35
01(Father)
n=5
A B
Reduced yield
on PCR (50-70%)
01(Mother)
n=4
C D
Children
3(n=5)
B C
4(n=5)
B D
5(n=5)
B D
6(n=4)
A D
7(n=5)
B C
8(n=4)
A C
9(n=5)
B D
10(n=4)
A D

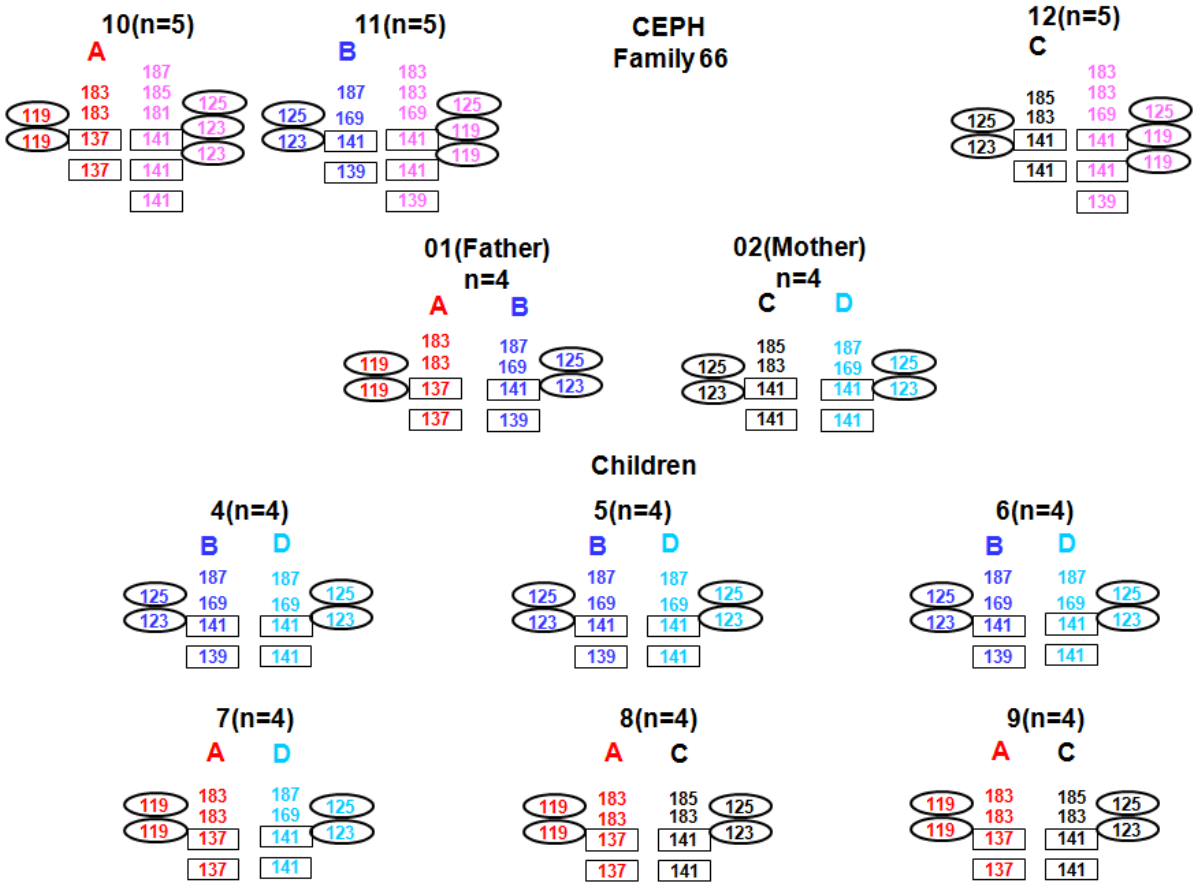
CEPH
Family 37

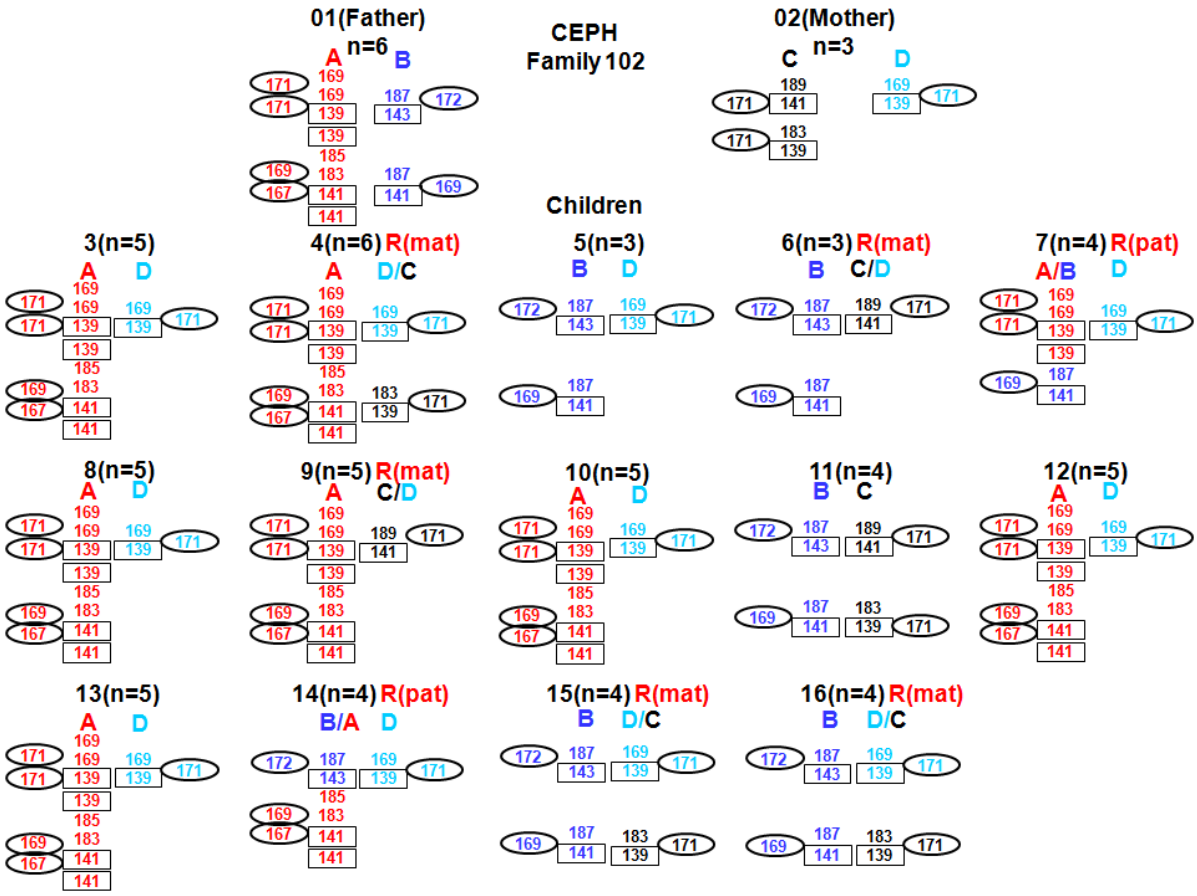


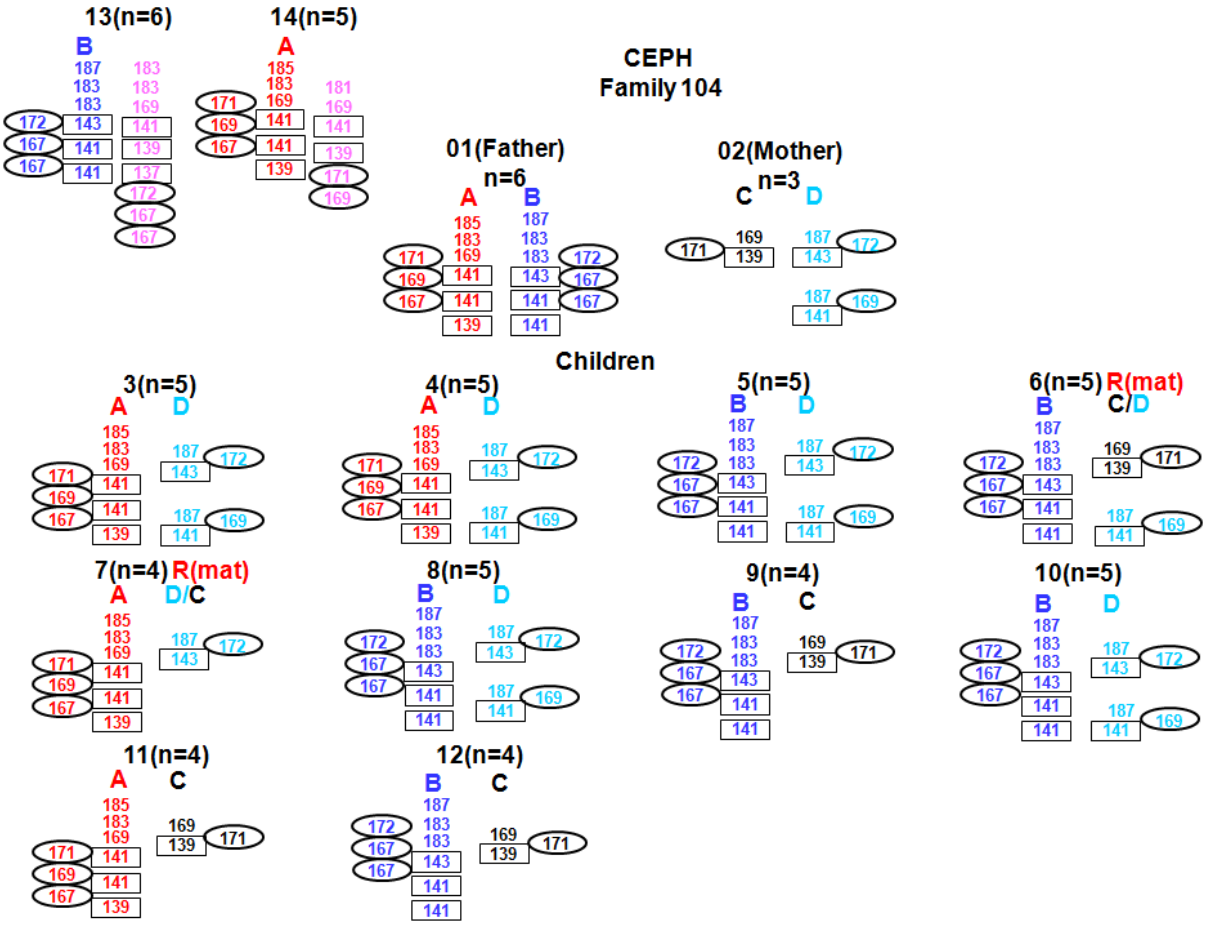
Children

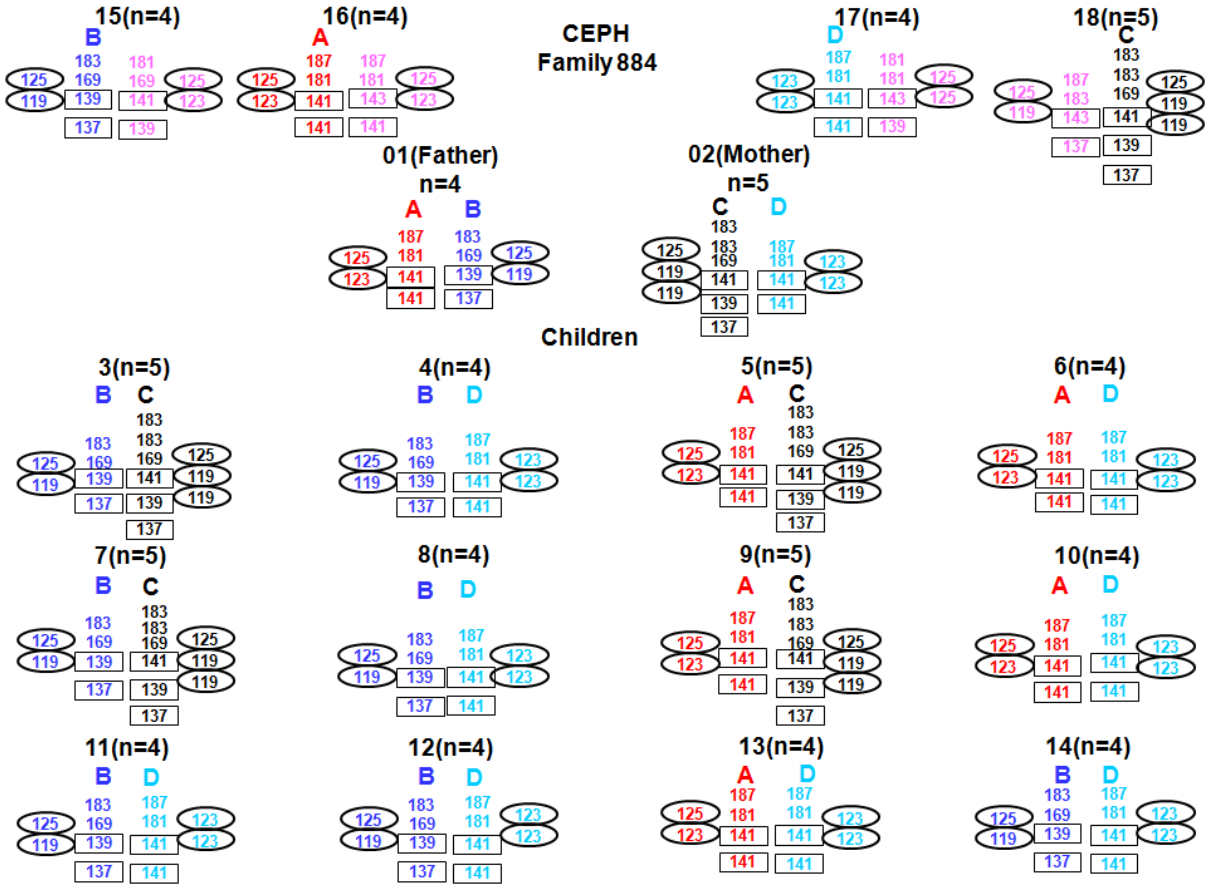


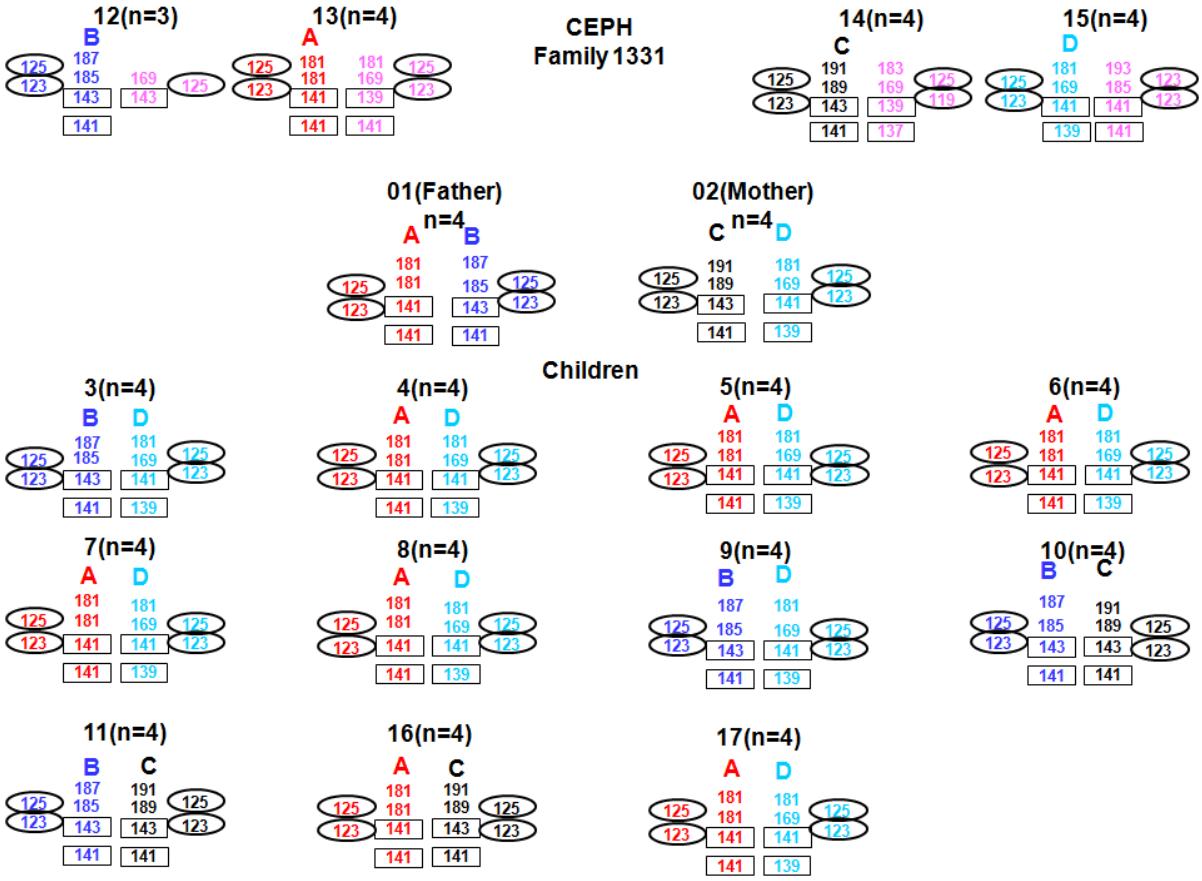

CEPH
Family 45
01(Father)
n=3
A
B
181
169
125
123
141
139
169
125
139
02(Mother)
n=5
C
D
181
169
125
125
169
123
141
139
139
183
181
123
141
119
133
Children
3(n=4)
A
D
181
169
125
123
141
139
183
181
123
141
119
133
4(n=5)
A
C
181
169
125
123
141
139
181
169
169
125
141
125
139
123
139
5(n=4)
A
D
181
169
125
123
141
139
183
181
123
141
119
133
6(n=3)
B
D
125
169
139
183
181
123
141
119
133
7(n=4)
B
C
125
169
139
181
169
169
141
139
139
125
125
123
8(n=4)
B
C
125
169
139
181
169
169
141
139
139
125
125
123
9(n=5)
A
C
181
169
125
123
141
139
181
169
169
141
139
139
125
125
123
10(n=3)
B
D
125
169
139
183
181
123
141
119
133

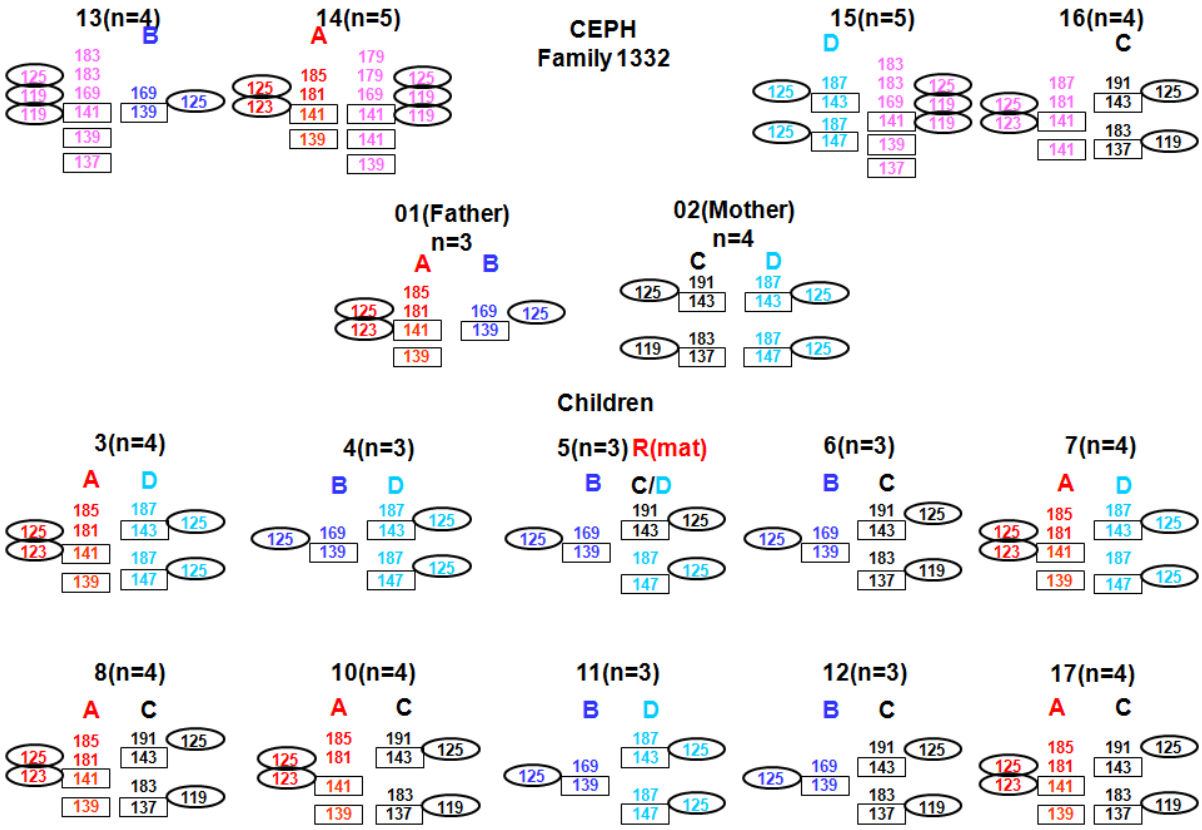

CEPH
Family 66
10(n=5)
A
119 119
183 183
137
187 185 181
141 141
125 123 123
141
11(n=5)
B
125 123
187 169
141
183 183 169
141 141
125 119 119
139 141
139
12(n=5)
C
125 123
185 183
141
183 183 169
141 141
125 119 119
139
01(Father)
n=4
A B
119 119
183 183
137
137
187 169
141
125 123
139
02(Mother)
n=4
C D
125 123
185 183
141
141
187 169
141 141
125 123
Children
4(n=4)
B D
125 123
187 169
141
139
187 169
141 141
125 123
141
5(n=4)
B D
125 123
187 169
141
139
187 169
141 141
125 123
141
6(n=4)
B D
125 123
187 169
141
139
187 169
141 141
125 123
141
7(n=4)
A D
119 119
183 183
137
137
187 169
141 141
125 123
141
8(n=4)
A C
119 119
183 183
137
137
185 183
141
125 123
141
9(n=4)
A C
119 119
183 183
137
137
185 183
141 141
125 123
141

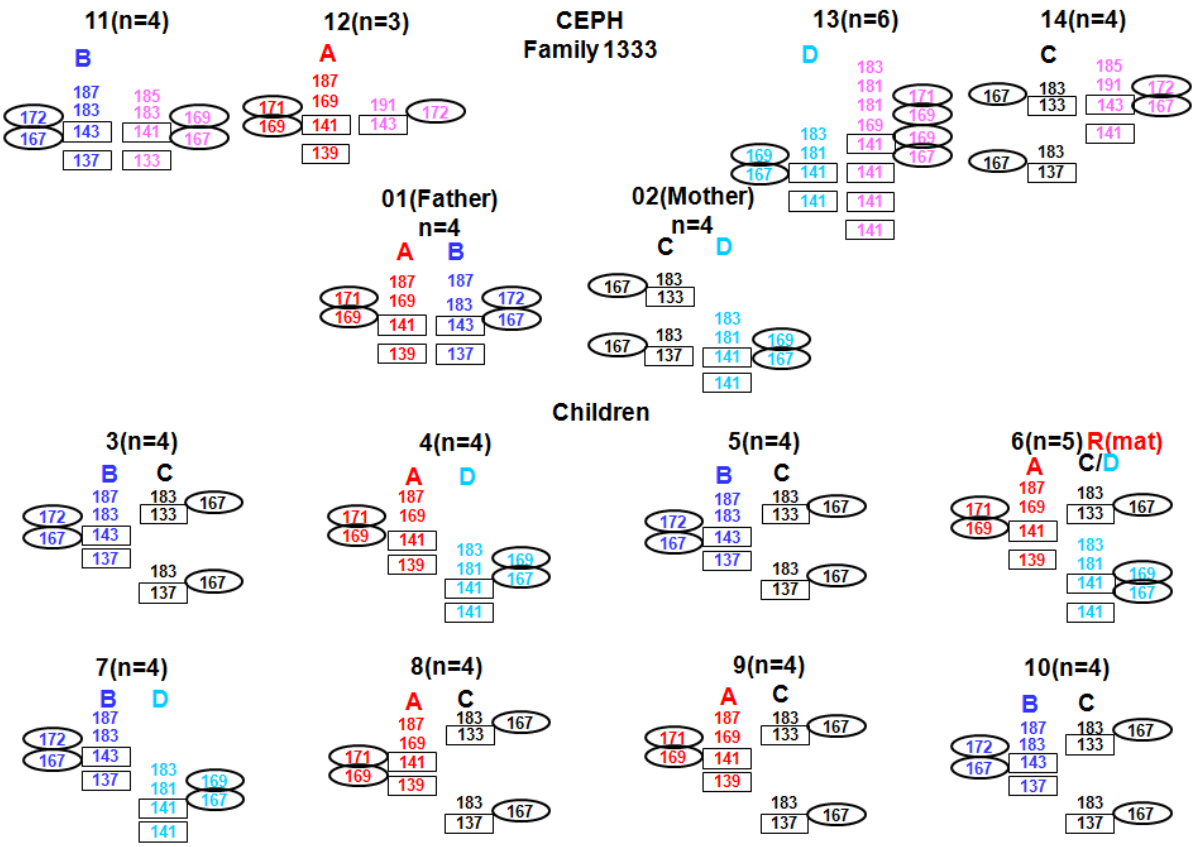

01(Father)
n=6
A
B
CEPH
Family 102
02(Mother)
n=3
C
D
Children
3(n=5)
A
D
4(n=6) R(mat)
A
D/C
5(n=3)
B
D
6(n=3) R(mat)
B
C/D
7(n=4) R(pat)
A/B
D
8(n=5)
A
D
9(n=5) R(mat)
A
C/D
10(n=5)
A
D
11(n=4)
B
C
12(n=5)
A
D
13(n=5)
A
D
14(n=4) R(pat)
B/A
D
15(n=4) R(mat)
B
D/C
16(n=4) R(mat)
B
D/C

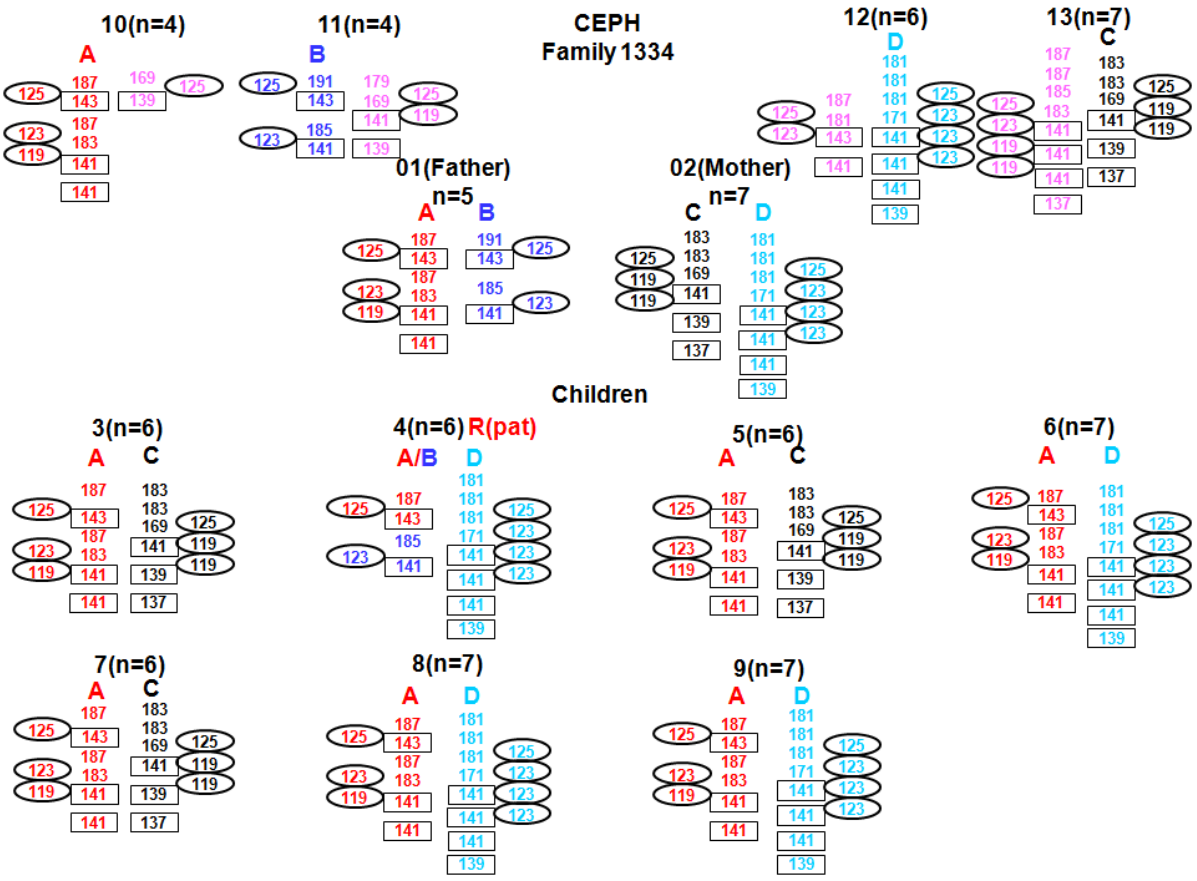

13(n=6)
B
187 183 183
172 167 167
143 141 141
183 183 169
141 139 137
172 167 167
14(n=5)
A
185 183 169
171 169 167
141 141 141
181 169 141 139 171 169
CEPH
Family 104
01(Father)
n=6
A
185 183 169
171 169 167
141 141 139
B
187 183 183
172 167 167
143 141 141
02(Mother)
n=3
C
171
169 139
D
187 172
143
187 169
141
Children
3(n=5)
A
185 183 169
171 169 167
141 141 139
D
187 172
143
187 169
141
4(n=5)
A
185 183 169
171 169 167
141 141 139
D
187 172
143
187 169
141
5(n=5)
B
187 183 183
172 167 167
143 141 141
D
187 172
143
187 169
141
6(n=5) R(mat)
B
187 183
172 167
143 141 141
C/D
169 171
139
187 169
141
7(n=4) R(mat)
A
185 183 169
171 169 167
141 141 139
D/C
187 172
143
8(n=5)
B
187 183 183
172 167 167
143 141 141
D
187 172
143
187 169
141
9(n=4)
B
187 183 183
172 167 167
143 141 141
C
169 171
139
10(n=5)
B
187 183 183
172 167 167
143 141 141
D
187 172
143
187 169
141
11(n=4)
A
185 183 169
171 169 167
141 141 139
C
169 171
139
12(n=4)
B
187 183 183
172 167 167
143 141 141
C
169 171
139

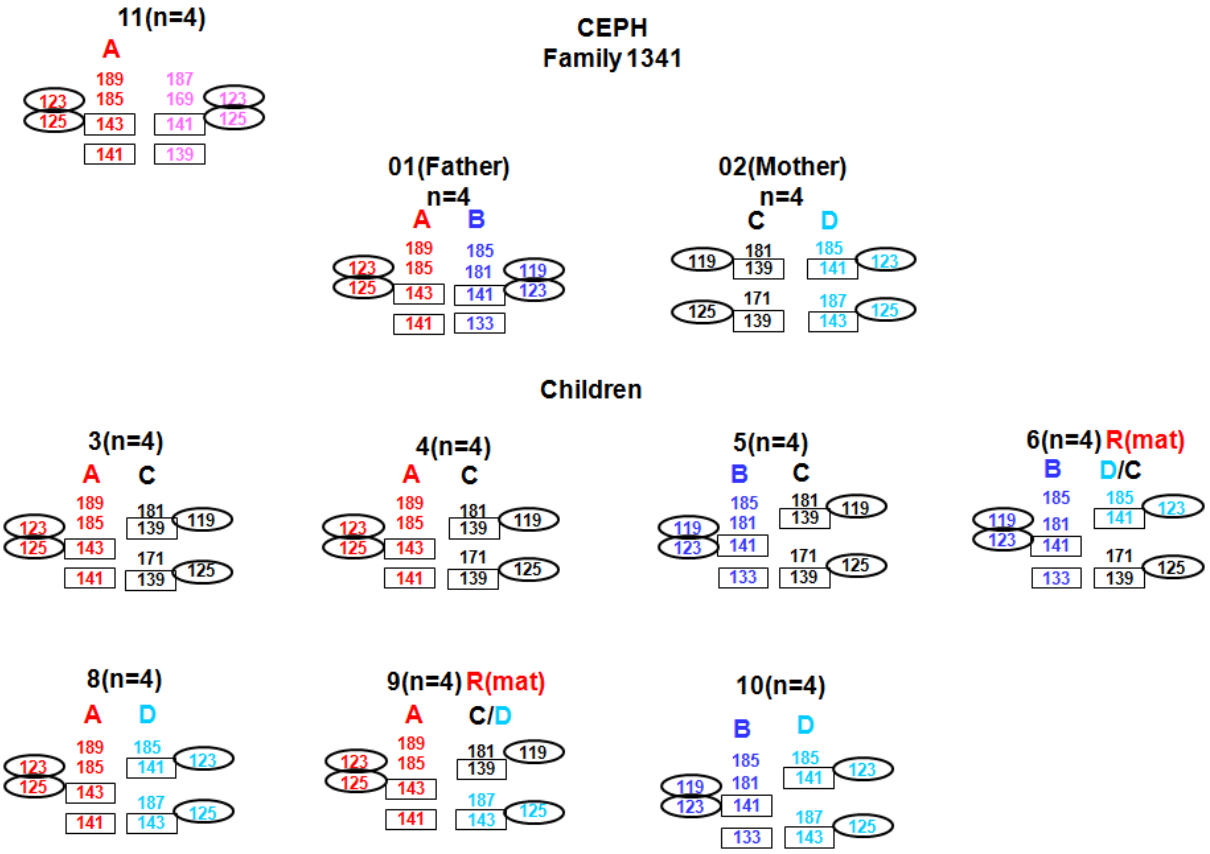

15(n=4)
B
16(n=4)
A
CEPH
Family 884
17(n=4)
D
18(n=5)
C
01(Father)
n=4
A B
02(Mother)
n=5
C D
Children
3(n=5)
B C
4(n=4)
B D
5(n=5)
A C
6(n=4)
A D
7(n=5)
B C
8(n=4)
B D
9(n=5)
A C
10(n=4)
A D
11(n=4)
B D
12(n=4)
B D
13(n=4)
A D
14(n=4)
B D

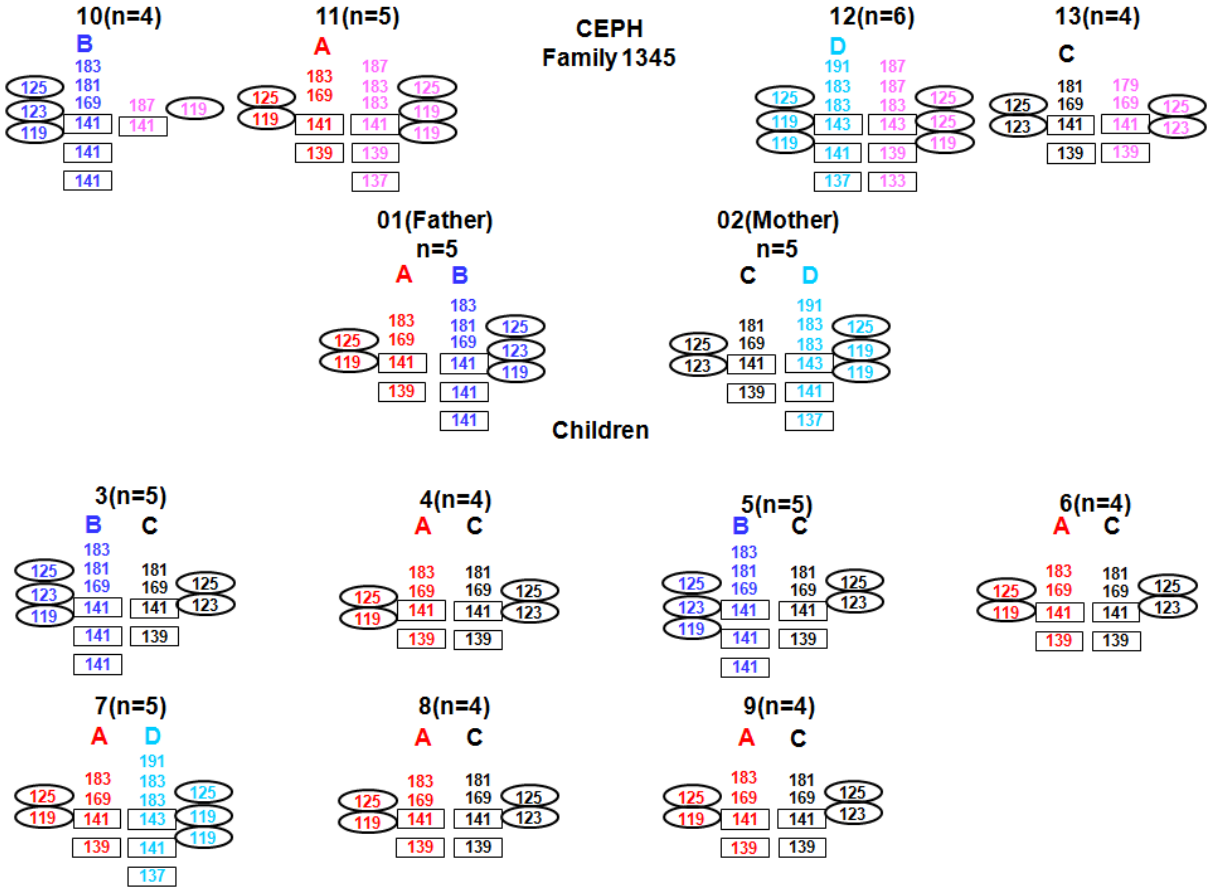

12(n=3)
B
125 123
187 185
143
169 143
125
141
13(n=4)
A
125 123
181 181
141
181 169
139
125 123
141 141
CEPH
Family 1331
14(n=4)
C
125 123
191 189
143
183 169
139
125 119
141 137
15(n=4)
D
125 123
181 169
141
193 185
141
123 123
139 141
01(Father)
n=4
A B
125 123
181 181
141
187 185
143
125 123
141 141
02(Mother)
n=4
C D
125 123
191 189
143
181 169
141
125 123
141 139
Children
3(n=4)
B D
125 123
187 185
143
181 169
141
125 123
141 139
4(n=4)
A D
125 123
181 181
141
181 169
141
125 123
141 139
5(n=4)
A D
125 123
181 181
141
181 169
141
125 123
141 139
6(n=4)
A D
125 123
181 181
141
181 169
141
125 123
141 139
7(n=4)
A D
125 123
181 181
141
181 169
141
125 123
141 139
8(n=4)
A D
125 123
181 181
141
181 169
141
125 123
141 139
9(n=4)
B D
125 123
187 185
143
181 169
141
125 123
141 139
10(n=4)
B C
125 123
187 185
143
191 189
143
125 123
141 141
11(n=4)
B C
125 123
187 185
143
191 189
143
125 123
141 141
16(n=4)
A C
125 123
181 181
141
191 189
143
125 123
141 141
17(n=4)
A D
125 123
181 181
141
181 169
141
125 123
141 139

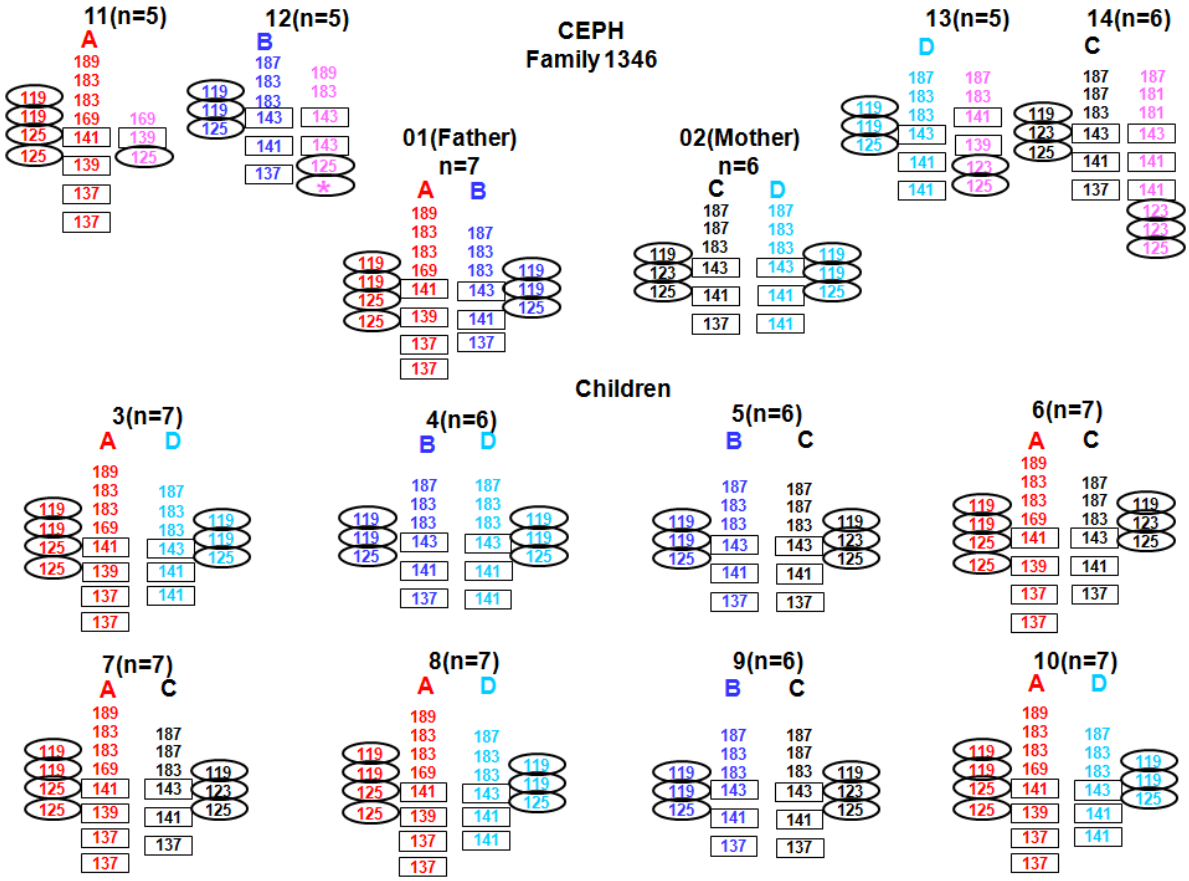

13(n=4)
B
183
183
169
141
139
137
125
119
119
169
139
125
14(n=5)
A
185
181
141
139
125
123
179
179
169
141
141
139
125
119
119
CEPH
Family 1332
15(n=5)
D
183
183
169
141
139
137
125
119
119
125
187
143
125
187
147
16(n=4)
C
191
143
125
187
181
141
125
123
183
137
119
141
01(Father)
n=3
A
B
185
181
141
139
125
123
169
139
125
02(Mother)
n=4
C
D
191
143
125
187
143
125
183
137
119
187
147
125
Children
3(n=4)
A
D
185
181
141
139
125
123
187
143
125
187
147
125
4(n=3)
B
D
169
139
125
187
143
125
187
147
125
5(n=3)
R(mat)
B
C/D
169
139
125
191
143
125
187
147
125
6(n=3)
B
C
169
139
125
191
143
125
183
137
119
7(n=4)
A
D
185
181
141
139
125
123
187
143
125
187
147
125
8(n=4)
A
C
185
181
141
139
125
123
191
143
125
183
137
119
10(n=4)
A
C
185
181
141
139
125
123
191
143
125
183
137
119
11(n=3)
B
D
169
139
125
187
143
125
187
147
125
12(n=3)
B
C
169
139
125
191
143
125
183
137
119
17(n=4)
A
C
185
181
141
139
125
123
191
143
125
183
137
119

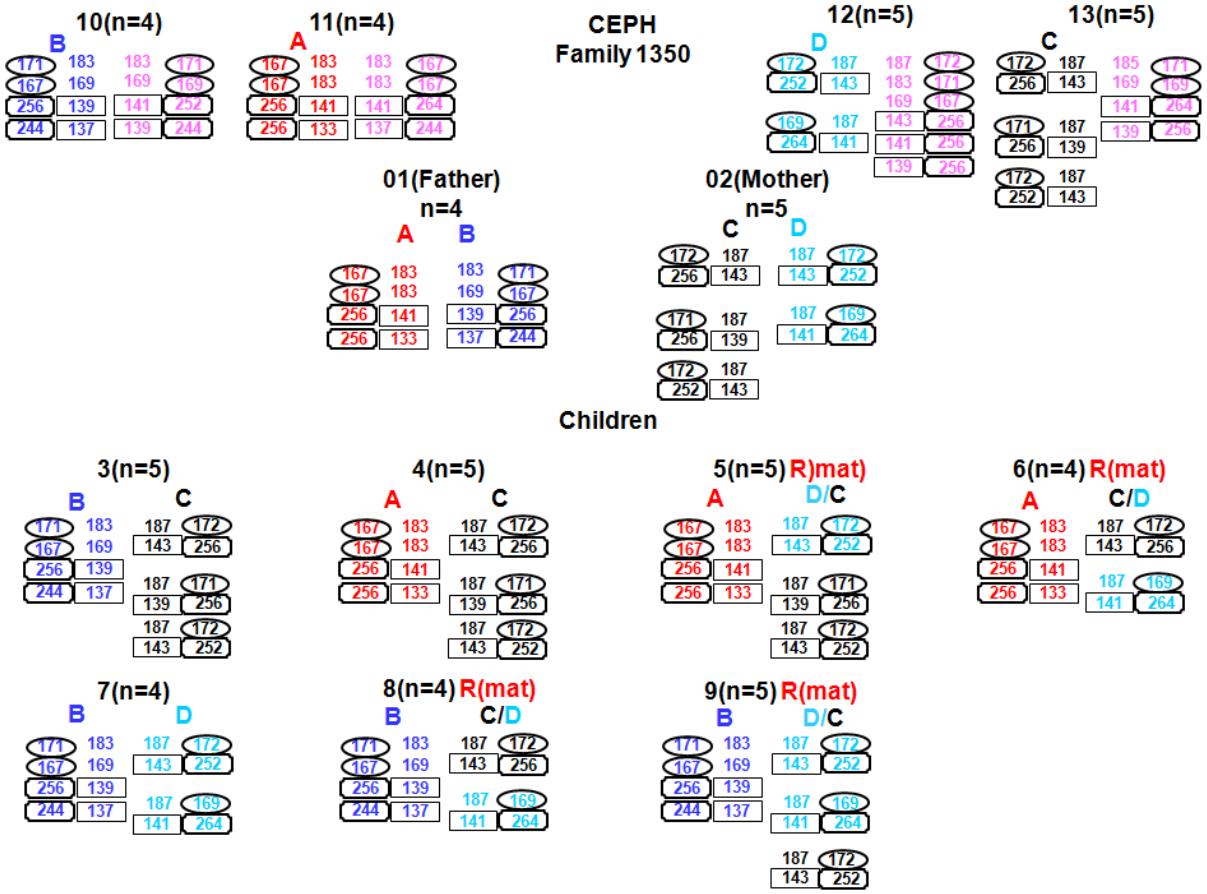

11(n=4)
B
187
183
172
167
143
137
185
183
141
133
169
167
12(n=3)
A
187
169
171
169
141
139
191
143
172
CEPH
Family 1333
13(n=6)
D
183
181
181
171
169
169
141
169
183
181
169
167
141
141
141
141
141
14(n=4)
C
167
183
133
185
191
143
172
167
141
167
183
137
01(Father)
n=4
A
B
187
169
171
169
141
139
187
183
143
137
172
167
02(Mother)
n=4
C
D
167
183
133
167
183
137
183
181
141
141
169
167
Children
3(n=4)
B
C
187
183
172
167
143
137
183
133
167
183
137
167
4(n=4)
A
D
187
169
171
169
141
139
183
181
141
141
169
167
5(n=4)
B
C
187
183
172
167
143
137
183
133
167
183
137
167
6(n=5)
R(mat)
A
C/D
187
169
171
169
141
139
183
133
167
183
181
141
141
169
167
7(n=4)
B
D
187
183
172
167
143
137
183
181
141
141
169
167
8(n=4)
A
C
187
169
171
169
141
139
183
133
167
183
137
167
9(n=4)
A
C
187
169
171
169
141
139
183
133
167
183
137
167
10(n=4)
B
C
187
183
172
167
143
137
183
133
167
183
137
167

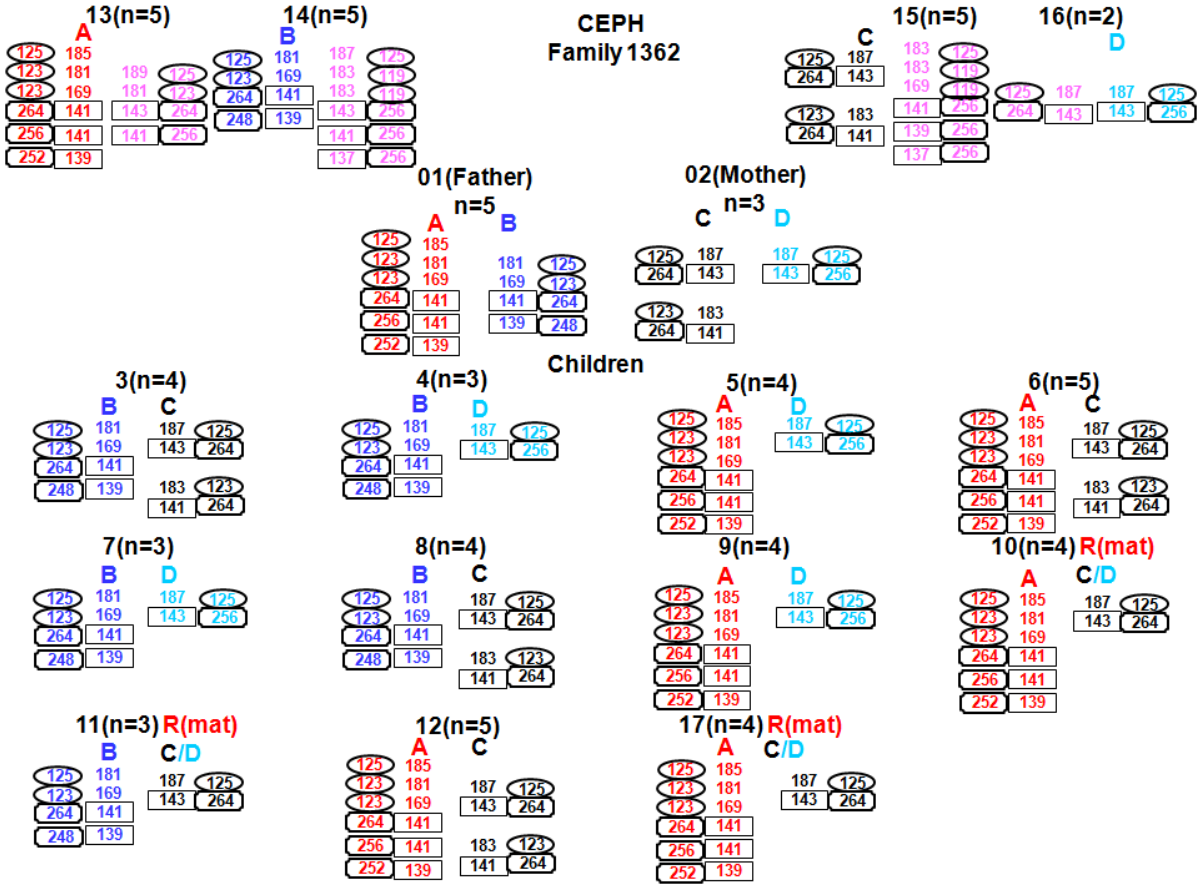

CEPH
Family 1334
10(n=4)
A
11(n=4)
B
12(n=6)
D
13(n=7)
C
01(Father)
n=5
A
B
02(Mother)
n=7
C
D
Children
3(n=6)
A
C
4(n=6)
A/B
D
R(pat)
5(n=6)
A
C
6(n=7)
A
D
7(n=6)
A
C
8(n=7)
A
D
9(n=7)
A
D

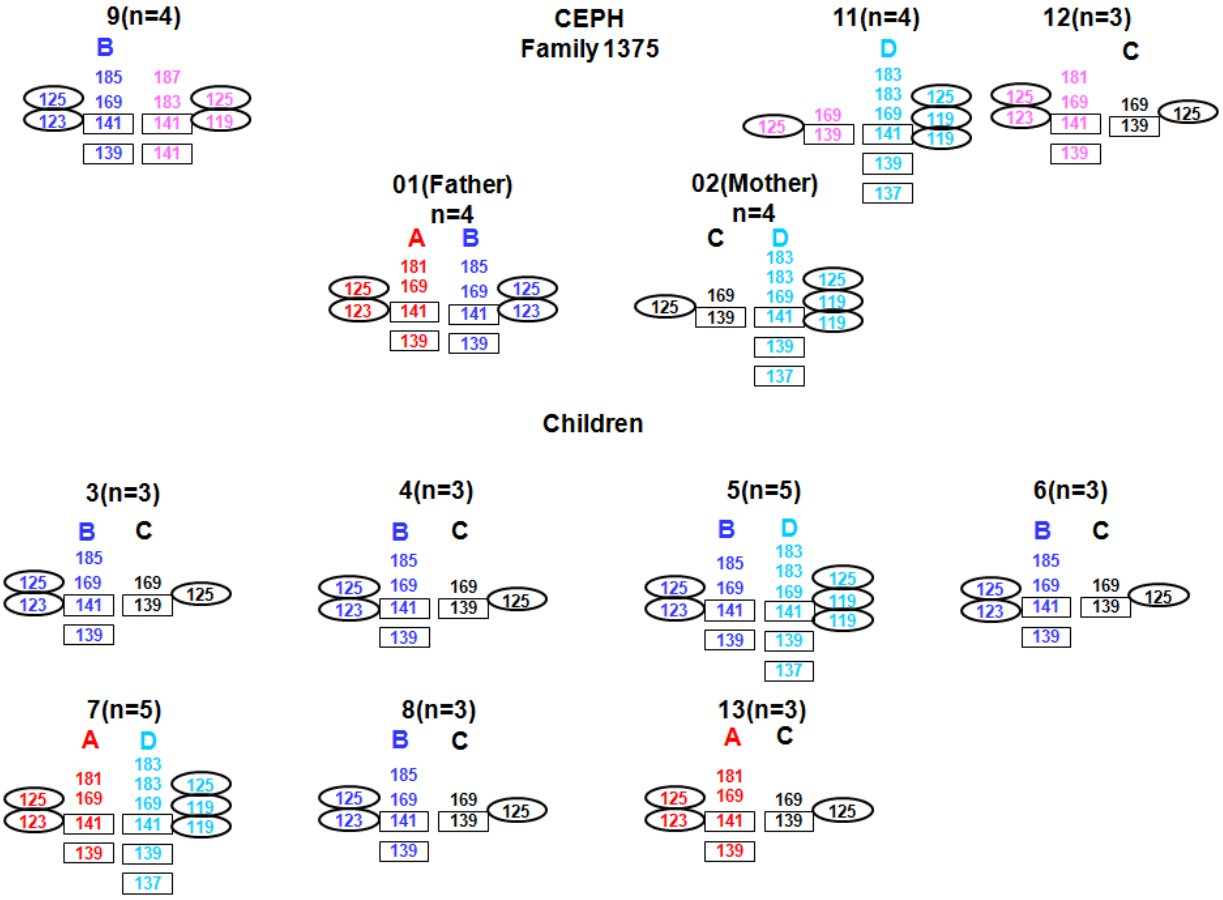

11(n=4)
A
189 187
185 169
123 123
125 143 141 125
141 139
CEPH
Family 1341
01(Father)
n=4
A B
189 185
185 181
123 119
125 143 141 123
141 133
02(Mother)
n=4
C D
181 185
119 139 141 123
171 187
125 139 143 125
Children
3(n=4)
A C
189 181
185 139 119
123
125 143 171
141 139 125
4(n=4)
A C
189 181
185 139 119
123
125 143 171
141 139 125
5(n=4)
B C
185 181
181 139 119
119
123 141 171
133 139 125
6(n=4) R(mat)
B D/C
185 185
181 141 123
119
123 141 171
133 139 125
8(n=4)
A D
189 185
185 141 123
123
125 143 187
141 143 125
9(n=4) R(mat)
A C/D
189 181
185 139 119
123
125 143 187
141 143 125
10(n=4)
B D
185 185
181 141 123
119
123 141 187
133 143 125

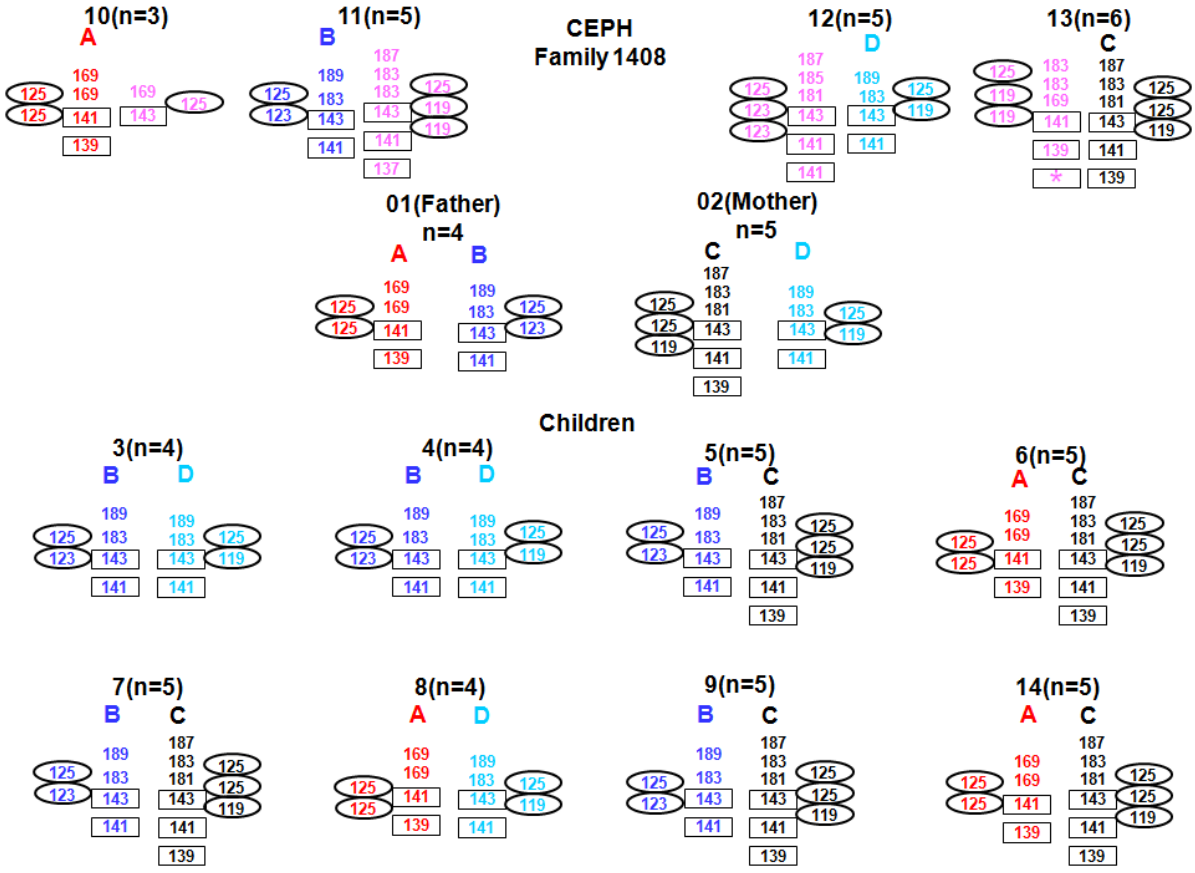

CEPH
Family 1345
10(n=4)
B
183
181
169
125
123
119
141
187
141
119
141
141
11(n=5)
A
183
169
187
183
183
125
125
119
119
141
141
139
139
137
12(n=6)
D
191
183
187
183
187
183
125
125
119
143
143
125
119
141
139
119
137
133
13(n=4)
C
181
169
179
169
125
125
123
141
141
123
139
139
01(Father)
n=5
A
B
183
169
183
181
169
125
125
123
119
141
141
139
141
141
02(Mother)
n=5
C
D
191
181
183
169
125
125
123
119
141
143
119
139
141
137
Children
3(n=5)
B
C
183
181
169
181
169
125
125
123
123
119
141
141
141
139
141
4(n=4)
A
C
183
169
181
169
125
125
119
141
141
123
139
139
5(n=5)
B
C
183
181
169
181
169
125
125
123
123
119
141
141
141
139
141
6(n=4)
A
C
183
169
181
169
125
125
119
141
141
123
139
139
7(n=5)
A
D
191
183
169
183
125
125
119
141
143
119
139
141
137
8(n=4)
A
C
183
169
181
169
125
125
119
141
141
123
139
139
9(n=4)
A
C
183
169
181
169
125
125
119
141
141
123
139
139

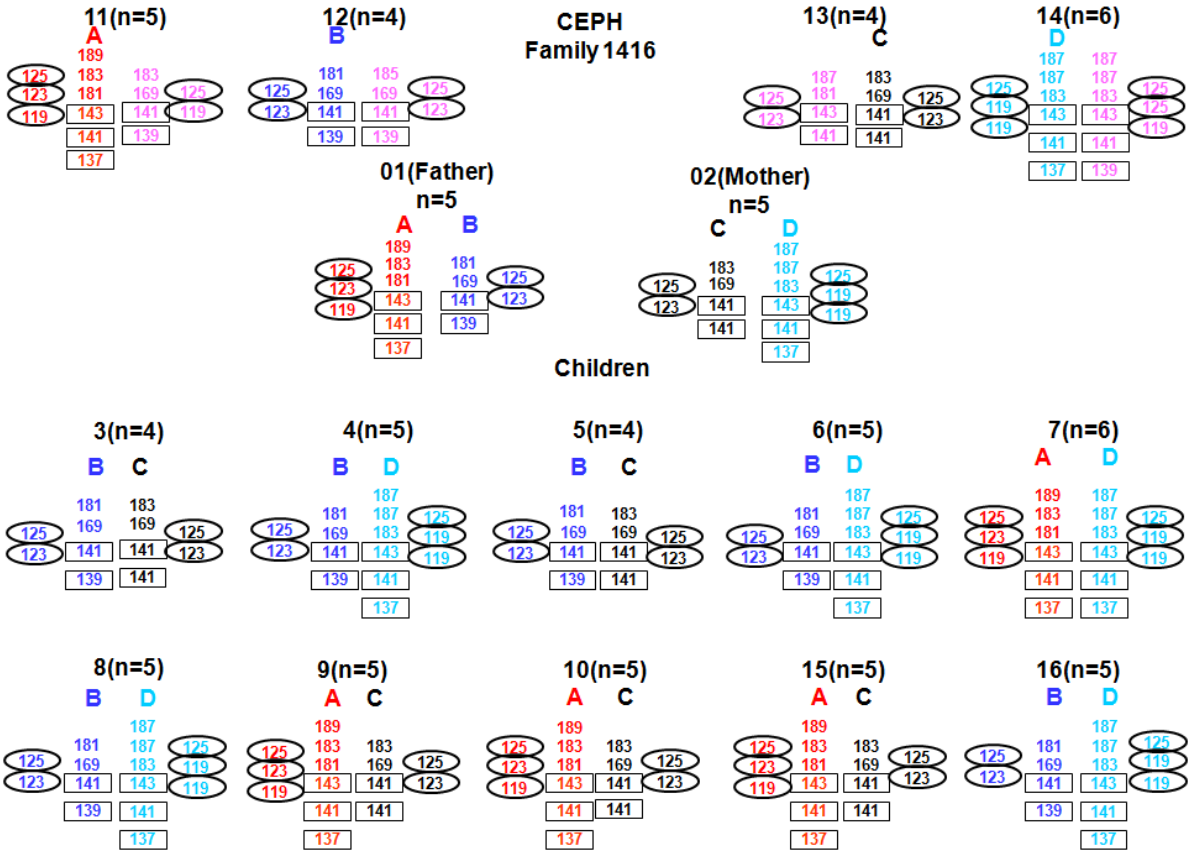

CEPH
Family 1346
11(n=5)
A
12(n=5)
B
01(Father)
n=7
A B
02(Mother)
n=6
C D
13(n=5)
D
14(n=6)
C
Children
3(n=7)
A D
4(n=6)
B D
5(n=6)
B C
6(n=7)
A C
7(n=7)
A C
8(n=7)
A D
9(n=6)
B C
10(n=7)
A D

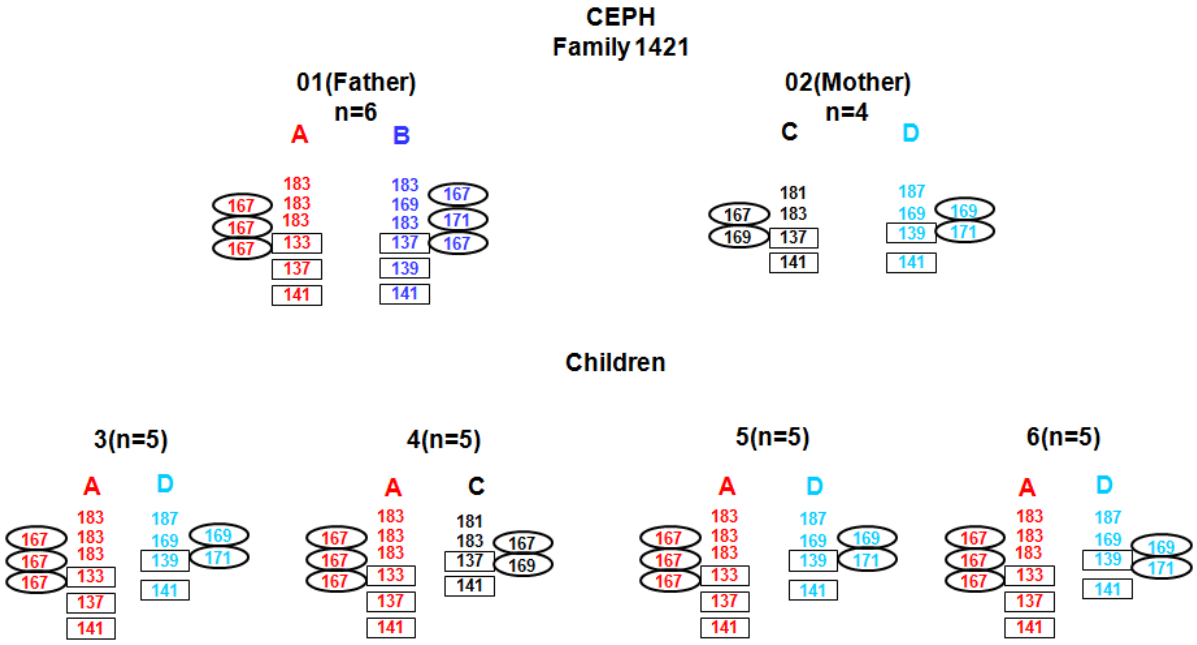

CEPH
Family 1350
10(n=4)
B
11(n=4)
A
12(n=5)
D
13(n=5)
C
01(Father)
n=4
A B
02(Mother)
n=5
C D
Children
3(n=5)
B C
4(n=5)
A C
5(n=5) R)mat)
A D/C
6(n=4) R(mat)
A C/D
7(n=4)
B D
8(n=4) R(mat)
B C/D
9(n=5) R(mat)
B D/C

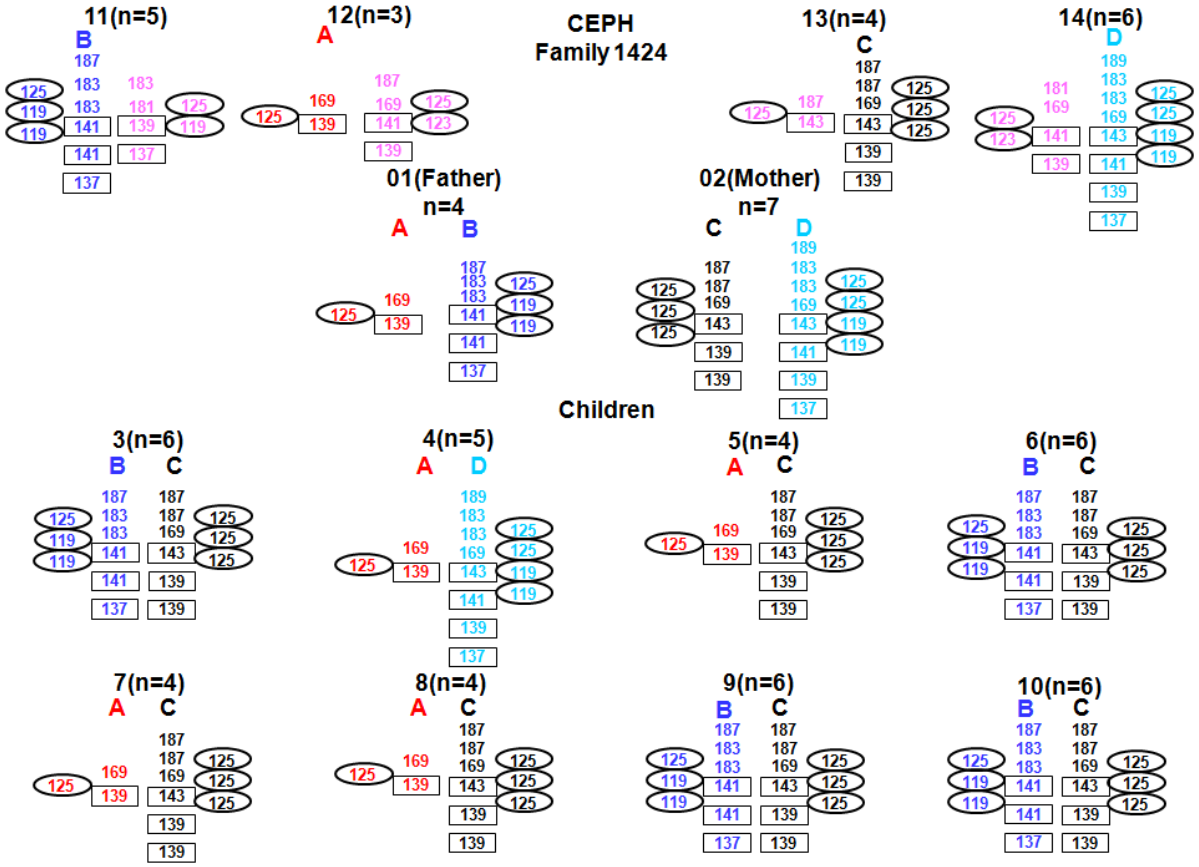

CEPH
Family 1362
13(n=5)
14(n=5)
15(n=5)
16(n=2)
01(Father)
n=5
02(Mother)
n=3
Children
3(n=4)
4(n=3)
5(n=4)
6(n=5)
7(n=3)
8(n=4)
9(n=4)
10(n=4) R(mat)
11(n=3) R(mat)
12(n=5)
17(n=4) R(mat)

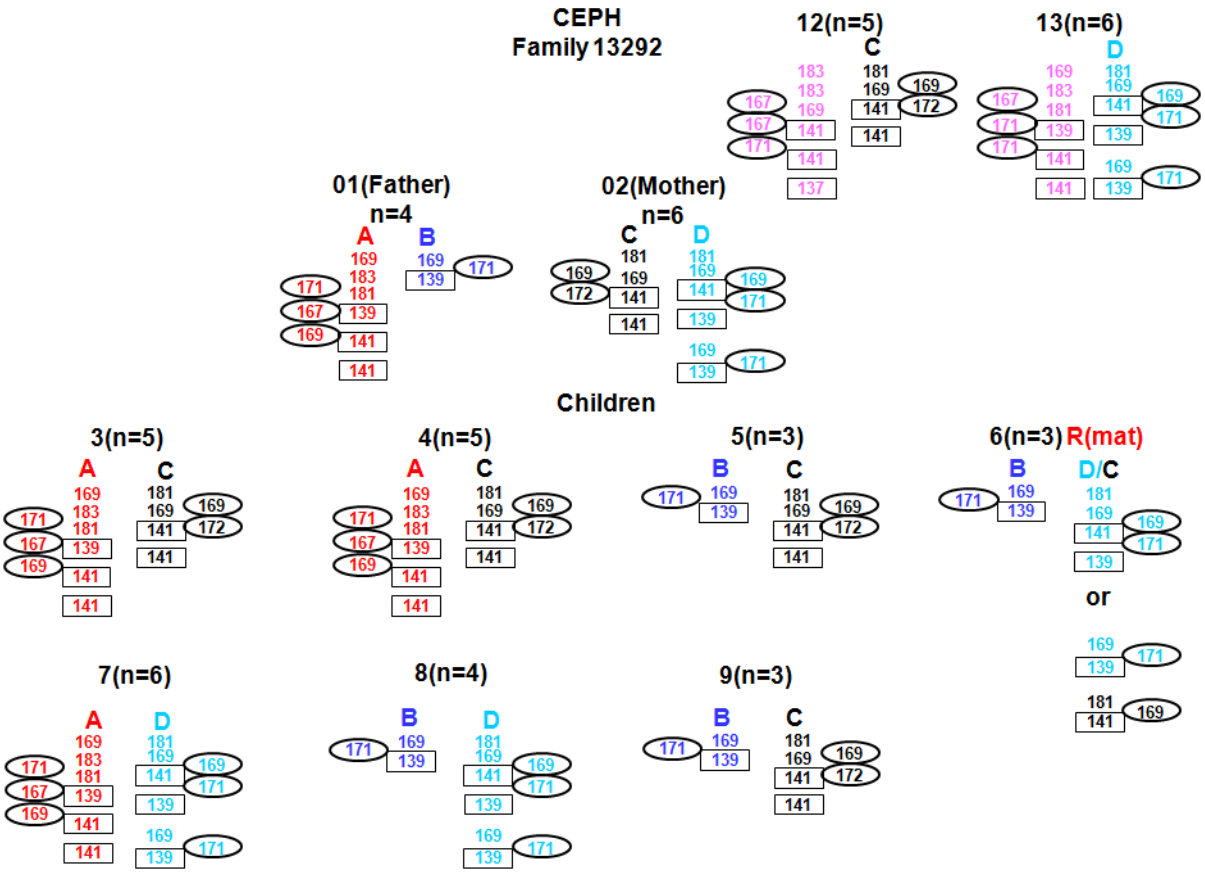

CEPH
Family 1375
9(n=4)
B
185 187
125 169 183 125
123 141 141 119
139 141
11(n=4)
D
183
183 125
125 169 169 119
139 141 119
139
137
12(n=3)
C
181
125 169 169 125
123 141 139
139
01(Father)
n=4
A B
181 185
125 169 169 125
123 141 141 123
139 139
02(Mother)
n=4
C D
183
183 125
125 169 169 119
139 141 119
139
137
Children
3(n=3)
B C
185
125 169 169 125
123 141 139
139
4(n=3)
B C
185
125 169 169 125
123 141 139
139
5(n=5)
B D
185 183
183 125
125 169 169 119
123 141 141 119
139 139
137
6(n=3)
B C
185
125 169 169 125
123 141 139
139
7(n=5)
A D
183
181 183 125
125 169 169 119
123 141 141 119
139 139
137
8(n=3)
B C
185
125 169 169 125
123 141 139
139
13(n=3)
A C
181
125 169 169 125
123 141 139
139

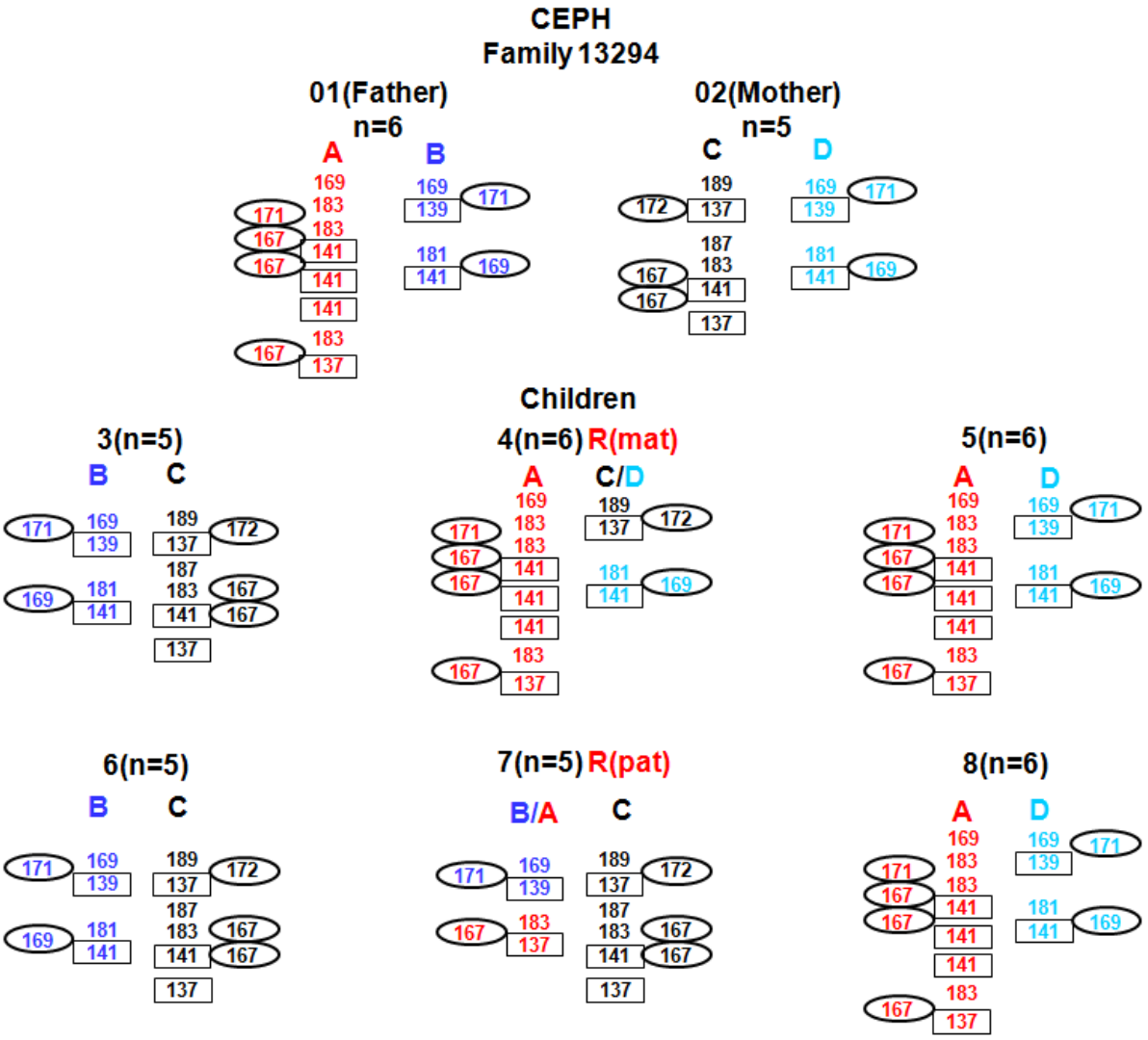

CEPH
Family 1408
10(n=3)
A
169
169
125
125
141
139
169
143
125
11(n=5)
B
187
183
183
125
189
183
125
123
143
119
143
119
141
141
137
12(n=5)
D
187
185
181
189
183
125
125
123
143
119
123
143
141
141
141
13(n=6)
C
183
187
183
183
125
169
183
125
119
181
125
141
143
119
139
141
*
139
01(Father)
n=4
A
B
169
189
169
183
125
125
125
143
123
141
139
141
02(Mother)
n=5
C
D
187
183
189
181
183
125
125
143
143
119
125
119
141
141
139
Children
3(n=4)
B
D
189
189
125
183
183
125
123
143
143
119
141
141
4(n=4)
B
D
189
189
125
183
183
125
123
143
143
119
141
141
5(n=5)
B
C
187
189
183
125
183
125
123
181
125
143
143
119
141
141
139
6(n=5)
A
C
187
169
183
125
169
183
125
125
141
181
125
139
143
119
141
139
7(n=5)
B
C
187
189
183
125
183
125
123
181
125
143
143
119
141
141
139
8(n=4)
A
D
169
189
169
183
125
125
141
143
119
139
141
9(n=5)
B
C
187
189
183
125
183
125
123
181
125
143
143
119
141
141
139
14(n=5)
A
C
187
169
183
125
169
183
125
125
141
181
125
139
143
119
141
139


CEPH
Family 1416
11(n=5)
12(n=4)
13(n=4)
14(n=6)
01(Father)
n=5
02(Mother)
n=5
Children
3(n=4)
4(n=5)
5(n=4)
6(n=5)
7(n=6)
8(n=5)
9(n=5)
10(n=5)
15(n=5)
16(n=5)


CEPH
Family 1421
01(Father)
n=6
A
B
183
183
183
133
137
141
167
167
167
183
169
183
137
139
141
167
171
167
02(Mother)
n=4
C
D
181
183
167
169
137
141
187
169
169
139
171
141
Children
3(n=5)
A
D
183
183
183
133
137
141
167
167
167
187
169
139
141
169
171
4(n=5)
A
C
183
183
183
133
137
141
167
167
167
181
183
137
141
167
169
5(n=5)
A
D
183
183
183
133
137
141
167
167
167
187
169
139
141
169
171
6(n=5)
A
D
183
183
183
133
137
141
167
167
167
187
169
139
141
169
171


CEPH
Family 1424
11(n=5)
12(n=3)
13(n=4)
14(n=6)
01(Father)
n=4
02(Mother)
n=7
Children
3(n=6)
4(n=5)
5(n=4)
6(n=6)
7(n=4)
8(n=4)
9(n=6)
10(n=6)


CEPH
Family 13292
12(n=5)
C
183 181
183 169 169
169 169 141 172
167 141 141
167 141
171 141
137
13(n=6)
D
169 181
183 169
181 141 169
167 141 171
171 139 139
141 141
141 169 171
139
01(Father)
n=4
A
169
183
181
171 139
167 141
169 141
141
B
169 171
139
02(Mother)
n=6
C
181
169 169
172 141
141
D
181
169
169 141
171
139
169 171
139
Children
3(n=5)
A
169
183
181
171 139
167 141
169 141
141
C
181
169 169
141 172
141
4(n=5)
A
169
183
181
171 139
167 141
169 141
141
C
181
169 169
141 172
141
5(n=3)
B
171 169
139
C
181
169 169
141 172
141
6(n=3)
R(mat)
B
171 169
139
D/C
181
169 169
141 171
139
or
169 171
139
181 169
141
7(n=6)
A
169
183
181
171 139
167 141
169 141
141
D
181
169 169
141 171
139
169 171
139
8(n=4)
B
171 169
139
D
181
169 169
141 171
139
169 171
139
9(n=3)
B
171 169
139
C
181
169 169
141 172
141


CEPH
Family 13294
01(Father)
n=6
A
B
169
183
171
167
183
141
167
141
141
183
167
137
169
171
169
139
181
169
141
02(Mother)
n=5
C
D
189
169
171
172
137
169
139
187
181
183
167
141
169
167
141
137
Children
3(n=5)
B
C
171
169
139
189
172
137
187
169
181
183
167
141
141
167
137
4(n=6) R(mat)
A
C/D
169
189
171
183
172
137
167
183
141
181
167
141
169
141
141
183
167
137
5(n=6)
A
D
169
169
171
183
139
171
167
183
141
181
167
141
169
141
141
183
167
137
6(n=5)
B
C
171
169
139
189
172
137
187
169
181
183
167
141
141
167
137
7(n=5) R(pat)
B/A
C
171
169
139
189
172
137
167
183
187
137
183
167
141
167
137
8(n=6)
A
D
169
169
171
183
139
171
167
183
141
181
167
141
169
141
141
183
167
137