Human β-defensin gene copy number variation and consequences in disease and evolution

Raquel Rodrigues Palla

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

July 2012

This PhD was supported by "Fundação para a Ciência e a Tecnologia" (FCT) (Foundation for Science and Technology) by a individual PhD scholarship with the reference SFRH / BD / 29753 / 2006, funded by the program POPH - QREN - Typology 4.1 – Advanced formation, co-participated by the European Social Fund and the Portuguese Ministry of Science, Technology and High Education.

FCT Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

ABSTRACT

Research on human genetic variation has shown that the human genome is not a fixed, invariant framework, but that there can be extensive structural variation. This variation includes copy number variation (CNV), which can lead to changes in DNA dosage contributing significantly to variation between individual human genomes and heritable traits.

Human β-defensins are small, secreted antimicrobial peptides encoded by *DEFB* genes located in a cluster of at least seven genes on 8p23.1. These genes are highly variable in copy number but accurate measurement of multiallelic copy number variants is challenging, particularly for high copy numbers, and has not been intensively studied until recently. A new PRT-based (Paralogue Ratio Test) triplex assay was developed to accurately measure the multiallelic β-defensin copy number variation. The Triplex assay was demonstrated to be an accurate and powerful method to measure copy number variation in large case-control association studies. This method was used to study the β-defensin CNV in psoriasis disease, showing that high β-defensin copy number is associated with susceptibility to psoriasis in Caucasians.

Studying population variation of CNV showed that variation in copy number of β-defensin is not significantly different across human populations. To understand the evolutionary history of beta-defensin CNV in the primate lineage, the study of CNV at this locus was carried out in great apes. β-defensin genes are copy variable in human and chimpanzee, but not in gorilla, suggesting that variation in copy number of beta-defensin genes may have arisen in the human-chimpanzee lineage after the divergence with gorilla.

ACKNOWLEDGEMENTS

I would like to start by thank my supervisor, Professor John A. L. Armour, for giving me the opportunity to integrate his research group and for the support and guidance during my time in his laboratory. Moreover, I also would like to thank for his special dedication to teaching and for the numerous linguistic lessons through these years.

My thank also goes to all members of C10 laboratory, past and present, for making the laboratory a cheerful and enjoyable place to do research: Fayeza Khan, Somwang Janyakhantikul, Danielle Carpenter, Tamsin Majerus, Sugandha Dhar, Holly Black, Dibo Pughikumo, Omniah Mansouri and the former members Ionnis Ladas, Emma Dannhauser, Susan Walker and Ed Hollox. A particular thank to Suhaili Abu Bakar with whom it was a joy to share the beta-defensin work and for being such a "happy go lucky" person, and to Jess Tyson for her support and advice, but most of all for being the "backbone" of the group and always be willing to help and listening.

I would like to thank to the Institute of Genetics and School of Biology for all the technical support given during my PhD and to the "Fundação para a

Ciência e a Tecnologia" (Foundation for Science and Technology) for the essential financial support.

I would like to express gratitude to my family who ever believed in this long-term project and who provide me with good family moments that helped me to clear my mind. To my parents, Fátima Rodrigues and Carlos Palla, I am extremely grateful for their continuous support and encouragement during my education and for making possible to realise my goals. They have always supported my scientific career and showed me that everything is possible. Lastly, I would like to dedicate a special thank to João Ferreira, for his endless support, understanding and an always positive willing. Thank you for sharing with me all these years and never letting me give up during this demanding project.

TABLE OF CONTENTS

TABLE OF FIGURES

ABBREVIATIONS

a-CGH Array-Based Comparative Genomic Hybridisation

BAC Bacterial Artificial Chromosome

BSA Bovine Serum Albumin

CEPH Centre d'Étude du Polymorphisme Humain

CD Crohn's Disease

CGH Comparative Genomic Hybridization

CNPs Copy Number Polymorphisms

CNV Copy Number Variation

CNVs Copy Number Variants

ddNTPs dideoxynucleotide triphosphates

DNA Deoxyribonucleic Acid

dNTP deoxynucleotide triphosphates

DSB double-stranded DNA breaks

EDTA Ethylenediaminetetraacetic acid

FISH Fluorescence In Situ Hybridisation

FoSTeS Fork Stalling and Template Switching

Indels Insertions and Deletions

kDa kiloDalton

LCRs Low Copy Repeats

LCVs Large-scale Copy Number Variations

Ld Low-dNTPs

LD Linkage Disequilibrium

MAPH Multiplex Amplifiable Probe Hybridization

MLPA Multiplex Ligation-dependent Probe Assay

MMBIR microhomology-mediated break-induced replication

MSVs MultiSite Variants

NAHR Non-Allelic Homologous Recombination

NHEJ Non-homologous end joining

PCR Polymerase Chain Reaction

PPRT Pyrosequencing-based Paralogue Ratio Test

PRT Paralogue Ratio Test

REPD REPeat Distal

REPP REPeat Proximal

REVDR Restriction Enzyme Digest Variant Ratios

RNA Ribonucleic Acid

RT-qPCR Real Time Quantitative PCR

SD Segmental Duplication

SNP Single Nucleotide Polymorphism

STRs Short Tandem Repeats

PUBLICATIONS RESULTING FROM THIS THESIS

Work from this thesis is reported in the following publications:

Armour, J. A. L., R. Palla, P. L. J. M. Zeeuwen, M. den Heijer, J. Schalkwijk and E. J. Hollox (2007). "Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats". Nucleic Acids Research **35**(3): e19. (Work included in this paper was executed before the start of the PhD).

Hollox, E. J., U. Hüffmeier, P. L. Zeeuwen, R. Palla, J. Lascorz, D. Rodijk-Olthuis, P. C. van de Kerkhof, H. Traupe, G. de Jongh, M. den Heijer, A. Reis, J. A. Armour and J. Schalkwijk (2008). "Psoriasis is associated with increased beta-defensin genomic copy number". Nature Genetics **40**(1): 23-25.

Aldhous, M. C., S. Abu Bakar, N. J. Prescott, R. Palla, K. Soo, J. C. Mansfield, C. G. Mathew, J. Satsangi and J. A. Armour (2010). "Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease". Human Molecular Genetics **19**(24): 4930-4938.

Stuart, P. E.*, U. Hüffmeier*, R. P. Nair*, R. Palla*, T. Tejasvi, J. Schalkwijk, J. T. Elder, A. Reis and J. A. L. Armour. "Association of β-defensin copy number and psoriasis in three cohorts of European origin". Manuscript submitted to the Journal of Investigative Dermatology.

* The authors wish it to be known that, in their opinion, the first 4 authors should be regarded as joint First Authors, and the last 3 authors as joint Senior Authors.

CHAPTER 1: INTRODUCTION

1.1 HUMAN GENETIC VARIATION

Ten years on after the first release of the human genome sequence (Human Genome Project) it is remarkable how our knowledge of the human genome has dramatically changed, bringing a new perspective into the real genetic variation that distinguishes and characterizes each individual identity, and defines us as humans.

Genomes can differ at many different scales, ranging from a single base to thousands of DNA nucleotide bases. Human genetic variation can be defined as the differences in DNA sequence observed between individuals. These genetic differences can be observed at large-scale (cytogenetic) level, comprising more than 3 Mb of DNA sequence and thus detectable by the light microscope, and at small-scale involving less than 3 Mb of DNA sequence and only detectable at the molecular (submicroscopic) level.

Until recently the most studied and best known types of genetic variation were at the sequence level, involving single nucleotides (indels, single nucleotide polymorphisms and point mutation), microsatellites and

minisatellites or at the chromosomal level, involving large chromosomal rearrangements. For a long time the knowledge of human genetic variation was limited by the detection methods available, and focused primarily on the heterochromatin polymorphisms large enough to be visible at the light microscope and to the sequence variation identified by traditional PCR-based DNA sequencing. Since then, new technologies filled the resolution gap between these earlier approaches to detect different classes of genetic variation, especially submicroscopic structural variation, contributing to a wider vision of the complex variation of the human genome. These technologies allowed the discovery of several genetic variants that play an important role in disease susceptibility. Since then, the study of human genetic diversity has become a major priority in human genetics as it is essential to understand the phenotypic differences observed between individuals, which either lead to "normal" human variation or confer susceptibility to disease.

1.1.1 Large-scale chromosomal variation

The evidence for human genetic variation at the DNA level started with the visualization of chromosomes under the light microscope (Tjio and Levan 1956; Speicher and Carter 2005). Since then, abnormal number or structure of chromosomes has been associated with disease phenotypes. The earlier cytogenetic techniques, without the use of staining methods, only allowed the visualization of differences in length, number and centromere position of the chromosomes, and therefore were only able to detect for example polyploidies, aneuploidies, marker chromosomes and isochromosomes (Jacobs and Strong 1959; Lejeune 1959). With the development of chromosome-banding techniques, such as G-banding (Rowley 1973; Caspersson *et al.* 1999) and fluorescence *in situ* hybridisation (FISH) (Bauman *et al.* 1980), more discrete chromosomal abnormalities could be detected, such as translocations, deletions, duplications and inversions (Speicher and Carter 2005). Since the first application of *in situ* hybridization using fluorescent microscopy by Bauman *et al.* (1980), *in situ* hybridization techniques have played a main role in mapping the position of genes on human chromosomes. Today, now that the Human Genome Project is complete, FISH and related *in situ* hybridization

methods are mainly used for clinical diagnoses. However, in other species, for which the entire genome has not been sequenced, this technique continues to be an essential tool to map the position of genes on chromosomes. The resolution of this technique has always been limited by the ability of light microscopy to distinguish between two points along the length of a chromosome. Typically the resolution for chromosomes at metaphase is expected to be resolution in the range of a few megabases (5-10 Mb), while for interphase chromosomes the resolution improves to the range of a few tens or hundreds of kilobases (50 kb to 2 Mb) (Levsky and Singer 2003; Speicher and Carter 2005). This method had a crucial impact in the understanding of chromosomal rearrangements, although the low resolution was a motivation to develop more powerful cytogenetic techniques such as fiber-FISH, which chromosome rearrangements can be identified on elongated chromatin strands (5 to 500 kb) (Florijn *et al.* 1995; Speicher and Carter 2005).

The numerical chromosomal variation which refers to the variation in the number of chromosomes relative to a diploid genome, can either involve the presence of more than two paired homologous sets of chromosomes, also known as polyploidy, or the presence of a numerical change in part of the chromosome set, aneuploidy, within the cells. Polyploidy can arise in metaphase I during meiosis in the germ line or during mitosis by abnormal cell division, leading to the formation of triploid (3n) or tetraploid (4n) genomes, respectively. The majority of embryos cannot survive with triploid or tetraploid genomes and are spontaneously aborted. Polyploidy of the entire genome is rare and normally lethal, but all humans have some polyploid cells, such as hepatocytes (2n to 8n) or cardiomyocytes (4n to 8n) (Korver *et al.* 1998).

Within the aneuploidies the most common numerical chromosomal abnormality is trisomy, which refers to the presence of three copies of a particular chromosome instead the normal two copies. However, aneuploidy can also involve the loss of one chromosome (monosomy) or a pair of chromosomes (nullisomy) or on the other hand, the gain of two or three chromosomes (tetrasomy, pentasomy), occurring either in autosomes or sex chromosomes. Aneuploidies normally arise when two homologous chromosomes fail to separate (disjoin) in anaphase, and migrate together into

the same gamete during mitosis or meiotic events, also called non-disjunction. Alternatively, aneuploidies can also result from the delayed movement (lagging) of a chromosome or chromatid, failing to be incorporated into one of the daughter cells during anaphase, referred as anaphase lag. Trisomies, in particular, have been described to be the cause of different autosomal syndromes, such as Down's syndrome, a trisomy of chromosome 21 (Lejeune 1959), Edward's syndrome, a chromosome 18 trisomy (Edwards *et al.* 1960) and Patau's syndrome, with an extra chromosome 13 (Patau *et al.* 1960). Down's syndrome was first described by Langdon Down in 1866 (Down 1866) as a clinical entity and one of the most common trisomies, possibly due to its high survival rate through adulthood. However, the majority of trisomic variations are not compatible with life; the larger and more gene-dense the chromosome involved the more likely the serious dysfunction. The occurrence of Down's syndrome has been associated with advanced maternal age and it occurs in about one in every 700 live births. Individuals with Down's syndrome share some mental and physical features, which can range from mild to severe symptoms. Often the individuals have learning disabilities and may have dementia. Another common feature is the high frequency of characteristic associated conditions like congenital heart disease and leukaemia.

Among the sex chromosomes, different syndromes have been described to be associated with chromosomal aneuploidies. An example is Turner's syndrome that is caused by the absence of all or part of the second sex chromosome. Frequently only one X chromosome is inherited with the other one lost during early development (2n=45, XO) (Jacobs *et al.* 1997). This syndrome occurs with a frequency of 1 in 2000 to 2500 live female births and is characterized by abnormal anatomical features, such as short stature and webbed neck, together with infertility and ovarian dysgenesis. Conversely, an additional X chromosome can be found in one of every 500 male newborns, causing Klinefelter's syndrome (47, XXY) (Jacobs and Strong 1959). Individuals with this condition mainly have problems in the reproductive system, namely infertility or reduced fertility, although some degree of developmental delay can also be observed.

Autosomal monosomies have far more lethal consequences than trisomies, which may be a direct consequence of the absence of expression of gene products encoded on the lost chromosome. On the other hand, the variation in the number of sex chromosomes has fewer negative effects in individuals than autosomal trisomies. This could be because the Y chromosome carries very few genes (as the major role is to determine the male sex) and because the X chromosome has an X chromosome inactivation mechanism, which results in dosage equalization (Ohno *et al.* 1959; Beutler *et al.* 1962).

1.1.2 Large-scale structural variation

Structural variation encompasses a group of genomic rearrangements which affect segments of DNA larger than 50 bp that can be microscopic (large-scale) or submicroscopic (small-scale). Depending on the size, structural variation normally involving DNA segments larger than 3 Mb (large-scale) can be detected by cytogenetic methods. Structural variation can be quantitative (comprising deletions, duplications, insertions, copy number variation, isochromosomes and marker chromosomes), positional (which refers to translocations) and orientational, as is the case of inversions. This section will discuss different types of large-scale structural variation.

Autosomal abnormalities involving the deletion of part of a chromosome are associated with recognized clinical syndromes, mainly associated with intellectual disability and several congenital malformations. Deletion of the short arm of chromosomes 5 and 4 was found to be the cause of Cri-du-chat and Wolf-Hirschhorn syndromes, respectively. Other deletions involving smaller segments of DNA (microdeletions) have also been associated with different syndromes, for example Prader-Willi (Ledbetter *et al.* 1981) and Angelman syndromes (Knoll *et al.* 1989). Patients with such syndromes have an interstitial deletion on the proximal portion of the long arm of chromosome 15 (15q11.2-13) either from paternal (Prader-Willi) or maternal derived chromosome (Angelman). The deletion commonly cover about 4 Mb of DNA sequence (Ledbetter *et al.* 1981; Knoll *et al.* 1989) but later studies in families defined an imprinting centre on human chromosome 15 estimated to span a minimum interval of 42 to 60 kb (Buiting *et al.* 1995). This structural

rearrangement involves the deletion of seven genes, which differ in expression according to paternal or maternal origin of the derived chromosome; Angleman syndrome is likely to be due to loss of maternal *UBE3A* function, whereas a more complex pattern of loss is necessary explain the features of Prader-Willi syndrome. Normally people inherit one copy of each gene from each parent but certain genes, for example the ones on chromosome 15, are expressed in a parent-of-origin-specific manner. This phenomenon of genomic imprinting explains the phenotypic differences associated with each syndrome (Ledbetter *et al.* 1981; Knoll *et al.* 1989).

Although many of the copy number variable regions are too small to be detected by cytogenetic methods, some CNVs showing large expansions of the repeat unit can be cytogenetically identified. Known examples are the copy number regions at 8p23.1, 15q11.2 and 16p11.2 (Barber *et al.* 1998; Ritchie *et al.* 1998; Barber *et al.* 1999; Hollox *et al.* 2003). The β-defensin gene cluster at 8p23.1 is highly polymorphic in copy number, with individuals showing between 2 and 12 copies per diploid genome. Using FISH, chromosomes with 7 or more copies of this repeat have been cytogenetically reported as euchromatic variants (Barber *et al.* 1998; Hollox *et al.* 2003). Similar to 8p23.1, at the 15q11.2 and 16p11.2 loci, chromosomes can contain up to ~20 and ~12 tandemly repeated copies, respectively, which appear as visible euchromatic variants (Ritchie *et al.* 1998; Barber *et al.* 1999).

Isochromosomes are a rare type of structural variation, consisting of a chromosome with two genetically and morphologically identical arms, either two long arms or two short arms. They may arise from recombination between sister chromatids due to the formation of an abnormal U-type structure. Isochromosomes have been observed in some females with Turner's syndrome, i(Xq), and in individuals with Down syndrome, i(21q) (Feuk *et al.* 2006). Marker chromosomes, also known as "supernumerary" chromosomes, are rare structurally abnormal chromosomes, and very little is known about their origin or importance, though their phenotypic effects are variable and depend on the genes contained within the marker (Feuk *et al.* 2006).

Structural chromosomal abnormalities can also result from misrepair of chromosome breaks or recombination between non-homologous regions of

chromosomes. Chromosome breaks occur as a result of damage or as part of the mechanism of recombination, but cells have natural mechanisms to prevent recombination between unrepaired chromosomes through chromosome break repair or apoptosis. Nevertheless, when these mechanisms fail, acentric (absence of centromere) or dicentric (two centromeres) chromosomes can be formed, which will normally be lost because they are not stable through mitosis (McClintock 1939; Stimpson *et al.* 2010). Conversely, chromosomes with only one centromere even with structural abnormalities, such as acrocentric (centromere located near one end of the chromosome) chromosomes, can stably recombine during meiosis frequently resulting in chromosomal translocations. Recombination of acrocentric chromosomes can occur between the proximal short arms of chromosomes resulting in the formation of a Robertsonian translocation and loss of the distal acentric parts of the two shorts arms (Robertson 1916; Hamerton *et al.* 1975). With the loss of one chromosome the karyotype will be reduced to 45 chromosomes. In Robertsonian translocations both centromeres are present but because they are very close together, they function as one, segregating normally. This translocation has been seen in all acrocentric chromosomes, namely 13, 14, 15, 21 and 22 (Stimpson *et al.* 2010). Another type of translocation is called reciprocal and results from the exchange of chromosome material of one of the arms of chromosomes between two non-homologous chromosomes (Evans *et al.* 1978). No phenotypic consequences are observed in the carriers of reciprocal or Robertsonian translocations, even though some genetic material has been lost at least in the Robertsonian translocations. For this reason such a translocation is regarded as a balanced structural variation. Nevertheless, carriers of translocations can produce unbalanced gametes and generate zygotes with chromosomal imbalances such as trisomy and monosomy. Moreover, some individuals show additional problems with infertility and recurrent spontaneous abortion (Berend *et al.* 1998).

Unlike other types of structural variation, inversions have been regarded as balanced rearrangements as they involve the change in orientation of a DNA segment without the loss of genetic material. However, Tuzun *et al.* (2005) demonstrated that inversions can often be not balanced, as they are frequently

accompanied by gain or loss of genetic material. Inversions can be pericentric if the inverted segment includes the centromere or paracentric if it does not contain the centromere. Although most inversions may not have a direct effect on the phenotype, there is increasing evidence that polymorphic inversions can predispose to further chromosomal rearrangements in subsequent generations (Sharp *et al.* 2006). A known example is the inversion flanked by the segmental duplicated olfactory receptor genes at 4p16 and 8p23. The occurrence of individuals with the recurrent translocation t(4;8)(p16;p23) is more frequent among those with parents who are carriers of the inverted heterozygous state of these inversions (Giglio *et al.* 2001). Therefore, inversion of regions flanked by segmental duplications can lead to an abnormal meiotic recombination and increased susceptibility to unequal non-allelic homologous recombination (NAHR) (Iafrate *et al.* 2004; Sebat *et al.* 2004; Sharp *et al.* 2005; Tuzun *et al.* 2005; McCarroll *et al.* 2006). Apart from translocations the inverted state may also be associated with inverted duplications and marker chromosomes (Sharp *et al.* 2006).

1.1.3 Small-scale structural variation

As described in the previous section, structural variation can also involve genomic rearrangements of smaller segments of DNA (<3 Mb). At this scale, genomic variation can involve one to several bases of DNA, including from single nucleotide polymorphisms (SNPs), short repetitive DNA sequences (microsatellites, minisatellites and mobile elements), and small insertions/deletions (less than 1 kb) to large copy number variants and inversions (Feuk *et al.* 2006). However, the detection of smaller and more abundant variation was only possible with the advent of new methods in molecular biology, in particular genome-scanning array technologies (Solinas-Toldo *et al.* 1997; Iafrate *et al.* 2004; Sebat *et al.* 2004) and comparative DNA-sequence analysis (Feuk *et al.* 2005; Tuzun *et al.* 2005). These techniques were especially important to detect intermediate-scale variation, since variation at the karyotype and nucleotide levels can be detected by cytogenetic and conventional sequence analysis, respectively. These intermediate variants ranging from ~1 kb to 3 Mb in size are defined as submicroscopic structural

variation (small-scale structural variation), which include copy number variants, segmental duplications or low copy repeats, deletions, insertions, inversions and translocations (Figure 1) (Feuk *et al.* 2006; Alkan *et al.* 2011). At first it was believed that small genetic variants (<1 kb), such as single base pair changes, would constitute most human genetic variation. However, in recent years hundreds of CNVs (Dhami *et al.* 2005; Sharp *et al.* 2005; Tuzun *et al.* 2005; Freeman *et al.* 2006; Redon *et al.* 2006; Conrad *et al.* 2010) and inversions (Stefansson *et al.* 2005; Tuzun *et al.* 2005; Visser *et al.* 2005) have been described. When referring to submicroscopic structural variation it is not directly implied that it has obvious phenotypic consequences, but because this type of variation encompasses millions of bases of DNA which can comprise entire genes, it is likely to influence gene dosage and therefore cause or give susceptibility to genetic diseases (Inoue and Lupski 2002; The Wellcome Trust Case Control Consortium 2010). In such cases, the term structural abnormality is used. Some of these structural rearrangements can occur in more than 1% of the population and can be called polymorphisms (Feuk *et al.* 2006; Alkan *et al.* 2011).

Insertion and deletion (indels) of DNA segments involve gain and loss of few hundred to several million nucleotides of the genome and are collectively known as copy number variants. Small indel events represent the most common type of structural variation in the human genome and are a major source of human biological diversity. In a recent study almost 2 million small indels were reported in the human genomes of 79 individuals, comprising between 1 bp to 10000 bp (Mills *et al.* 2011). This type of rearrangement was associated with earlier identified human genetic traits, such as colour blindness (Nathans *et al.* 1986; Vollrath *et al.* 1988) and α-thalassaemia (Vollrath *et al.* 1988; Higgs *et al.* 1989).

The red and green opsin pigment genes, *OPN1LW* ("opsin 1 long-wave-sensitive" for the red photopigment) and *OPN1MW* ("opsin 1 medium-wave-sensitive" for the green photopigment) are located head-to-tail in a tandem array on the human X chromosome and share a high sequence similarity, which in combination predispose to unequal recombination or gene conversion. This mechanism is responsible for the fusion of red and green gene pigments and the observed variation in green-pigment gene number among individuals with colour blindness (Nathans *et al.* 1986; Vollrath *et al.* 1988).

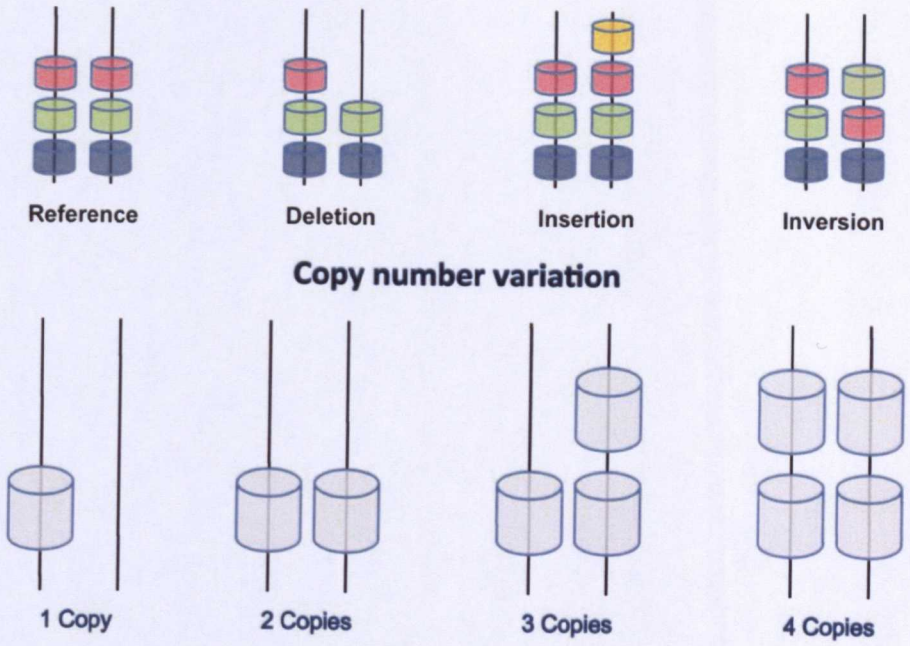


Figure 1: Schematic representation of examples of structural genomic variation involving rearrangements of one or more regions of DNA sequence. Top panel illustrates deletion, insertion and inversion of a DNA segment relative to the reference genome (each colour represents a different DNA sequence). On the bottom panel is illustrated the variation in copy number of a DNA segment of at least 1kb relative to the common (reference) situation of 2 copies per diploid genome.

Non-allelic homologous recombination between paralogous sequences is also responsible for the deletion of α-globin genes that result in α-thalassaemia. Normal individuals have four alpha-globin genes, with one copy of each highly homologous gene, alpha-1 (*HBA1*) and alpha-2 (*HBA2*), present on each chromosome 16. High sequence similarity between these genes was thought to drive unequal recombination and lead to recurrent deletion of one or two copies of the α-globin genes causing α-thalassaemia (Higgs *et al.* 1989). On the other hand, duplication of these genes was also observed in individuals from different ethnic backgrounds, showing up to five copies of alpha-globin genes with apparently no clinical or haematological abnormalities (Goossens *et al.* 1980). Another example is the case of the Rhesus blood group antigens. The *RHD* gene at chromosome 1p34.1-1p36 is flanked by two DNA segments called Rhesus boxes, with 98% homology. This molecular structure supports the occurrence of unequal crossing over between the Rhesus boxes accompanied by the frequent deletion of the entire *RHD* gene (depending on the population between 3% and 25% in Caucasians) (Ottolenghi *et al.* 1974; Wagner and Flegel 2000).

More recently two studies using dense SNP genotype data from family trios from the International HapMap Project identified several (around 500 each study) deletion polymorphisms demonstrating the importance and extent of copy number polymorphisms (CNPs) in the human genome (Conrad *et al.* 2006; McCarroll *et al.* 2006). The deletion polymorphisms identified range from 1 kb to 750 kb spanning several genes with roles in sex steroid metabolism, olfaction and drug response (McCarroll *et al.* 2006). In particular, deletions of 5 kb or larger were found to be widespread in the human genome proving that the abundance of this type of polymorphism had been underestimated (Conrad *et al.* 2006). This evidence is supported by Hinds *et al.* (2006) who used array-based comparative genomics to detect 215 deletions in 24 unrelated individuals. Overall, around 1.5 million indels were identified within known genes in the human genome, with about 5,500 of them located in the promoters and exons of genes where it is expected they could influence gene expression and function (Mills *et al.* 2006). Indel polymorphisms in several genes have also been reported to influence a variety of common phenotypes. Homozygous deletions of the glutathione S-transferase genes (*GSTM1* and *GSTT1*) for example are associated with increased risk of a variety of cancers (García-Closas et al. 2005), whereas duplication of the *CYP2D6* gene, a cytochrome P450 CYP2D6 drug-metabolizing enzyme involved in the metabolism of several drugs and xenobiotics, causes alteration of the metabolism of CYP2D6 substrates and is also a risk factor for lung or larynx cancer (Agúndez *et al.* 2001).

Another class of genetic variation involves changes in the number of repeated DNA sequences arranged in tandem arrays, generally known as variable number of tandem repeats (VNTRs). This class covers different scales of variation in which microsatellites and minisatellites, tandem arrays with variable repeat units typically between 2 and 100 bp in length and also satellite DNA, which comprises large tandem arrays spanning from hundreds of kilobases to several megabases in size, are included. Microsatellites, also known as short tandem repeats (STRs), are tandem DNA arrays composed of repeat units 1 to 6 bp in length. Typically these arrays contain between 10 and 30 repeats spanning not more than approximately 200 bp. On the other hand, the

repeat array length of minisatellites normally spans from 100 bp to 20 kb with each repeat unit between 8 and 100 bp in length.

The substantial length polymorphism of VNTRs, in particular of microsatellites and minisatellites, makes these useful and very informative genetic markers to study genetic diversity. The highly polymorphic length of minisatellites, due to allelic variation in repeat copy number, was discovered in 1985 by Jeffreys *et al.* (1985a; 1985b) and was subsequently used in DNA fingerprinting. The highly variable microsatellite profiles found in the human genome can be used in individual identification, paternity testing and forensic studies. While some satellite DNA normally lies at the centromeres and telomeres, composing functional components of chromosomes, some microsatellites and minisatellites lie within the coding regions of genes or in regulatory regions. Although some tandem repeat loci are seen as neutral markers, others can play a functional role and variation can affect gene expression and function and therefore have phenotypic effects (Armour 2006). Well known examples are the "triplet repeat diseases" in which expansions of microsatellite arrays are associated with single-gene inherited disorders. One example is fragile-X syndrome, which is associated with expansion in the number of tandem CGG trinucleotide repeats upstream of the human *FMR1* gene, normally from 5-50 repeats to 50-4000 repeats, leading to reduced or silenced transcription (Pieretti *et al.* 1991; Hornstra *et al.* 1993). Other well known examples of triplet repeats are those caused by CAG (glutamine) tandem repeats within the coding region of a gene. The polyglutamine expansions were described in Huntington disease and in a number of spinocerebellar ataxias (SCAs) and cause a dominant gain-of-function neurotoxicity (MacDonald *et al.* 1993; Michalik and Van Broeckhoven 2003). The cause of the autosomal dominant neuromuscular disorder, facioscapulohumeral muscular dystrophy (FSHD) has also been linked to satellite repeat markers. Increased evidence suggests that an epigenetic mechanism causes FSHD involving the contraction of a tandem DNA repeat (D4Z4) on chromosome 4q35. However, is still unclear how this deletion causes the disease (van der Maarel and Frants 2005). In addition, minisatellite polymorphisms were also reported to have influence on phenotype. Upstream

of the insulin gene *INS* (=*IDDM2,* insulin-dependent diabetes mellitus 2) on human chromosome 11p15.5, there is a polymorphic minisatellite which was described as a susceptibility factor for type I diabetes (Bennett *et al.* 1995). Transcription of the insulin gene was correlated with allelic variation of this minisatellites (Bennett *et al.* 1995).

Analysis of DNA sequence data reveals that the human genome comprises an estimated 5.2% of duplicated sequences (segmental duplications, SDs), also termed low-copy repeats (LCRs) (Bailey *et al.* 2001; Bailey *et al.* 2002; Cheung *et al.* 2003; She *et al.* 2004). Segmental duplications are blocks of DNA that range from 1 to 400 kb in length of highly homologous DNA sequences sharing a high degree of sequence identity, approximately 98% (Eichler 2001). SDs are interspersed throughout the human genome, but they tend to occur at pericentromeric and subtelomeric regions. They can be duplicated within a single chromosome (intrachromosomal duplication) or in non-homologous chromosomes (inter or transchromosomal duplication) (She *et al.* 2004; Linardopoulou *et al.* 2005). The presence of large segmental duplications predisposes these regions to recurrent structural rearrangements via non-allelic homologous recombination (NAHR), resulting in the formation of different structural polymorphisms, including deletion, duplication or inversion of intervening DNA segments (Iafrate *et al.* 2004; Sebat *et al.* 2004; Sharp *et al.* 2005; Tuzun *et al.* 2005; McCarroll *et al.* 2006). The location of SDs has been associated with regions of chromosomal instability or at the breakpoints of syntenic blocks in the human genome suggesting their involvement in evolutionary rearrangements (Samonte and Eichler 2002; Armengol *et al.* 2003; Bailey *et al.* 2004). These segments can contain dosage-sensitive genes or regulatory regions, and as such they have been named as potential mediators in a number of disease phenotypes, designated genomic disorders (>25 recurrent genomic disorders) (Lupski 1998; Ji *et al.* 2000; Inoue and Lupski 2002; Stankiewicz and Lupski 2002). One example involves the recombination (NAHR) between the *CMT1A* (Charcot-Marie-Tooth disease type 1A) repeats on chromosome 17p containing the *PMP22* gene, which encodes a myelin protein. Duplication of this region (three copies of *PMP22* gene) leads to Charcot-Marie-Tooth disease, while the reciprocal deletion is correlated with

liability to pressure palsy (*HNPP*) (Lupski 2009). Further examples include the Smith-Magenis syndrome and Potocki-Lupski syndrome which result from similar rearrangement mechanisms (Lupski 2009). NAHR has also been involved in the formation of large inversion polymorphisms. Two studies, using different comparative genomic approaches identified several inversion polymorphisms with the majority (>75%) mapping to sites of segmental duplications (Feuk *et al.* 2005; Tuzun *et al.* 2005). One of the best known examples of inversion polymorphisms reported to date is the 900 kb inversion at chromosome 17q21.31 identified by Stefansson *et al.* (2005). The inverted haplotypes appear at increased frequency in the European population (21%) but at very low frequency in Africans (6%) and Asians (1%) with female carriers of the inversion showing a small increase in fertility (Stefansson *et al.* 2005). Although inversion polymorphisms have not been associated with alterations in copy number and subsequently have not been suggested to be a direct cause of phenotypic variation, many of them confer susceptibility or are the substrate to further chromosomal rearrangements in the following generations (Sharp *et al.* 2006). Inversion polymorphisms were frequently observed in parents of deletion carriers that confer different disease phenotypes, such as Angelman syndrome (Gimelli *et al.* 2003), Sotos syndrome (Osborne *et al.* 2001) or Williams-Beuren syndrome (Visser *et al.* 2005). Therefore, inversion of the segment flanked by SDs may predispose to abnormal meiotic pairing leading to unequal NAHR and subsequently the formation of several chromosomal rearrangements that confer disease phenotypes.

1.1.3.1 Copy number variation and its consequences

In the past few years copy number variation (CNV) has been demonstrated to be a prevalent form of structural variation in the human genome and an important source of genetic variation (Iafrate *et al.* 2004; Sebat *et al.* 2004; Redon *et al.* 2006; The Wellcome Trust Case Control Consortium 2010). CNV can be defined as a DNA segment 1 kb or greater in size, which is present at variable number of copies (Feuk *et al.* 2006). Numerous copy number polymorphisms (CNPs) have been associated with the presence of segmental duplications (Sharp *et al.* 2005). Genomic regions enriched for

segmental duplications are prone to chromosomal rearrangements through NAHR and are considered hotspots of genomic instability and therefore prone to copy-number variation (Sharp *et al.* 2005). Copy number variation can involve just simple tandem duplications (also referred as low copy repeats or segmental duplications) or more complex arrangements, such as gains or losses of homologous sequences at different sites in the genome (Redon *et al.* 2006).

Three studies published between 2004 and 2006 strongly contributed to the perception that copy number variation has a widespread distribution throughout the human genome accounting for a large portion of the genetic variation found among human genomes, which may reflect their importance in genetic diversity and evolution (Iafrate *et al.* 2004; Sebat *et al.* 2004; Redon *et al.* 2006). Iafrate *et al.* (2004) identified 200 large-scale copy-number variants (LCVs) in the human genome, of which 24 are present in >10% of the studied individuals. However, by studying 270 individuals from four populations with different ancestry (Europe, Africa and Asia from the HapMap collection) Redon *et al.* (2006) identified 1447 copy number variable regions (CNVRs) encompassing as much as 360 megabases of the genome (~12%), containing hundreds of genes with important biological functions.

CNV has been described to influence gene expression, phenotypic variation and adaptation by disrupting genes and altering gene dosage (Iafrate *et al.* 2004; McCarroll *et al.* 2006), which can lead to disease phenotypes or confer susceptibility to complex traits such as psoriasis (Hollox *et al.* 2008b) or glomerulonephritis (Aitman *et al.* 2006). Many copy number variants directly affect the protein levels, as in the case of the α-synuclein (*SNCA*) gene (Miller *et al.* 2004); however for α-defensin genes the expression of *DEFA1/DEFA3* mRNA is not simply correlated with the gene copy number (Aldred *et al.* 2005). Studying the relative impact of CNV on gene expression phenotypes in 210 unrelated individuals of the International HapMap project, Stranger *et al.*, found that 17.7% of the detected genetic variation in gene expression is due to copy number variation (Stranger *et al.* 2007), suggesting that copy number variation is not limited to intergenic or intronic regions. Also reporting similar conclusions, a more recent study using the mouse as a model system showed that CNVs shape tissue transcriptomes and therefore phenotypic variation. The

Genome-wide expression data generated from six major organs confirmed the influence that CNVs have on the expression of genes within the CNV and in their vicinity, an effect that could be extended up to half a megabase (Henrichsen *et al.* 2009). Despite the direct effect that large-scale copy number variants have in disease phenotypes, these variants are common among the genomes of phenotypically normal individuals, suggesting an indirect role in the phenotype through position effects or predisposing to further chromosomal rearrangements in the following generations with potentially deleterious effects. Therefore, they have been proposed as a driving force for genome evolution and phenotypic variation (Sebat *et al.* 2004; Feuk *et al.* 2006).

Perhaps the best and simplest example of copy number variation is the well known deletion of the *RHD* gene responsible for the rhesus negative blood group. The Rh (Rhesus) blood group antigens are products of two genes, *RHD* and *RHCE*, located on chromosome 1p34-36.2. *RHD* carries the D antigen, one of the most potent immunogenic blood groups. The deletion of the whole *RHD* gene occurs with an approximate frequency of 40% in Europeans and corresponds to the rhesus-negative phenotype and complete absence of the D antigen expression (Wagner and Flegel 2000).

One of the commonest copy number variants described contains the human salivary amylase gene (*AMY1*) at chromosome region 1p13.3, that spans between 150 kb and 425 kb (Groot *et al.* 1989; Iafrate *et al.* 2004). Relative copy number gains and losses of this locus were detected at about the same frequency within a population, but the CNV was found to vary between different populations sampled from Africa, Asia and Europe, according to the type of diet (Iafrate *et al.* 2004; Perry *et al.* 2007). Populations with a diet rich in starch were found to have more copies of the *AMY1A* gene than populations with "low starch" diets (Perry *et al.* 2007).

Copy number variation in the human genome has been described as an important determinant for susceptibility to various autoimmune diseases. Aitman *et al.* (Aitman *et al.* 2006), described that variation in copy number of the *FCGR3B* (Fcy-receptor-IIIB) gene at chromosome 1q23, which encodes an Fc receptor for IgG, confers susceptibility to immunologically mediated glomerulonephritis, a major cause of kidney failure and morbidity in systemic

lupus erythematosus (SLE). Diploid copy number varies between zero and four copies, but low copy number of *FCGR3B,* in particular complete *FCGR3B* deficiency, was associated with glomerulonephritis in patients with systemic lupus erythematosus (SLE) (Aitman *et al.* 2006; Fanciulli *et al.* 2007). SLE was recently also associated with the complement component *C4* gene copy number variation, which ranges from zero to six copies per diploid genome. After comparison with healthy individuals, a clear high frequency of low copy number was detected among the patients with SLE (Yang *et al.* 2007). The risk ([OR]=6.514; p=0.00002) of SLE disease susceptibility significantly increased among individuals with only two or fewer copies of *C4* gene, while in comparison higher copy numbers (3 or more copies) confer protection against SLE (OR=0.574; p=0.012) (Yang *et al.* 2007).

The CCL3L1 locus is a multiallelic copy variable region on the q-arm of chromosome 17. In Europeans *CCL3L1* gene copy number is commonly polymorphic, varying between zero to four copies, but in African and Asian populations it is even more polymorphic with copy numbers as high as fourteen recorded in African individuals (Walker *et al.* 2009). Variation in copy number at *CCL3L1* has been reported to be associated with disease phenotypes such as SLE (Mamtani *et al.* 2008), Rheumatoid Arthritis (McKinney *et al.* 2008) and HIV-1/AIDS (human immunodeficiency virus/acquired immunodeficiency syndrome) (Townson *et al.* 2002; Gonzalez *et al.* 2005). *CCL3L1* gene encodes the CC chemokine ligand 3 like-1 (CCL3L1), which binds to several pro-inflammatory cytokine receptors, including the CC chemokine receptor 5 (CCR5) a major co-receptor for HIV (Human immunodeficiency virus). Low copy number of *CCL3L1* gene, relative to population average of the ethnic group, was associated with HIV-1/AIDS susceptibility and progression (Townson *et al.* 2002; Gonzalez *et al.* 2005). This finding is strongly controversial and in recent studies this association failed to replicate (Shao *et al.* 2007; Urban *et al.* 2009). Nevertheless, different studies have shown evidence of the relationship between gene CNV and susceptibility to human complex disease related with immunity.

Copy number variants have also been associated with other mendelian and complex diseases, such as Parkinson's and Alzheimer's diseases. A study

in patients with Parkinson disease, reported the frequent occurrence of genomic duplications or triplications of the alpha-synuclein gene (*SNCA*), which cause early-onset Parkinsonism with dementia (Chartier-Harlin *et al.* 2004), whereas other studies reported the importance of the duplication of amyloid precursor protein (APP) in susceptibility to Alzheimer's disease (Sleegers *et al.* 2006; McNaughton *et al.* 2012).

There have been several reported examples of genomic copy number variation, either involving simple deletion and duplication (diallelic) of a gene or more complex copy number involving multiple copies of genomic segments (multiallelic) encompassing several genes, parts of genes or intronic regions (Wain *et al.* 2009). One of such multiallelic copy number variant is the human β-defensin locus at chromosome 8p23.1. This locus is highly polymorphic in copy number and is a fascinating example of structural variation in the human genome. However, not many studies have reported the variation at this locus, although considering the antimicrobial activity (Ganz 2003) and the anti-HIV properties (Klotman and Chang 2006) of β-defensins, a growing interest can be expected in the way the β-defensin CNV is associated with some immune diseases, such as psoriasis (Hollox *et al.* 2008b). The study of β-defensin copy number variation and the phenotypic relevance of this variation has been the focus of my project and will be explored in more detail in further sections.

1.1.3.2 Human 8p23.1 locus

The human 8p23.1 locus, located towards the telomeric end of chromosome 8p has a complex genomic structural organization showing several types of structural variation. The region spans approximately 6.5 Mb and contains two large segmental duplications (SDs), the distal repeat (REPD) and the proximal repeat (REPP), about 5 Mb apart, sharing a 95-97% sequence identity (Giglio *et al.* 2001). More than 50 different genes have been found at 8p23.1, including defensin genes, olfactory receptor genes and FAM90A cluster genes (Figure 2). These gene clusters are copy number variable and are located in the REPD but no defensin genes were reported at REPP until 2009 (Hollox *et al.* 2003; Aldred *et al.* 2005; Bosch *et al.* 2008). The genome assembly shows defensin genes, which include alpha (α), beta (β) and theta (θ) defensins at

REPD. However, in 2009, Abu Bakar *et al.* mapped a β-defensin genomic locus to REPP, which is not annotated on the genome assembly, through the analysis of crossover breakpoints in CEPH segregation data (Abu Bakar *et al.* 2009). The β-defensin cluster at both locations is highly polymorphic in copy number and varies independently from the other gene clusters (Groth *et al.* 2008).

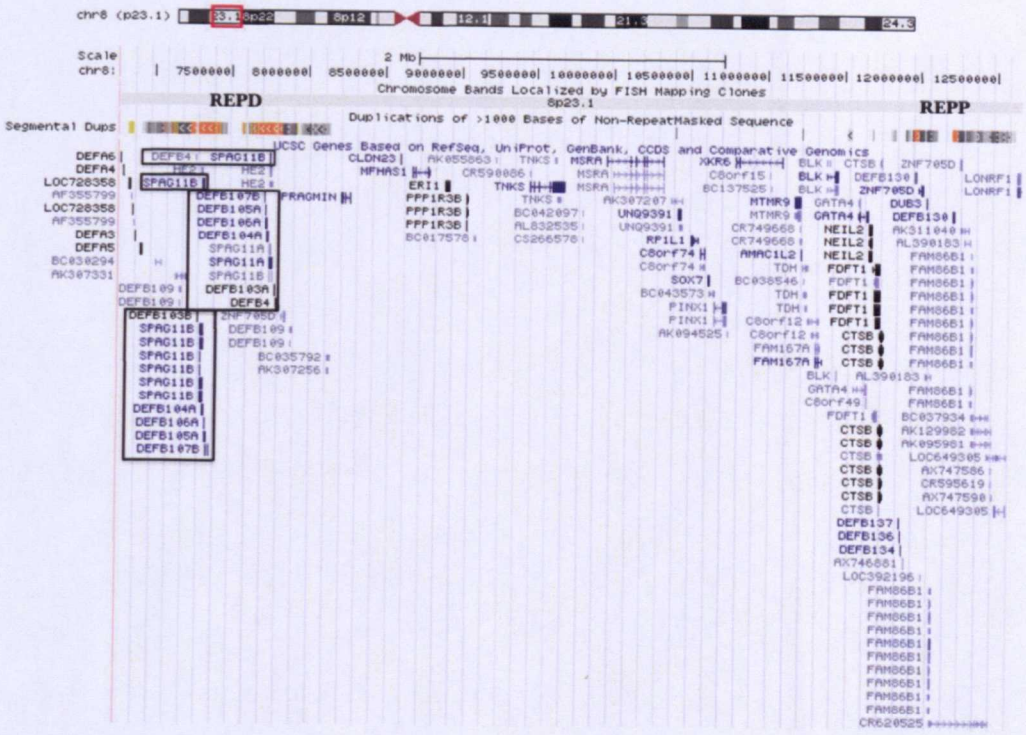


Figure 2: Chromosome 8p23.1 locus spanning approximately 6.5 Mb, showing two pairs of segmental duplications, REPD (left) and REPP (right) represented by grey and orange shading. Genes annotated in the UCSC Human Genome Browser are shown relative to their position in the chromosome (β-defensin genes are highlighted). At the REPD site two β-defensin clusters are shown in inverted orientation relative to one another (Figure from UCSC Genome Browser on Human March 2006 (NCBI36/hg18) assembly).

SDs can mediate frequent chromosomal rearrangements leading to the formation of several types of structural variants, such as copy number variation and inversion polymorphisms by homologous unequal recombination between the two duplicated regions at 8p (Giglio *et al.* 2002; Sugawara *et al.* 2003). Although this suggests the importance of SDs in the formation of the rearrangements observed at 8p23.1, the genome assembly sequence is still incomplete for these loci.

Allelic recombination between the two distinct genomic locations of β-defensins was demonstrated to be the main mechanism for β-defensin copy

number formation, showing a germ-line rate of copy number change of approximately 0.7% per gamete (Abu Bakar *et al.* 2009). At the end points of the β-defensin CNV repeat, spanning the region between REPD and REPP, a large inversion polymorphism (~4.7 Mb) was first reported by Giglio *et al.* (2001). This polymorphism is present in 25% of Europeans and approximately 33% of Japanese are heterozygous for the inversion polymorphism (Giglio *et al.* 2001; Giglio *et al.* 2002). Moreover, higher inversion frequencies were estimated in CEPH families (60%) (Chen *et al.* 2006a) and in three HapMap populations (42.6%), which showed higher frequencies, between 50% and 60% in Europeans and Africans but considerably lower frequency in Asians, only 12.5% (Antonacci *et al.* 2009). Taking into consideration the relative high frequency of the inverted orientation it was suggested that the human reference genome actually represents the minor allelic configuration, corresponding to the non-inverted orientation (Chen *et al.* 2006a). Analysis of multiallelic length polymorphisms at the 8p23.1 locus provided evidence that the inverted state of this polymorphism has arisen more than once (Abu Bakar *et al.* 2009).The inversion polymorphism will influence the type of rearrangements generated when the two loci recombine, which will be dependent on the orientation status of the sequence spanning the interval between REPD and REPP. Therefore, the inversion will have an impact on the allelic recombination between the two loci and on the *de novo* generation of copy number haplotypes. The two types of structural variation seem to be closely associated and both might be involved in the formation of copy number variation during recombination.

The presence of SDs seems to be the substrate for the formation of intrachromosomal rearrangements involving chromosome 8p by unequal recombination between REPD and REPP, leading to the formation of different chromosomal rearrangements, including the common inversion polymorphism (Giglio *et al.* 2002; Sugawara *et al.* 2003).

1.1.3.3 Mechanisms of copy number formation

Increasing evidence shows that CNVs often occur in regions of segmental duplications or flanked by segmental duplications (Iafrate *et al.* 2004; Sharp *et*

al. 2005; Tuzun *et al.* 2005). Several studies suggest that segmental duplications are hot-spots for chromosomal rearrangements, since segmental duplications increase the chance of non-allelic homologous recombination (NAHR)(Figure 3A) (Lupski 1998; Sharp *et al.* 2006). Segmental duplications by definition share more than 90% sequence similarity and cover at least 1 kb. They can arise either by tandem repeats of a particular DNA fragment or through a duplicative transposition-like process that results in the copying of a DNA segment and its transposition from one location to another (Eichler 2001).

During NAHR, highly similar duplicated sequences on the same chromosome can easily recombine and generate rearrangements, such as copy number changes of the segmental duplicated regions (Inoue and Lupski 2002). The recombination of segmental duplicates can occur between paralogous sequence on the same chromatid (which do not lead to deletions or duplications), between sister chromatids or between homologous chromosomes. Repeats that are further apart are also less likely to experience NAHR than closer repeats. When duplicated sequences are in direct orientation, they can result in deletions or duplications, but when occurring in inverted orientation, recombination will lead to inversions (Tuzun *et al.* 2005; Turner *et al.* 2008). A comparison of the *de novo* mutation rate of deletions against duplications in the male germline reported a higher rate of deletion generation (Turner *et al.* 2008). However, these observations are not concordant with the fact that deletions and duplications are present at similar frequency in the genome. Such facts might support the hypothesis that deletions and duplications are subjected to different selective pressures, resulting in directional selection favouring duplications (Freeman *et al.* 2006).

Although extensively described as a possible origin of CNV, segmental duplications overlap just 24% of the total copy number variants and therefore cannot account for the formation of several human CNV regions (Redon *et al.* 2006). Another possible mechanism of CNV formation appears to be related to non-homology based mutational mechanisms, which involve non-β DNA structures. Such DNA structures can promote chromosomal rearrangements and seem to be the dominant mechanism among smaller copy repeats (Bacolla

and Wells 2004). Nevertheless, little evidence of the real consequences of these events in the construction of new copy number variants has been reported.

More recently other mechanisms have been proposed to explain part of the non-recurrent genomic rearrangements including copy number variation: non-homologous end joining (NHEJ) (Figure 3B) and the Fork stalling and template switching/microhomology-mediated break induced replication (FoSTeS/MMBIR) (Figure 3C). NHEJ is a frequent mechanism used by human cells to repair DNA double-stranded breaks (DSB), involving the modification and rejoining of the two DNA break ends. In order to repair ends, modification of the ends needs to occur to make them compatible and ligatable, which can involve addition or deletion of bases of the segment between the two DSBs. Conversely to NAHR, this mechanism does not require the presence of LCRs or segments sharing high sequence similarity (Lieber *et al.* 2003; Gu *et al.* 2008).

A replication error-based mechanism was proposed to explain the origin of some complex non-recurrent rearrangements (e.g. deletions and/or duplications interrupted by either normal or triplicated genomic segments), such as the ones reported in the dosage-sensitive proteolipid protein 1 (*PLP1*) gene that cause Pelizaeus-Merzbacher disease (PMD), which have been difficult to explain by either NAHR or NHEJ recombination mechanisms (Lee *et al.* 2007; Zhang *et al.* 2009b). The FoSTeS/MMBIR event occurs during DNA replication. When replication fork stall, the 3' primer end of a DNA strand can change templates to a ssDNA template in a nearby replication fork, depending of microhomology (4-15 bp) at the 3' end, "priming" and DNA synthesis reinitiation (Lee *et al.* 2007). Depending on the location of the new replication fork, upstream or downstream of the original replication fork, a deletion or duplication will occur. Similarly, the orientation of the incorporated fragment, direct or inverted, is dependent on whether it is the lagging or leading strand, respectively, that is used as a template in the new replication fork (Hastings *et al.* 2009a). DNA replication-based mechanism of FoSTeS/MMBIR may be a important mechanism for generating structural variation, namely non-recurrent CNVs and complex genomic rearrangements in the human genome (Hastings *et al.* 2009b; Zhang *et al.* 2009a).

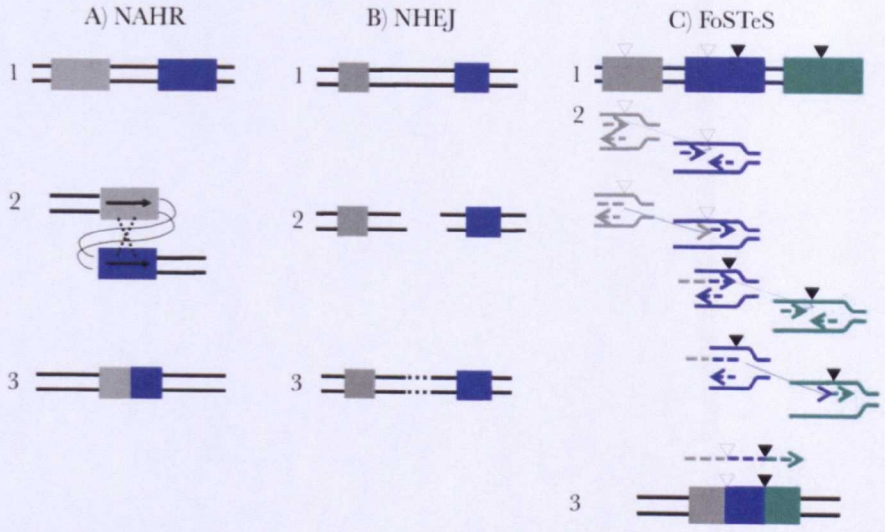


Figure 3: Comparison of three major mechanisms underlying human genomic rearrangements and CNV formation. A) Intrachromatid NAHR (non-allelic homologous recombination) event between two directly orientated low copy repeats (LCRs) sharing high homology (~97%) (1), represented by grey and blue rectangles. Alignment and recombination at non-allelic positions (2) leads to deletion or duplication of part of the LCRs and deletion of the segment between them (3). B) NHEJ (non-homologous end joining) event can occur between two sequences with no homology (1) (represented by grey and blue rectangles) when double stranded DNA breaks are created (2). The DSBs are repaired via NHEJ through a mechanism that includes modification of the ends to make them compatible for final ligation, resulting in the deletion or insertion of few bases at the DNA segment that separates the two sequences (3). C) FoSTeS (Fork staling and template switching) event only requires microhomology (2 to 5 base pairs) between the genomic fragments (1) (represented by grey, blue and green rectangles, with open triangles representing the microhomology sites between the grey and blue sequences and the filled triangles bearing sites for blue and green sequences). The replication forks of each sequence are shown in the same colour. The leading strand (2) (in the top grey replication fork) invades the right site replication fork (top blue replication fork), followed by DNA synthesis (dotted lines) using the new replication fork as a template. This event can happen several times, as illustrated, causing deletion of the two fragments flanked by each pair of microhomology sites (3). Juxtaposition of genomic sequences from multiple distinct regions leads to complex rearrangements. Figure adapted from Gu *et al.* (2008).

Scientific knowledge up to now indicates NAHR as the major mechanism responsible for the origin of large structural variation in the human genome (~48%) (Kidd *et al.* 2008), but conversely events mediated by NAHR are relatively rare (~14%) (Korbel *et al.* 2007). This discrepancy may be due to technical limitations of next-generation sequencing, which is a relatively poor technique for detecting structural variants within duplication regions of the genome (Kidd *et al.* 2008). Although NAHR has been described as the prevalent mechanism responsible for the formation of recurrent and some non-recurrent rearrangements, other non-recurrent rearrangements have been

proposed to occur due to NHEJ or by FoSTeS. So, the mechanisms responsible for copy number formation are still relatively unclear and further investigation is essential to understand the formation of *de novo* copy number mutations.

1.1.3.4 Methods to measure copy number variation

Copy number variation is an important and frequent type of structural variation in the human genome, which can contribute to phenotypic variation and even lead to disease (Aitman *et al.* 2006; McCarroll and Altshuler 2007; de Smith *et al.* 2008; Hollox *et al.* 2008b). Such characteristics increased the interest in human copy number variation, leading to the establishment of diverse and distinct quantitative methodologies to measure copy number. Such techniques allow the detection and cataloguing of human copy number variants but are also used to assess the association of copy number variants with biological function, recent human evolution and common and complex diseases (de Smith *et al.* 2008).

Until recently, SNPs were thought to account for the majority of genomic variability observed in the human genome but recent research showed that CNV are as common, although their discovery has been limited by the technology available, which was especially designed for SNP genotyping (McCarroll and Altshuler 2007). Accurate typing of copy number is difficult and more challenging than SNP genotyping, since it requires the ability to quantitatively distinguish different copies of the same repeat relative to the rest of the genome. As a consequence, copy number techniques should be precise and accurate to achieve confident and reliable measurements (McCarroll and Altshuler 2007). Without a high level of accuracy and precision, differences between cases and controls can be biased and result in false positive associations (Clayton *et al.* 2005).

Among the techniques that have been applied to measure gene copy number the more frequent typing methods are: Quantitative real-time PCR, Multiplex Ligation-dependent Probe Amplification (MLPA), array Comparative Genome Hybridization (array-CGH), SNP genotyping arrays and PRT (Paralogue ratio test). Apart from those, Multiplex Ligation dependent Genome Amplification (MLGA), Multiplex Amplifiable Probe Hybridization

(MAPH) and more recently, molecular combing have also been used to measure copy number but not extensively in association studies.

Quantitative real-time PCR describes the amplification and simultaneous quantification of a targeted DNA molecule during the early phases of the PCR reaction. It is the most popular method to measure copy number variation and has been used in several copy number variable regions including the β-defensin region (Linzmeier and Ganz 2005; Chen *et al.* 2006b; Fellermann *et al.* 2006; Bentley *et al.* 2010). For copy number determination, usually two pairs of primers are used to amplify the target locus and a single copy of a reference locus (for example the human serum albumin, ALB gene). Additionally in some studies, a calibrator plasmid, containing a copy of each of the reference and target gene, and a control plasmid, incorporating 2 copies of the target gene and one copy of the reference gene, are amplified simultaneously in order to calibrate the experiment. The detection and quantification of PCR products occurs during the exponential phase by fluorescence emission collected in a laser detector during the course of the reaction (Heid *et al.* 1996; Barrois *et al.* 2004; Chen *et al.* 2006b). Quantification can be achieved through two common methods that either use fluorescent dyes (e.g. SYBR Green Dye) that intercalates with the double-strand DNA (Dall'Ozzo *et al.* 2003; Linzmeier and Ganz 2005; Aitman *et al.* 2006; Fellermann *et al.* 2006) or modified DNA oligonucleotide probes (e.g. TaqMan Probe) that fluoresce after hybridization with complementary DNA (Heid *et al.* 1996; Bentley *et al.* 2010). The analysis of the results can be carried out using the standard curve method, comparative threshold cycle or by melt analysis.

Real-Time PCR has been used in different copy number variable regions, such as *BRCA1* (Barrois *et al.* 2004), *FCGR3* (Dall'Ozzo *et al.* 2003; Aitman *et al.* 2006) and *CCL3L1* (Gonzalez *et al.* 2005). At chromosome 8p23, Linzmeier and Ganz (2005) used quantitative real-time PCR to analyse the independent copy number variation of α- and β-defensin genes. They successfully addressed the question of independent variation between α- and β-defensin but just 24 samples were typed, giving only weak information about the reliability of this method to accurately measure copy number variation (Linzmeier and Ganz 2005). In a study developed by Chen *et al.* (Chen *et al.* 2006b), the copy

number polymorphisms in human β-defensin genes were screened using a modified quantitative real-time PCR. Again, the study did not include a large set of samples (44 samples) and did not compare the accuracy of the method for high copy numbers, such as 8, 9 and 10 but also, no analysis of raw data was shown (Chen *et al.* 2006b). Quantitative real-time PCR (TaqMan assay) was also used in association studies, to evaluate the correlation between β-defensin copy number and Crohn's disease (CD) (Fellermann *et al.* 2006; Bentley *et al.* 2010). Although these studies assessed the β-defensin copy number in a considerably larger cohort than previous studies, they do not report the unrounded copy number values and copy numbers were generally allocated into three main copy number bins, corresponding to lower than, equal to or higher than 4 copies. The accuracy of quantitative real-time PCR was not extensively tested in any of the studies above for high copy numbers, a crucial point in measuring the variable β-defensin cluster that can vary up to 12 copies. Nevertheless, there are some advantages in performing real-time PCR: the ability to quantify the PCR products during the exponential phase and the reduced probability of contamination when compared with other methodologies (Heid *et al.* 1996). The method itself relies on the use of plasmids containing one copy each of the target gene and the reference gene, as a control; however this does not exactly represent the real human genome. Instead, real reference samples should be applied to calibrate the experiment. Furthermore, using a multiplex PCR that uses a different pair of primers to amplify the test and reference locus will reduce the overall specificity and accuracy of the method.

In a separate group, there are the methods based on multiplex targeted copy-number analysis, such as Multiplex ligation-dependent probe amplification (MLPA) and Multiplex ligation dependent genome amplification (MLGA). MLPA involves the hybridization and ligation of two half-probes, followed by quantitative amplification of the ligated products, requiring just 20ng of genomic DNA (Schouten *et al.* 2002). MLPA has been largely applied to detect common deletions and duplications in genes with clinical importance, such as the genes for Duchenne muscular dystrophy (DMD), Becker muscular dystrophy (BMD) and the low-density lipoprotein (LDL) receptor (Holla *et al.*

2005; Janssen *et al.* 2005). Is also a rapid and effective assay for routine diagnostics of aneuploidies and trisomies (Slater *et al.* 2003; Herodez *et al.* 2005). Many studies corroborate their precision to detect common deletions and duplications, but these methods were never used to measure high copy number variants, as 6, 7, 8 and 9 copies of a gene or variable region.

MAPH (Multiplex amplifiable probe hybridization) is another method used in CNV measurement. This approach relies on sequence-specific probe hybridization to genomic DNA, followed by amplification of the hybridized probes and semi-quantitative analysis of the resulting PCR products (Armour *et al.* 2000). MAPH probes hybridize with genomic DNA fixed to a membrane, instead of genomic DNA free in solution, as occurs in MLPA and MLGA (Figure 4). The probes are created from any sequence, except highly copy number variants as *Alus*, CG-rich or GC-poor regions, by cloning target sequences into a plasmid vector, which are then amplified with specific vector primers. The membranes are intensively washed to remove the unbound probes, while the bound probes are separated from the membrane and amplified with universal primers. After separation by capillary electrophoresis, the PCR products are quantified in terms of peak area or height. In this method the amount of amplified product is directly proportional to the copy number in the genomic DNA (Armour *et al.* 2000; Armour *et al.* 2002; Hollox *et al.* 2002a; Hollox *et al.* 2002b). MAPH has been successfully applied to measure α- and β-defensin copy variable genes, which show a large range of variation up to 12 copies. In different studies, MAPH was combined with other methods to either clarify the genomic organization of defensin clusters or to characterize the copy number variation at such loci (Hollox *et al.* 2003; Aldred *et al.* 2005). When applied in association studies, the copy number reported from each probe was equivalent and the method showed comparable accuracy to other methodologies (Hollox *et al.* 2005; Hollox *et al.* 2008b). Additionally, MAPH has been used to detect *BRCA1* duplications in breast and ovarian cancer (Rad *et al.* 2001), subtelomeric abnormalities related with idiopathic mental retardation in children (Sismani *et al.* 2001) and a large deletion in *TBX5* gene that causes the Holt-Oram syndrome (Akrami *et al.* 2001).

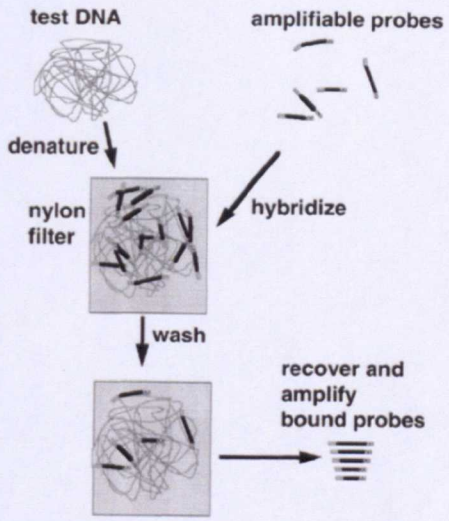


Figure 4: General principle of Multiplex Amplifiable Probe Hybridization (MAPH). Test DNA is denatured and hybridised with a set of amplifiable probes, each recognising a unique region in the genome. After intensive washing, the bound probes are retained in the membrane and amplified using a common primer pair, being quantified after capillary electrophoresis. Figure taken from Armour et al. (2000).

Furthermore, it was also used to screen for deletions in the Duchenne muscular dystrophy (DMD) gene (White *et al.* 2002) and to screen both deletions and duplications in *PMP22* gene, which can be involved in hereditary neuropathy with liability to pressure palsies (HNPP) and Charcot-Marie-Tooth disease type 1A (CMT1A), respectively (Akrami *et al.* 2003). MAPH is a reliable and reproducible method to measure and detect copy number variants, even high copy number states, at multiple loci (~110) simultaneously and has been used for screening cancer genes (Tyson *et al.* 2009; Tyson *et al.* 2010). Conversely, this assay is very time consuming, due to the careful construction of a set of probes and the labour involved in the washing step and separation of the bound probes from the membrane. To perform this assay about 1µg of genomic DNA is necessary, an amount that even if it is not prohibitive, is much higher than employed by other techniques nowadays, compromising its applicability in large scale (Armour *et al.* 2002; Hollox *et al.* 2002a).

Copy number variation can also be identified using DNA microarrays by Comparative Genome Hybridization, a method based on DNA arrays. Essentially, test and reference DNA are differentially labelled and hybridized together to the array previously spotted with specific DNA fragments (bacterial artificial chromosome - BAC, cDNA, PCR fragments or oligonucleotides). After hybridization, the resultant fluorescent ratio between the two DNA

samples is determined revealing genomic differences, such as copy number variants (Solinas-Toldo *et al.* 1997; Pinkel *et al.* 1998). Initially developed from fluorescent *in situ* hybridization (FISH) to compare the gene expression of multiple targets in a single experiment (Kallioniemi *et al.* 1996), array-CGH has become a common method to study genome-wide copy-number changes in the human genome (Sharp *et al.* 2005; Redon *et al.* 2006; Wong *et al.* 2007). With the growing interest in CNV, a number of studies have identified new variations using this technology. Studying normal human populations, Iafrate *et al.* (2004) identified 255 loci, among 55 individuals, containing genomic imbalances, while Sebat *et al.* (2004) found 76 copy number polymorphisms between 20 individuals. In a recent study by Barber *et al.* (2007), a duplication of 550 kb at 8p23.1 was revealed by array CGH, which has been associated with a characteristic facial phenotype including a prominent forehead and arched eyebrows. Array-CGH is a powerful method to screen the entire genome in a single experiment with complete coverage; however, in most implementations it has low resolution (down to ~50 kb for BAC array-CGH) and cannot define efficiently the rearrangement breakpoints (Sharp *et al.* 2006).

SNP genotyping arrays, originally developed for determining single base polymorphisms, can be used to measure genomic copy number by comparative hybridization intensities of genomic DNA to the oligonucleotide probes. For CNV detection, intensity ratios with respect to a reference genome are examined, allowing the determination of the relative number of copies per locus (Redon *et al.* 2006; Carter 2007; McCarroll *et al.* 2008). SNP platforms use probes that are specific to detect single base differences either by single-base extension methods (Illumina) or differential hybridization (Affymetrix) (Alkan *et al.* 2011). Both platforms show a good resolution, between 2 kb and 2.5 kb, but probes are not uniformly distributed across the genome with very poor representation in regions of segmental duplications and copy number variation (Carter 2007). Moreover, these platforms are mainly able to identify simple deletion and duplication polymorphisms, but are very poor in detecting high copy number variants. Most multiallelic copy number variants are not tagged by SNPs because these regions are only sparsely covered by unique SNPs, and the use of probes with SNPs compromise the use of CGH for

detecting and quantifying copy number. The ability of array-CGH to detect copy number changes depends greatly on signal-to-noise ratio of hybridization of probes. It is important, particularly for CNV detection, that measurement variance in hybridization methods, which lead to false positive and false negative results are accurately assed. However, cross-hybridization of probes to different loci frequently leads to incorrectly called regions as CNVs (Carter 2007).

Molecular combing (MC) is a method for single DNA molecule analysis which allows the direct visualization and size measurement of the relevant loci on numerous individual molecules at 1 kb resolution (Lebofsky and Bensimon 2003). Essentially, an array of combed single DNA molecules is prepared by stretching molecules attached by their extremities to a silanised glass surface with a receding air–water meniscus. Followed by fluorescent hybridisation on combed DNA, genomic probe position can be directly visualised, making possible the construction of physical maps and detection of micro-rearrangements (Lebofsky and Bensimon 2003). This method has not been extensively used in copy number measurement but has proved very useful to investigate the allelic combinations associated with Facioscapulohumeral dystrophy disease (FSHD) (Nguyen *et al.* 2011). FSHD is associated with a reduction in the copy number of the *D4Z4* repeat array at subtelomeric region of 4q35 and a particular genomic organization different from its chromosomal counterpart at 10q26. Accurate discrimination and description of the 4qter and 10qter *D4Z4* repeat arrays were detectable by MC, which is extremely important in FSHD diagnosis (Nguyen *et al.* 2011). MC can also be applied to other copy-number-variants or repeat expansion arrays associated with human diseases.

Paralogue ratio test is a newly developed method for copy number measurement. PRT is a comparative PCR-based method, where the same primer pair amplifies test and reference loci (Deutsch *et al.* 2004; Armour *et al.* 2007). This important technical detail, which clearly distinguishes PRT from multiplex PCR-based assays (e.g. real-time PCR) is the amplification of test and reference loci at very similar kinetic properties using a single pair of primers. This reduces the reproducible differences commonly observed in multiplex PCR, when test and reference loci are amplified by 2 pairs of primers. Although PRT is limited by the precise design of primer sequences that

exclusively match test and reference loci, many copy number variable regions have been shown to have these characteristics (Armour *et al.* 2007). Moreover, the precision and accuracy of PRT are equivalent to MLPA and MAPH, with the advantage of requiring smaller amounts of genomic DNA (10-20ng) (Armour *et al.* 2007). For these reasons, PRT allows high-throughput analysis of copy number variation and has been successfully applied to measure different CNVs, such as the β-defensin genes (Armour *et al.* 2007; Hollox *et al.* 2008b; Aldhous *et al.* 2010), α-defensins genes (*DEFA1/DEFA3*)(Fayeza Khan, personal communication) both at 8p23.1, *CCL3L1/CCL4L1* copy number at 17q12 (Walker *et al.* 2009) and the immunoglobulin-receptor genes *FCGR3A* and *FCGR3B* on chromosome 1q23.3 (Hollox *et al.* 2009).

The knowledge of copy number variants and their importance in the risk of common human diseases led to the development of several technologies to measure copy number of variable regions. Analytical methods that use either global or targeted approaches have been used to address the copy number, but many of them show characteristics that either compromise the accuracy or the applicability on large scale, due to the amount of genomic DNA required or the cost involved. The β-defensin cluster is an interesting copy number variable repeat region, showing high copy number states, with a complex genomic architecture (Hollox *et al.* 2003). Many of the methods described do not appear to be accurate for measuring such high copy numbers. On the other hand, the methods with improved accuracy at this level are very difficult to perform on a large-scale or the cost can compromise their applicability. For these reasons, the development of an improved technology to measure the copy number, at this or other loci highly polymorphic in gene number, will be essential in this field.

1.2 DEFENSINS

The innate immune response in plants and animals is mediated by "natural antibiotic peptides". These peptides are present among eukaryotes, including mammals, birds, amphibians, insects, plants and protozoa (Gabay 1994; Ganz 2003). In mammals, these peptide antibiotics form a large family of small cationic antimicrobial peptides, with no more than 50 amino acids, generally known as defensins. The defensins have an important role in innate immunity, acting against microbial invasion of several pathogens. On the other hand by signalling to chemokine receptors on dendritic cells and T-cells, they can also contribute to the regulation of host adaptive immunity (Ganz 1999a; Yang *et al.* 1999; Yang *et al.* 2002). Defensins are produced by cells and tissues that are involved in host defence against microbial infection, mainly leukocytes and mucosal epithelial cells, which are active against a range of bacteria, including Gram-positive and Gram-negative species, virus and fungal pathogens (Ganz 2003; Klotman and Chang 2006).

The defensins are small peptides of around 3.5-6 kDa with characteristic β-sheet structures. Six cysteine residues stabilized by three intramolecular disulphide bonds characterize their structure. According to the spatial arrangement of the disulfide bridges between the cysteine residues, the mammalian defensin family is divided into three subfamilies, the alpha, beta and theta-defensins. In α-defensins the six cysteine residues are linked in the 1-6, 2-4 and 3-5 pattern (Cys1-Cys6, Cys2-Cys4 and Cys3-Cys5) while in β-defensins, the linkages are between the 1-5, 2-4 and 3-6 (Cys1-Cys6, Cys2-Cys4 and Cys3-Cys5) residues. In the θ-defensins, due to its circular structure, the disulphide pairing has a distinct configuration from the α- and β-defensins, with a 1-6, 2-5 and 3-4 (Cys1-Cys6, Cys2-Cys5 and Cys3-Cys4) pattern (Tang *et al.* 1999; Ganz 2003; Klotman and Chang 2006).

In humans, *DEFA* and *DEFB* genes, mainly located in a cluster on chromosome 8p23.1 locus, encode the α- and β-defensins respectively. Additionally, they can also be found in a cluster on chromosome 6 (6p12) and in two clusters on chromosome 20; 20p13 and 20q11.1 (Linzmeier et al. 1999; Ganz 2003; Taudien et al. 2004). The θ-defensin genes (DEFT) are only present

in humans as non-functional copies (pseudogenes), which have been identified on chromosome 8p23 and chromosome 1 (Nguyen et al. 2003; Carsten Münk 2004).

1.2.1 Human defensins: alpha, beta and theta

Five different genes encode the α-defensins, *DEFA1* to *DEFA6*. Alpha-defensins encoded by *DEFA1* to *DEFA4* genes are constitutively contained in neutrophil granules (Ganz *et al.* 1985; Linzmeier *et al.* 1993), while *DEFA5* and *DEFA6* are expressed in Paneth cells of the small intestine (Ghosh *et al.* 2002), (Table 1). The majority of these genes are located in a cluster at the telomeric end of chromosome 8p23.1 (Aldred *et al.* 2005). *DEFA1* and *DEFA3* encode different peptides, human neutrophil-derived alpha-defensin 1 (HNP-1) and 3 (HNP-3), respectively, which differ just by the first (N-terminal) amino acid of the mature peptide due to a single-nucleotide polymorphism, C3400A (Ganz and Lehrer 1995). This polymorphism is human specific suggesting a recent origin to the human lineage (Aldred *et al.* 2005). The HNP-2 peptide is highly homologous to HNP-1 and -3, missing only the first residue, but no gene has been identified for this peptide. It has been suggested that HNP-2 is a proteolytic product of one or both of DEFA1 and DEFA3 peptides (Ganz 2003; Linzmeier and Ganz 2005).

Alpha-defensins are synthesized as prepropeptides of 90-100 amino-acid precursor sequences composed of an amino (N)-terminal signal sequence (~19 amino acids), an anionic propiece (~45 amino acids) and a carboxy (C)-terminal mature cationic defensin (~30 amino acids), after which translational proteolytic cleavage produces the mature peptide, including the HNP-2 (Harwig *et al.* 1992). Human α-defensin-1, -2, -3 and -4 are synthesized by neutrophil precursor cells, promyelocytes, in the bone marrow and stored in neutrophil granules before mature neutrophils are released into the blood, where they may migrate to sites of inflammation (Ganz 2003). In contrast, for human α-defensin-5 (HD5) the maturation process takes place outside the cell (Ghosh *et al.* 2002).

DEFA1 (HNP-1) and *DEFA2* (HNP-2) show a strong antimicrobial activity against *Candida albicans* (Chertov *et al.*, 1996), as well as Gram-negative bacteria, *Enterobacter aerogenes* and *Escherichia coli,* and Gram-positive bacteria, *Staphylococcus aureus* and *Bacillus cereus* (Ericksen *et al.*, 2005).

Additionally, *DEFA1* (HNP-1) was also described to have a strong anti-HIV-1 activity with a direct effect on the virus (Chang *et al.*, 2004). Apart from humans, α-defensins have been found in different mammals, including non-human primates, Rhesus macaque and some rodent species, in which they participate in similar functions to those described for humans (Ganz 2003).

Table 1: Summary of defensins present in the Human genome.

Human defensins	Chromosomal location	Gene symbol	Expression location	Function
α-defensin	8p23.1	DEFA1 and 3	Neutrophils granules	Antimicrobial activity against Gram-positive and Gram-negative bacteria
		DEFA4, 5 and 6	Paneth cells	
β-defensin	8p22-p23	DEFB1, 4, 103-107, 109, 130, 135-137 and SPAG11	Mainly in epithelial tissues for eg.: keratinocytes; respiratory and gastrointestinal tract; urogenital and respiratory system; amniotic fluid	Antimicrobial activity against Gram-positive and Gram-negative bacteria, cytokine-like properties and pro-inflammatory activity (particularly DEFB103)
	6p12.3	DEFB110, 112-114 and 134		
	20q11.21	DEFB115,116,118-124		
	20p13	DEFB125-129,132		
θ-defensin	8p23.1	Pseudogenes: DEFT1P,1P1 and 1P2	Bone marrow, spleen, thymus, testis and skeletal muscle	Unknown: inactivated peptides due to a premature stop codon
	1q23.1	Pseudogene: DEDD		

Human β-defensins are encoded by *DEFB* genes, mainly located on chromosome 8p23-p22, that include thirteen different β-defensin genes, *DEFB1*, *DEFB103-107, DEFB109, DEFB4, DEFB130, 135-137* and *SPAG11* (*HE2/EP2*). Seven of these genes (*DEFB103-107*, *DEFB4* and *SPAG11*) are clustered in a repeat unit at chromosome 8p23.1, which is highly polymorphic in copy number (Hollox *et al.* 2003; Groth *et al.* 2008; Abu Bakar *et al.* 2009). Three other β-defensin gene clusters have been identified by *in silico* analysis, one at chromosome 6p12.3 (*DEFB110, 112-114 and DEFB134*) and two at chromosome 20, 20q11.21 (*DEFB115, 116, 118-124*) and 20p13 (*DEFB125-129, 132*) but they

do not appear to exhibit copy number variation (Schutte *et al.* 2002; Rodriguez-Jimenez *et al.* 2003; Hollox and Armour 2008), (Table 1). To date, more than 30 β-defensin genes have been identified that are mainly expressed in epithelial tissues, including the skin and respiratory tract (Schutte and McCray 2002; Schutte *et al.* 2002).

The genomic structure of the majority of β-defensin genes is composed of two exons separated by one intron of variable length, except for the *DEFB105* gene, which has an extra exon and intron. The first exon of β-defensin genes includes the 5' untranslated region and encodes the signal and pro-sequence, whereas the second exon encodes the mature six-cysteine defensin peptide, which normally has between 38 and 42 amino acids (White *et al.* 1995; Ganz 2003; Lehrer 2004). For *DEFB1*, the first exon encodes both the pre-peptide and the mature peptide (Liu *et al.* 1997).

Tracheal antimicrobial peptide (TAP) was the first member of the β-defensin family to be described in mammals by Diamond *et al.* (1991). This peptide was isolated from the bovine tracheal mucosa and shows a broad-spectrum of antimicrobial activity similar to other β-defensins, namely the human β-defensin 2 (human homologue) (Diamond *et al.* 1991). In humans, hBD-1 (*DEFB1* gene) was the first β-defensin to be identified and isolated from human plasma (Bensch *et al.* 1995). *DEFB1* shows high homology with TAP and is predominantly expressed in leukocytes and epithelial cells from the respiratory tract (Bensch *et al.* 1995; Fulton *et al.* 1997; Ali *et al.* 2001). Apart from its antimicrobial activity, *DEFB1* is a potential tumor suppressor gene for urological cancers, inducing apoptosis of renal cancer cells (Sun *et al.* 2006). A second human β-defensin, hBD-2, encoded by *DEFB4*, was initially isolated from psoriatic lesions. hBD-2 is expressed in skin, mainly in keratinocytes, and epithelia of the airway system, such as respiratory tract, gastrointestinal tract and urogenital system, where it contributes to its antimicrobial defence (Harder *et al.* 1997b). Nevertheless, hBD-2has been identified in other sites, such as in the amniotic fluid, which accounts for the host defence against microorganisms within the amniotic cavity (Soto *et al.* 2007). Another antimicrobial peptide, human β-defensin-3, hBD-3 (*DEFB103*), was also isolated from human psoriatic scales and cloned from keratinocytes. hBD-3 is expressed in different tissues

including skin, tonsils and lung (Harder *et al.* 2001). hBD-2 antimicrobial activity is especially directed against Gram-negative bacteria such as *Escherichia coli* and *Pseudomonas aeruginosa* (Harder *et al.* 1997a) but other β-defensins, as hBD-3, can also show strong antimicrobial activity for Gram-positive bacteria, such as *Staphylococcus aureus* (Harder *et al.* 2001).

hBD-2 and hBD-3 also have cytokine-like properties. Both defensins were described to be specific human chemoattractants for cytokines mediating the recruitment of neutrophils to sites of inflammation and infection. This finding reveals their capacity to mediate the innate and adaptive immune responses (Harder *et al.* 2001; Niyonsaba *et al.* 2004). Recent evidence from Niyonsaba *et al.* (2007) enhances the functional role of β-defensins, hBD-2, -3 and -4, but not hBD-1, in skin immunity. These peptides promote keratinocyte migration and proliferation and stimulate cytokine/chemokine production (Niyonsaba *et al.* 2007). In addition to their pro-inflammatory action, human β-defensin 3 (hBD3) was recently described to have anti-inflammatory properties (Semple *et al.* 2010). hBD-3 can effectively inhibit the accumulation of pro-inflammatory cytokines TNF-α and IL-6 after stimulation by lipopolysaccharide (LPS). The novel functions of hBD-3 suggests a role in the resolution of inflammation and in tissue repair damage caused by effectors of antimicrobial action (Semple *et al.* 2010).

In most mammals coat colour patterns are controlled by the melanocortin system, a seven transmenbrane-domain receptor and its extracellular ligand, encoded respectively by the *melanocortin 1 receptor* (*Mc1r*) and *Agouti* genes. This process, commonly known as pigment "type-switching", controls the production of two types of pigments: yellow (eumelanin) and black (phaeomelanin) (Ollmann *et al.* 1998). However, in dogs the production of these pigments is also controlled by *CBD103*, an orthologue of human *DEFB103*. A study undertaken in domestic dogs reported that a mutation in *CBD103*, controls pigmentation through the melanocortin 1 receptor and correlates with black coat colour in 38 different breeds. In the presence of the wild-type alleles for all three genes the Agouti binds to Mc1r resulting in yellow coating. However, dogs carrying the dominant black allele of *CBD103*, the respective peptide binds with high affinity to the melanocortin 1 receptor (Mc1r) competing with Agouti and affecting the pigment type-switching

mechanism in domestic dogs and transgenic mice (Candille *et al.* 2007). This discovery associated a new function to the β-defensins in mammals and suggests a similar new role for β-defensins, not yet investigated, in humans.

Three epididymis-specific β-defensins, hBD-4, hBD-5 and hBD-6 (encoded by *DEFB104*, *105* and *106* genes) were identified by Basic Local Alignment Search and annotated at the 8p23.1 locus (Yamaguchi *et al.* 2002). These peptides are specifically expressed in human epididymis, which is highly vulnerable to microbial invasion, thus suggesting the role of β-defensins in host defence against bacterial pathogens, protecting the spermatozoa (Yamaguchi *et al.* 2002). In addition to these β-defensins, *SPAG11* located within the 8p23.1 copy variable repeat unit, also encodes a epididymis-specific secretory peptide that was described as a conserved component of the innate epididymal epithelial defence system in primates (Horsten *et al.* 2004).

Although the β-defensins were firstly known for their antimicrobial activity against a range of microorganisms, several studies have demonstrated their additional functional activity, including a role in the adaptive immune system as signalling molecules. As such, they are referred as a multifunctional gene family.

The θ-defensin subfamily is the most understudied and so far, only six θ-defensin (*DEFT*) genes have been identified in the human genome; five on chromosome 8p23 and one on chromosome 1 (Table 1). They appear to be expressed pseudogenes that encode antimicrobial peptides, also called retrocyclin (Nguyen *et al.* 2003; Carsten Münk 2004). Although θ-defensin mRNA transcripts have been found in human bone marrow, spleen, thymus, testis and skeletal muscle they are inactivated due to a premature stop codon that disables their translation (Cole *et al.* 2002).

Homologues of the human theta-defensin genes have been found in chimpanzees and gorillas, containing the same premature stop codon, whereas active genes were found in several Old World monkeys and orangutans (Nguyen *et al.* 2003). Human theta-defensin peptides are homologous to rhesus macaque (*Macaca mulatta*) circular minidefensins. These are termed rhesus theta-defensin-1 (RTD-1), RTD-2 and RTD-3 and were isolated from granules of neutrophils and monocytes (Tang *et al.* 1999; Cole *et al.* 2002). The

biosynthesis of the mature peptides involves the fusion of two alpha-defensin-related nonapeptides (van der Maarel and Frants 2005) followed by cyclization through the formation of two new peptide bonds (Tang *et al.* 1999). This result in a cyclic octadecapeptide molecule stabilized by three disulphides bonds, the only cyclic peptide known in animals (Selsted and Ouellette 2005). It has been suggested that *DEFT* genes and theta-defensins arose by mutation of a pre-existing α-defensin gene in Old World monkeys. The inactivation of theta-defensins in humans occurred after the orangutan and hominid lineage divergence (Nguyen *et al.* 2003). All defensin peptides share similar antimicrobial properties but for the theta-defensins, retrocyclin, Cole *et al.* (2002) suggested an additional strong anti-HIV-1 infection activity. *In vitro*, retrocyclin protects human $CD4^+$ cells from infection by T-tropic and M-tropic strains of HIV-1.

1.2.2 Defensins and immunity: mode of action

In higher vertebrates, such as mammals, the immune system is composed by two types of immunity; innate and adaptive immunity. Whilst innate immunity is activated immediately after microbial invasion providing a rapid and direct protection against foreign agents, the adaptive immune response mediated by T and B cells can take several days to unfold. Important components of the innate immune system are the antimicrobial substances, which include the defensins.

The defensins have a broad spectrum of antimicrobial activity *in vitro* against Gram-positive and Gram-negative bacteria, yeast, fungi and enveloped virus and are important players in innate immunity (Ganz and Lehrer 1995; Ganz 2003). The activity of α-, β- and θ-defensins becomes effective at concentrations of 0.5-5μM and for most α- and β-defensins their activity is regulated by physiological concentrations of NaCl, divalent cations and serum components (Selsted and Ouellette 2005). At high ionic concentrations of salt (NaCl) these defensins have reduced antimicrobial activity. Therefore, the direct antimicrobial activity of α- and β-defensins *in vivo* is likely to occur in the phagocytes and epithelial tissues, where the salt concentrations are low (Ganz and Lehrer 1998; Lehrer and Ganz 1999; Schroder 1999). Conversely, these factors seem not to influence the activity of θ-defensins, which have been

suggested to correlate with its cyclic conformation (Tang *et al.* 1999; Tran *et al.* 2002). The antiviral activity of defensins is not influenced by salt concentration or the absence of serum components (Klotman and Chang 2006). In the particular case of viruses the mechanism of antiviral activity by defensins can occur in the absence of serum with the inactivation of enveloped virus particles by disruption of viral envelope or by interaction with viral glycoproteins (Klotman and Chang 2006). For some β-defensins, such as human β-defensin 1, the antimicrobial activity is regulated by redox conditions. hBD-1 showed potent antimicrobial activity under reducing conditions that characterize anaerobic environments. After reduction of disulphide-bridges, hBD-1 becomes an effective antimicrobial peptide against Gram-positive *Bifidobacterium* and *Lactobacillus* species and fungus *Candida albicans* (Schroeder *et al.* 2011). Anaerobic environments can be found in mucosal membranes of the intestine, vagina and oral cavity as well as in cutaneous sweat, sebaceous glands and infectious sites where hBD-1 is constitutively produced and can provide an effective barrier against colonisation by anaerobic pathogens and commensals (Schroeder *et al.* 2011).

Although defensin antimicrobial mechanisms are not completely understood, it is agreed that it may involve the disruption of structural elements of the microbial cell membrane. The main mechanism of action is membrane depolarization and permeabilization, which is dependent on differences in membrane composition of both host and microorganism, followed by disruption of some physiological processes (Yang *et al.* 2002; Selsted and Ouellette 2005). In contrast to host cells, most of the microorganism's membranes lack cholesterol and are composed of negatively charged phospholipids. Defensin peptides are cationic and amphiphilic, allowing them to insert within the phospholipids and interact with the negatively charged (anionic) microbial membrane (Lehrer and Ganz 1999; Kamysz *et al.* 2003; Selsted and Ouellette 2005). Subsequently, some defensins can aggregate and form ion channel pores or cover the microbial membrane, forming a carpet-like arrangement which leads to the permeabilization and disruption of membrane integrity and function resulting in the lysis of the microorganism (Schroder 1999; Hoover *et al.* 2000). At high concentrations (15-30μM),

defensins can have cytotoxic activity against tumor cell growth *in vitro* (Lichtenstein *et al.* 1986). Moreover, it was suggested that the differences observed in membrane composition of tumor cells that are rich in phosphatidylserine, makes them more susceptible to membrane permeabilization by defensin peptides. Defensin peptides are expressed in several tumor cell lines and depending on concentrations they can exert a necrotic or mitogenic activity (Kamysz *et al.* 2003).

The defensins have also been linked to the adaptive immune system. The first findings focused on the role of defensin in innate immunity were in alpha-defensins (HNP-1 and HNP-2), which were shown to be chemotactic for human monocytes and naive T-cells (Territo *et al.* 1989; Chertov *et al.* 1996; Yang *et al.* 2000). Later, β-defensin hBD-3 and hBD-4 were also described to have chemoattractant properties for monocytes, while hBD-2 mediates the recruitment of memory T-cells and immature dendritic cells via human chemokine receptor 6 (CCR6) (Yang *et al.* 1999; Conejo García *et al.* 2001; Garcia *et al.* 2001). These studies seem to indicate that CCR6 is a receptor for both CC-chemokine ligand 20 (CCL20 or chemokine macrophage inflammatory protein 3α) and β-defensins (Yang *et al.* 1999). In addition, HNP-1, hBD-3, and to a lower extent hBD-1, are chemotactic for immature dendritic cells, with hBD-3 able to recruit macrophages as well (Yang *et al.* 1999; Yang *et al.* 2000; Garcia *et al.* 2001). As just described, defensins can share the same receptors with some chemokines to mediate the recruitment of immune cells for sites of infection, suggesting that defensin and chemokines can have similar functions. Other findings showed that some chemokines have defensin-like antimicrobial activities, despite their distinct amino acid structure (Cole *et al.* 2001). Moreover, defensins stimulate the production of several cytokines, soluble proteins of low molecular weight produced by all immune cells, which regulate the innate and adaptive immune responses. Defensins were found to stimulate the production of TNF-α and IL-1 in monocytes and IL-8 in lung epithelial cells (Van Wetering *et al.* 1997; Chaly *et al.* 2000). In immature dendritic cells the production of IL-2 can be induced by HNP-1, hBD-2 and hBD-3 (Yang *et al.* 2002).

The first evidence that defensins could have mitogenic properties was given by Murphy *et al.* (1993), who showed that, at concentrations that would

allow antimicrobial activity, defensins could also stimulate fibroblast and epithelial cell division *in vitro*. The same range of defensin concentrations is expected *in vivo* during wound healing, suggesting a dual role of defensins in cell growth and wound healing. Such ideas were later confirmed by Aarbiou *et al.* (2004) who showed the same role of defensins in lung epithelial cells. Human neutrophil peptides (HNP) (α-defensins) can also regulate angiogenesis by affecting endothelial cell adhesion, migration, proliferation and differentiation (Chavakis *et al.* 2004). These studies provide new insight into the role of defensins in the regulation of inflammation by stimulating the restoration of epithelial tissue and neovascularisation, a critical part of inflammation and wound repair after a microbial invasion.

Defensins are chemoattractant for monocytes and macrophages, which are phagocytes involved in the direct immune response of the organism against invading foreign particles, such as bacteria. Thus, defensins also indirectly enhance innate immunity and phagocytosis through the recruitment of cells from the adaptive immune system (Yang *et al.* 2002). Given the capacity of defensins to enhance phagocytosis, promote the recruitment of several adaptive immune cells and stimulate the production of cytokines, they have been described as regulators/modulators of the innate host immune defence and as adjuvant of the adaptive immune response against microorganism invasion. The rapid expression of antimicrobial defensins and the recruitment of adaptive immune cells with a prolonged cellular and humoral response to a potential pathogen show the dual role of defensins in immunity linking the two types of immune response, innate and adaptive (Yang *et al.* 2002; Selsted and Ouellette 2005). Defensins have a concentration-dependent activity that at lower concentrations stimulates cytokine production or become chemoattractant for immune cells, while at higher concentrations are involved in microbial and tumor cell lysis (Kamysz *et al.* 2003).

1.2.3 Copy number variation of human defensin

Copy number variation of defensin genes was first reported in somatic cell hybrids mapping chromosome 8, for genes encoding the human neutrophil defensins HP-1 and HP-3 (*DEFA1* and *DEFA3*), which comprised between two

and three copy repeats (Mars *et al.* 1995). Copy number variation of α-defensins was later confirmed by Aldred *et al.* (2005) and Linzmeier and Ganz (2005) but the extension of copy number variation proved to be more complex than initially reported. *DEFA1* and *DEFA3* are located in a multi-copy array at 8p23.1, showing variation in both number and position between normal individuals (Aldred *et al.* 2005; Linzmeier and Ganz 2005). The α-defensin repeats appear in arrays of 19 kb, in a "Full repeat", containing one α-defensin gene and the DEFT1 pseudogene, and in a 9.5 kb "partial repeat" with only one α-defensin gene. The interchangeable position of *DEFA1* and *DEFA3* in the repeat unit suggest that both genes are located at the same locus, subsequently renamed *DEFA1A3* (Aldred *et al.* 2005). The diploid copy number varies between 3 and 11 copies in Europeans but in Africans and Asians the copy number can go up to 17 copies (Fayeza Khan, personal communication). Other α-defensin genes, *DEFA4*, *DEFA5* and *DEFA6*, are not commonly copy number variable (Aldred *et al.* 2005). The *DEFA3* gene is absent in about 10% of the UK population and in 37% of the sub-Saharan African population (Yoruba), although the effects that this common deletion might have in immunity are not yet known (Ballana *et al.* 2007). Analysis of expression levels in neutrophils indicated that HNP-1, HNP-2 and HNP-3 peptide levels are proportional to *DEFA1* and *DEFA3* copy number (Linzmeier and Ganz 2005). However, although Aldred *et al.* (2005) also suggested a correlation between the relative proportions of *DEFA1*:*DEFA3* mRNA and gene copy number, the combined mRNA levels of *DEFA1*+*DEFA3* were not proportional to the total gene copy number.

The presence of multiple copies of human β-defensins at chromosome band 8p23.1 was first reported as a cytogenetic duplication of the short arm of chromosome 8 by FISH (Barber *et al.* 1998). The β-defensin gene cluster at 8p23.1 includes six defensin genes, *DEFB103*, *DEFB104*, *DEFB105*, *DEFB106*, *DEFB107* and *DEFB4* and a defensin-like gene, *SPAG11*, in a repeat unit, which is highly polymorphic in copy number (Figure 5) (Hollox *et al.* 2003; Groth *et al.* 2008; Abu Bakar *et al.* 2009). The size of the repeat unit is unknown, but from pulse-field gel analysis it was demonstrated to span at least 260 kb (Hollox *et al.* 2003). Moreover, the β-defensin genes also vary

concordantly in copy number as a block (tandem repeat) with no internal duplications or deletions (Groth *et al.* 2008). Diploid copy numbers commonly range between 2 and 7 copies, but copy number up to 12 was found in some individuals (Hollox *et al.* 2003; Abu Bakar *et al.* 2009). Higher copy numbers between 9 and 12 can be visualized as euchromatic variants by cytogenetic analysis (Barber *et al.* 2005). Information about the haploid copy number states was recently investigated in CEPH reference pedigrees through segregation analysis. The study found between one and five copies of β-defensin genes per haploid genome (Abu Bakar *et al.* 2009). Null alleles are very rare, but they can be found at a very low frequency, about 0.2%. Functional consequences associated with the null homozygote might be strongly deleterious and lead to its elimination from the population by purifying selection (Hollox *et al.* 2008a). Analysis of mRNA transcript levels in lymphoblastoid cells by semi-quantitative RT-PCR showed a positive correlation with the genomic copy number of *DEFB4* (Hollox *et al.* 2003).

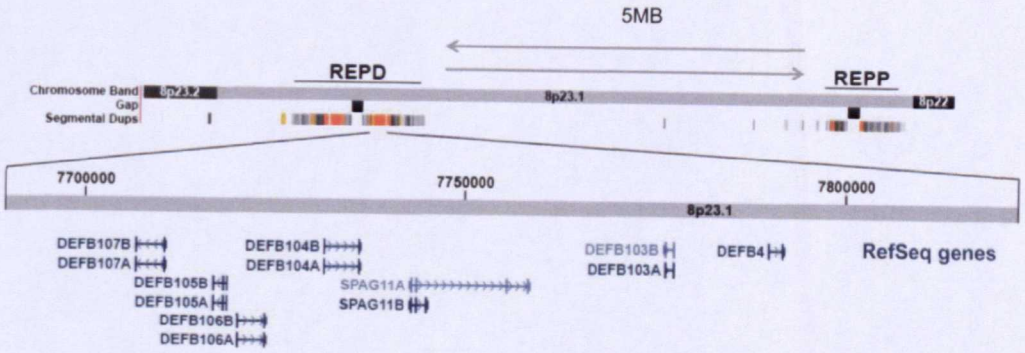


Figure 5: Genomic organization of the β-defensin locus at 8p23.1. On the top panel is represented the whole 8p23.1 locus with the REPD and REPP loci. Segmental duplications and gaps are represented in each locus. The grey arrows represent inverted and non-inverted orientation of the region that spans between REPD and REPP. On the bottom panel is highlighted the genome organization of each repeat with the seven genes that compose the β-defensin copy number variation repeat. (Figure adapted from UCSC Genome Browser on Human March 2006 (NCBI36/hg18) assembly).

1.2.4 Defensins in human disease

Defensins constitute a gene family with multiple functions; however, their role in immunity has highlighted defensins as candidate genes for inflammatory and autoimmune human diseases.

Several studies have been focused on the consequences of genetic variants and copy number variation of defensin genes in disease susceptibility. Defensins have attracted attention in cancer research since a link has been established between inflammation and cancer development (Karin *et al.* 2006; Lu *et al.* 2006). The presence of tissue damage caused by microbial infection, chemical irritation or wounding, initiates an inflammatory process which involves the recruitment of a wide variety of immune cells (neutrophils, mast cells, macrophages, leukocytes and lymphocytes) to the site of infection, as well as the release of various proinflammatory cytokines, chemokines and growth factors. These cells will contribute to antibacterial tissue breakdown and are important to maintain the defence against infection (Coussens and Werb 2002; Philip *et al.* 2004). The regulation of inflammation by apoptosis and phagocytosis of inflammatory cells is mediated by anti-inflammatory molecules. However, dysregulation of these processes lead to chronic inflammation, which will maintain the activation of immune cells and the recruitment of proinflammatory cytokines, chemokines and growth factors. These are responsible for the persistent tissue damage, cell proliferation and differentiation and release of reactive oxygen and nitrogen species that may cause DNA damage, thus predisposing to neoplasia (Lu *et al.* 2006). The development of cancer from inflammation may be mediated by inflammatory cells and regulators, including cytokines and chemokines that facilitate angiogenesis and promote cell growth, invasion and metastasis of tumor cells (Lu *et al.* 2006). In prostate cancer, the presence of persistent bacterial colonization in epithelial cells of prostate gland has been suggested to induce carcinogenesis through the associated inflammatory response (De Marzo 2007). Antimicrobial peptides such as the β-defensins have been suggested to play a role in the development and progression of this and other types of cancer (Schullerus *et al.* 1999; Donald *et al.* 2003; Zheng *et al.* 2004; Huse *et al.* 2008). However, no evidence found so far, supports this hypothesis. Although the impact of β-defensins in carcinogenesis is not yet clear, they may play a role in the pathogenesis of several cancers either by acting directly as a barrier against pathogens or/and by activating inflammatory response.

Chronic inflammation is also typically observed in chronic obstructive pulmonary disease (COPD). COPD is a progressive inflammatory disease of the airways characterized by narrowed airways leading to limitation of the airflow to and from the lungs. COPD is caused by toxic particles or gas, most frequently from tobacco smoking, which triggers an abnormal inflammatory response in the lung (Rabe *et al.* 2007). A recent finding shows the association between higher genomic copy number variation of β-defensin, especially *DEFB4* and COPD, revealing the potential functional role of β-defensins in airway epithelial cells and triggering inflammation. As chemoattractant peptides, the increased copy number of β-defensin may enhance the chronic inflammatory cascade contributing to the pathogenesis of COPD (Janssens *et al.* 2010).

Bacterial infection can also trigger the pathogenesis of other inflammatory diseases, such as cystic fibrosis. Chronic bacterial colonization of the lungs, especially by *Pseudomonas aeruginosa*, is a major cause of morbidity in patients with cystic fibrosis. Given the expression of the human β-defensin 2 peptide (encoded by *DEFB4*) in the lung epithelia and its strong antimicrobial activity against *Pseudomonas aeruginosa*, it has been suggested that *DEFB4* can be a potential candidate modifier gene for cystic fibrosis (Schutte and McCray 2002). However no association was found that would support the hypothesis of β-defensin copy number as a modifier for CF (Hollox *et al.* 2005).

More recently, using quantitative real-time PCR approaches, two studies found association between β-defensin copy number variation and Crohn's disease (Fellermann *et al.* 2006; Bentley *et al.* 2010). Crohn's disease is a chronic inflammatory bowel disease mediated by T lymphocytes and characterized by intestinal ulceration affecting mainly the ileum and colon. The aetiology remains unclear but genetic factors, host immune response, environmental factors, nutrition and enteric microflora all contribute to Crohn's disease susceptibility. However, it is suggested that Crohn's disease "arises in genetically susceptible individuals as a result of a breakdown in the regulatory constraints on mucosal immune responses to enteric bacteria" (Shanahan 2002). Given the functional role of β-defensins in immune system, as antimicrobials and cytokines (Ganz 2003; Niyonsaba *et al.* 2004), it is possible that they could play a role in Crohn's disease susceptibility. In 2006,

Fellermann *et al.* (2006) reported an association between low β-defensin copy number and Crohn's disease of the colon, suggesting that a reduced antimicrobial barrier in the gut could trigger the development of Crohn's disease. Conversely, high β-defensin copy number was also associated with Crohn's disease by Bentley *et al.* (2010), supporting the idea that increased dosage of *DEFB4* gene leads to chronic inflammatory response that has also been described as a possible factor causing Crohn's disease. However, neither of the associations previously reported have been replicated in subsequent studies carried out using a PRT-based assay to assess the β-defensin copy number in more than 1500 samples (Aldhous *et al.* 2010), or oligonucleotide array-CGH in more than 2000 cases and 3000 controls (WTCCC 2009).

Defensins were first reported in 1993 to have antiviral activity against HIV-1, inhibiting its replication *in vitro* (Nakashima *et al.* 1993). Theta-defensins are not actively present in humans, but synthetic putative ancestral human peptide, retrocyclin, has the ability to inhibit proviral DNA formation and protects $CD4^+$ lymphocytes cells from *in vitro* infection by HIV-1 (Cole *et al.* 2002). On the other hand, alpha-defensins can exhibit anti-HIV activity by directly inactivating virus particles and by targeting $CD4^+$ T cells and thus affecting the virus replication (Mackewicz *et al.* 2003; Chang *et al.* 2005). β-defensins, in particular hBD2, are expressed at higher levels than α-defensins 1 and 3 in oral epithelium making them strong candidates for innate resistance to oral HIV infection. hBD1 (*DEFB1*), hBD2 (*DEFB4*) and hBD3 (*DEFB103*) were all reported to inhibit, to some extent, HIV-infection by direct inactivation of virions and suppressing viral replication in human oral cavity (Quiñones-Mateu *et al.* 2003; Sun *et al.* 2005). The defensins appear to have an important role in innate and adaptive resistance to viral infection through both their antiretroviral effect and chemokine-like effect against HIV.

An important inflammatory disease associated with β-defensins is psoriasis. The β-defensins, especially hBD-2 (*DEFB4* gene), which was first isolated from psoriatic lesions, has recently been highlighted as a possible candidate gene for psoriasis. The role of β-defensins in the immune system and the aetiology of psoriasis strongly support such a hypothesis, but until 2008 no study had investigated this relationship. Understanding the role of β-defensin, especially

the consequences of β-defensin copy number variation in psoriasis was a central aim of my PhD project, which I will explore in more detail.

1.2.4.1 Immuno-genetics of psoriasis

Psoriasis is a chronic inflammatory skin disorder with 2 to 3% prevalence in the European population and a complex aetiology strongly influenced by genetics (Lebwohl 2003). In other human populations psoriasis is less common with an incidence of only about 0.1% in Asians and even lower values in Africans (Bowcock 2005). Psoriasis is a T-cell-mediated inflammatory disease characterized by uncontrolled proliferation of keratinocytes and recruitment of T-cells into the skin (Valdimarsson *et al.* 1995). Long described as an autoimmune disease, its pathogenesis fits the classical definition of an autoimmune disease: "clinical syndrome caused by the activation of T-cells and/or B-cells, in the absence of an ongoing infection or other discernable cause" (Davidson and Diamond 2001).

Psoriasis can be present in different forms - Psoriasis Vulgaris, Guttate, Flexural and Pustular, varying according the age of onset, severity and the part of the body affected. Both sexes are affected equally. Psoriasis Vulgaris can be further sub-divided into type I, if onset occurs earlier in life (<=40 years), often associated with a positive family history and inheritance of particular HLA alleles that predispose to a higher incidence of streptococcal infection triggering psoriasis, and type II, if it is associated with a late onset (>40 years) and absence of the family history and without predisposing HLA antigens (Bhalerao and Bowcock 1998; Weisenseel *et al.* 2002).

The most common type of psoriasis, psoriasis vulgaris or plaque psoriasis, is characterized by epidermal hyper-proliferation, vascular remodelling and inflammation resulting in the formation of elevated, scaly, erythematous and thickened plaques normally located on the elbows, knees, scalp, nails and joints. It can nevertheless also develop in any other part of the body (Lebwohl *et al.* 2003). About 15% of the individuals with psoriasis also develop psoriatic arthritis (PA), normally within the first 10 years after the diagnosis of psoriasis (Ibrahim *et al.* 2009).

Psoriasis can occur in association with other complex diseases such as Crohn's disease (CD), type II diabetes (Wolf *et al.* 2008) and human immunodeficiency virus (HIV) infection (Duvic 1990). Psoriasis and Acquired immune deficiency syndrome (AIDS) are immune-mediated disorders and have been associated with HLA antigens (Duvic 1990). Psoriasis has also been reported to share genetic variants with both CD and type II diabetes, suggesting that some genes play a role in different inflammatory pathways and may contribute to the pathogenesis of multiple diseases (Capon *et al.* 2007; Cargill *et al.* 2007; Wolf *et al.* 2008).

The pathogenesis of psoriasis is complex and involves the activation and continuous stimulation of T-cells with excessive infiltration of many types of leukocytes, namely dendritic and T-cells, into the dermis and epidermis (Bowcock and Krueger 2005). The cellular and inflammatory pathway in psoriasis is initiated with T-cell activation through IL-12 and IL-23 cytokines that trigger the differentiation of naive $CD4^+$ T-cells (Th0) into different T-cell subsets followed by their consequent clonal expansion. The T-cell activation through IL-23 and IL-12 leads to the activation of a cascade pathway characterized by the synthesis of T-cell-derived pro-inflammatory cytokines that stimulates the proliferation of keratinocytes and endothelial cells, resulting in marked thickening of the epidermis and permanent inflammation recognized in psoriasis (Lowes *et al.* 2007; Duffin *et al.* 2010). The Th17 cells, a particular activated T-cell subset implicated in the pathogenesis of psoriasis, synthesize cytokines such as IL-17 and IL-22 which activate keratinocytes, resulting in their proliferation and release of pro-inflammatory cytokines, chemokines, TNF-α (tumour necrosis factor-alpha) and antimicrobial peptides (β-defensins) that promote inflammatory responses and are responsible for the persistent inflammation characteristic of psoriatic lesions (Lowes *et al.* 2007; Duffin *et al.* 2010).

T-cell activation requires the interaction with antigen-presenting cells (Langerhans cells), through a variety of cell-surface molecules, such as the MHC. They express many different pattern recognition receptors (PRR) in their surface that interact with a range of microorganisms, such as bacteria and viruses, leading to their maturation and migration to lymph nodes. Here,

antigen-presenting cells present antigens to the antigen-specific receptors of T-cells, leading to their activation and proliferation (Lebwohl 2003; Duffin *et al.* 2010). The factors leading to dendritic cell activation are not yet completely clear. However, some mechanisms have been suggested, which include activation through cytokines, pattern-recognition receptors, heat-shock proteins or direct interaction with counter-receptors in T-cells (Lebwohl 2003; Lowes *et al.* 2007). Many environmental factors have been linked with T-cell activation. These have been described to initiate and enhance the progression of psoriasis (Leung *et al.* 1995; Valdimarsson *et al.* 1995). The exact contribution and importance of both genetic and environmental factors to psoriasis susceptibility is not yet completely clear; however, it is commonly accepted that genetics may have a stronger role in the risk of developing psoriasis. Nevertheless, psoriasis is regarded as a multifactorial disease in which several genes interact with each other and where environmental factors can play a role in the development of the disease. This evidence is provided by studies in monozygotic and dizygotic twins. Such studies reveal that monozygotic twins have a higher rate of concordance for psoriasis (between 70-72%) than dizygotic twins (between 15-23%), supporting a genetic origin of psoriasis (Farber *et al.* 1974; Brandrup *et al.* 1978). However, the heritability of psoriasis in the Australian population is not so high, showing only 35% concordance among monozygotic twins and 12% concordance in dizygotic twins (Duffy *et al.* 1993). Since not all monozygotic twins share the disease, together with the fact that concordance of psoriasis never reaches 100% in any population studied, suggests that environmental factors can also play a role in psoriasis susceptibility.

In order to elucidate the exact contribution of genetics in the development of psoriasis, several linkage and association studies have been carried out in the last 30 years with the aim of mapping genes conferring susceptibility to psoriasis. As a result, psoriasis has been associated with ten different loci (*PSORS1 to PSORS10*), each of them normally including more than one gene. In the early 1970s, Russell *et al.*, first described an association between psoriasis and the human leukocyte antigens (HLA) in the major histocompatibility locus complex (MHC) on chromosome 6. The MHC locus has been consistently identified in many studies since then and is one of the major susceptibility loci

for psoriasis. This locus is also known as "psoriasis susceptibility 1" or *PSORS1* comprising a 300 kb region in the MHC on chromosome 6p21.3, which includes the genes that encode the human leukocyte antigens (HLA) (Russell *et al.* 1972). The *HLA-Cw6* (HLA) has been identified as the primary risk allele for psoriasis with 60% of patients with early onset psoriasis carrying this allele (Tiilikainen *et al.* 1980; Gudjonsson *et al.* 2003; Nair *et al.* 2006). However, only 10-15% of *HLA-Cw6* carriers develop psoriasis, highlighting the fact that other genetic factors may contribute to the risk of psoriasis (Duffin *et al.* 2010). The association of psoriasis with the *HLA-Cw6* allele suggests that MHC class I molecules are involved in the T-cell activation process through antigen-presenting cells (APC), namely in the APC binding to the respective ligands on T-cells (Gottlieb and Krueger 1990). The *PSORS1* locus contains several other candidate genes, such as CCHCR1 (HRC) and CDSN that encode for the coiled-coil α-helical rod protein 1 and corneodesmosin, respectively (Asumalahti *et al.* 2000; Bowcock and Krueger 2005). CCHCR1 is upregulated in psoriatic epidermis and might negatively regulate keratinocyte differentiation and proliferation. Corneodesmosin, a protein involved in keratinocyte cohesion and desquamation, is overexpressed in the epidermis of psoriatic patients and could be involved in the decreased (impaired) desquamation, contributing to the abnormal scaling process of psoriatic lesions (Bowcock and Krueger 2005).

The precise location of the *PSORS1* locus is highly controversial. It has been suggested that genes encoding the classical HLA class I molecules (HLA-A, HLA-B and HLA-C) are not included in this region which could mean that *HLA-Cw6* is a marker in linkage disequilibrium with *PSORS1* locus rather than the susceptibility allele itself (Nair *et al.* 2000; Orrù *et al.* 2005). On the other hand, several other studies exclude the CCHCR1 (HRC) and CDSN genes and map the *PSORS1* locus to the HLA-C region (Helms *et al.* 2005). Due to the strong linkage disequilibrium among the associated alleles of the *PSORS1* locus the exact extent of this region is difficult to disentangle. Despite the well documented evidence of the role of MHC class I molecules, in particular the *HLA-Cw6* allele, in the susceptibility to psoriasis, the exact genetic contribution of the *PSORS1* locus in psoriasis is still not completely clear (Enlund *et al.* 1999).

The first psoriasis–susceptibility locus other than the MHC locus was described on chromosome 17q24-q25 by Tomfohrde *et al.* (Tomfohrde *et al.* 1994). This psoriasis-susceptibility locus, known as *PSORS2,* includes three different genes, *SLC9A3R1* (solute-carrier family 9, isoform 3, regulator 1), *NAT9* (N-acetyltransferase 9) and *RAPTOR* (regulatory associated protein of mammalian target of rapamycin (MTOR)) (Tomfohrde *et al.* 1994; Helms *et al.* 2003). *SLC9A3R1* encodes for a binding phosphoprotein with two PDZ domains that interact with the cytoskeleton proteins ezrin, radixin and moesin. This phosphoprotein is involved in several epithelial membrane protein-protein interactions, especially in the interaction of antigen-presenting cells and T-cells. It has been proposed that the binding complex formed between the *SLC9A3R1* domains and cytoskeleton proteins leads to the immunological synapse formation and ultimately to T-cell activation (Helms *et al.* 2003; Bowcock and Krueger 2005). A single nucleotide polymorphism lying between *SLC9A3R1* and *NAT9* leads to the loss of RUNX1 binding, suggesting a defective regulation of *SLC9A3R1* and *NAT9* by RUNX1 (a transcription factor protein involved in DNA binding). Therefore, these genes have been identified as a susceptibility factor for psoriasis (Helms *et al.* 2003). Psoriasis has also been associated with variants in the *RAPTOR* gene (Helms *et al.* 2003). Since the variants were found in the non-coding region of the gene, this suggests its role as a regulatory gene of the MTOR protein. Therefore, alteration in the regulatory pathway of MTOR can lead to modifications in the normal cell growth and proliferation rate of T-cells and keratinocytes (Capon *et al.* 2004).

With the use of linkage analysis in large families, a "psoriasis susceptibility 3" locus was located on chromosome 4q25. This evidence was demonstrated through a genome-wide scan in which the microsatellite marker, *D4S1535,* shows a maximum pair wise LOD score of 3.03 (Matthews *et al.* 1996). Located just 50 kb from the *PSORS3* locus lies the human IFR2 gene (interferon regulatory factor 2) showing two markers, located in IFR2 exon 9, associated with type 1 psoriasis (Foerster *et al.* 2004). This gene is also involved in interferon (IFN) regulation and has a role as transcriptional repressor of α- and β-interferon target genes. Genetic variants, such as SNPs, may have an effect on the transcriptional regulation of IRF2 resulting in a deficient expression or

function of the gene, which can lead to hyperresponsiveness to type 1 interferon signalling, a process involved in the pathogenesis of psoriasis (Foerster *et al.* 2004).

The epidermal differentiation complex (EDC) spans about 2 Mb on chromosome 1q21 and includes several genes involved in epidermal differentiation and maturation. A *PSORS4* locus located within the EDC region was associated with psoriasis (Capon *et al.* 1999). Taking into consideration the role of the EDC genes, it is possible that genetic variants could affect keratinocyte proliferation and differentiation. Such a relationship can lead to the development of an abnormal inflammatory response, such as in psoriasis (Lowes *et al.* 2007). Within the EDC region are placed two different clusters of genes, the genes encoding S100 proteins (small proline-rich proteins) and the late cornified envelope (*LCE*) proteins. A deletion comprising the *LCE3B* and *LCE3C,* genes from the *LCE* gene cluster, is strongly associated with risk of psoriasis (p-value=1.38×10^{-8}) in the European and North American population. The *LCE3C_LCE3B_del* is in strong LD with a nearby SNP (rs4112788) 4.5 kb upstream, suggesting a single origin (de Cid *et al.* 2009). Moreover, the biological implication of these genes for psoriasis was also confirmed in a genome-wide association study in the Chinese population (Chinese Han ancestry). In this study, Zhang *et al.* identified another SNP within the *LCE* gene cluster that confers susceptibility to psoriasis (rs4085613, p-value=6.69×10^{-30}) (Zhang *et al.* 2009c). This SNP is found in the same linkage disequilibrium block of the rs4112788, reported by de Cid *et al.,* and they are in almost complete LD (Zhang *et al.* 2009c). These studies bring to attention the importance of genetic variants within *LCE* genes for the development of psoriasis. Since the LCE proteins play a role in the epidermal terminal differentiation of keratinocytes, genetic variation within these genes may lead to the formation of an abnormal cornified envelope with poor adherence causing the release of cells and the scaling appearance found in psoriatic lesions (de Cid *et al.* 2009; Zhang *et al.* 2009c).

The *PSORS5* locus maps to chromosome 3q21. So far only one candidate gene, *SLC12A8,* has been identified and associated with risk of psoriasis in this locus (Enlund *et al.* 1999; Hewett *et al.* 2002; Hüffmeier *et al.* 2005). *PSORS5*

encodes for a member of the solute carrier family 12 proteins (member A 8). These are integral membrane proteins with cation/chloride co-transporter functions that may play a role in the control of keratinocyte proliferation. In the study of a Swedish population, a five-SNP haplotype spanning the 3' half of the *SLC12A8* gene has been identified and associated with psoriasis (*p*-value of 3.8×10^{-5}) (Hewett *et al.* 2002). However, in a more recent study in a German population only one SNP (rs2228674) with very weak association with psoriasis was found within the *PSORS5* locus (Hüffmeier *et al.* 2005). Therefore, the variants that make *PSORS5* a psoriasis susceptibility locus still remain to be investigated.

A genome-wide linkage scan in a German population was the first study to identify a psoriasis susceptibility locus at 19q13 (*PSORS6*) (P=0.0002) comprising about 15 cM (Lee *et al.* 2000). The interval spanning the *PSORS6* region is occupied by several genes, and one of them is the gene coding for ICAM-1, an intercellular adhesion molecule-1 (Nickoloff *et al.* 1993). ICAM-1 is expressed in leukocytes, fibroblasts and endothelium, and mediates leukocyte migration into sites of infection and T-cell activation, suggesting that keratinocytes may have a role in the regulation of T-cell activation, a crucial process involved in the pathogenesis of psoriasis (Nickoloff *et al.* 1995). Moreover, the 19p13 psoriasis susceptibility locus is in interaction with the *HLA* region on chromosome 6p (Lee *et al.* 2000). The interaction between *PSORS6* and *PSORS1* loci is also evident in a study where the observed association of *PSORS6* with psoriasis is only true in type I psoriasis patients carrying the *PSORS1* risk allele, suggesting epistasis between these two loci (Hüffmeier *et al.* 2009).

A further susceptibility locus, known as *PSORS7*, was initially described by Veal *et al.*, at chromosome 1p (Veal *et al.* 2001). This psoriasis susceptibility locus has been reported in several genome-wide association scans carried out in large cohorts of Caucasian individuals. These studies reveal the association between psoriasis and SNPs within the *IL12B* (encoding the IL-12p40 subunits of two cytokines, *IL-12* and *IL-23*), *IL23R* (encoding a subunit of the *IL-23* receptor) and *IL23A* (encoding the p19 subunit of *IL-23*) genes (Cargill *et al.* 2007; Nair *et al.* 2009). While *IL23R* and *IL23A* are located within the *PSORS7* locus at 1p31.3, *IL12B* maps outside this locus on chromosome 5q31.1-q33.1

(Cargill *et al.* 2007). All these genes are involved in *IL-23* signalling, regulating the immune responses and promoting the survival and proliferation of Th17 as well as the release of IL-17, a key cytokine involved in the pathogenesis of psoriasis. *IL-23R* and *IL-23A* have also been associated with other autoimmune diseases, such as Crohn's disease, Ulcerative colitis and Psoriatic arthritis (Duerr *et al.* 2006; Cummings *et al.* 2007; Liu *et al.* 2008; Nair *et al.* 2009). From these studies is evident that several autoimmune diseases can share the same disease-risk alleles.

Although other *PSORS* loci (8, 9 and 10) have been mapped to chromosome 16q, 4q31 and 18p11, there is very little evidence to show association with psoriasis disease (Nair *et al.* 1997; Zhang *et al.* 2002; Asumalahti *et al.* 2003; Karason *et al.* 2003; Zhang *et al.* 2007).

In addition to the psoriasis susceptibility genes just described, other genes not mapped to any of the psoriasis susceptibility loci have been associated with psoriasis. At chromosome 20q13 is the SNP rs495337, a disease-associated variant (p-value=1.4×10^{-8}) mapping to the *SPATA2* gene (encodes the spermatogenesis-associated protein 2) (Capon *et al.* 2008). The function of this gene might not be particularly interesting for psoriasis, but the SNP rs495337 is in strong LD with 5 other SNPs in the *ZNF313* gene, which is expressed in skin, T-lymphocytes and dendritic cells. The function of *ZNF313* is unknown, but from homology studies it was suggested that *ZNF313* has similar functions to those included in a RING domain E3 ubiquitin ligase family, regulating T-cell activation (Capon *et al.* 2008). Such evidence puts the *ZNF313* gene on the long list of genes conferring susceptibility to psoriasis disease.

More recently, several genome-wide association studies, including large cohorts of individuals not just from European ancestry but also from a Chinese Han population, have identified new susceptibility loci associated with psoriasis. The *TRAF3IP2* gene and the *NOS2* gene were identified from GWAS. *TRAF3IP2* gene is involved in the IL-17 signalling, while *NOS2* gene encodes an inducible nitric oxide synthase, which is involved in the regulation of dendritic cells and many other immune system activated cells (Ellinghaus *et al.* 2010; Huffmeier *et al.* 2010; Stuart *et al.* 2010). Some psoriasis susceptibility loci appear to be population-specific, conferring susceptibility to the disease just

in a particular population but not for all human populations (Sun *et al.* 2010). While some susceptibility loci identified in candidate gene studies have been replicate in recent GWAS showing a genome-wide level of statistical significance, such as *HLA-C*, three genes involved in IL-23 signalling (*IL12B, IL23A, IL23R*), *SPATA2*, *TRAF3IP2* and *NOS2*, the association of other loci with psoriasis have not been supported by these studies (Nair *et al.* 2009; Ellinghaus *et al.* 2010; Huffmeier *et al.* 2010; Stuart *et al.* 2010). This might be explained by the small sample sizes used in association studies, which could lead to false positive associations, compared with the large sample sizes used nowadays in GWAS. The use of large cohorts in GWAS increase the power of the association, which leads to the more reliable discovery of loci genuinely involved in disease phenotypes.

The susceptibility loci described so far give support to the classification of psoriasis as a complex immune disease with a strong genetic background. Even so, many of these genetic factors only explain a small fraction of the risk of the disease. So, it still remains to be discovered which other genetic variants account for the risk of psoriasis. In addition to SNPs, copy number variation (CNV) has been identified as an important risk factor in the susceptibility for complex diseases, such as psoriasis (Hollox *et al.* 2008b). While the psoriasis-susceptibility loci based on SNP variants lead to alterations in gene function and regulation, copy number variation may lead to a change in gene dosage. This highlights the importance that genetic variants other than SNPs should be taken into consideration as risk factors in several complex diseases, such as psoriasis, Crohn's disease and diabetes.

1.2.5 Defensins and CNV in non-human primates

Defensins are present across all vertebrate species. In birds and some mammalian species (cows and pigs) only β-defensins (Gallinacin in birds) have been identified (Harwig *et al.* 1994; Zhao *et al.* 2001; Ganz 2003; Thouzeau *et al.* 2003). Peptides structurally homologous to β-defensin were found in the venom of reptile species (Nicastro *et al.* 2003) and β-defensin orthologues are present in different mammalian species, such as dog, rat, mouse and non-human primate species, for example chimpanzee (Patil *et al.* 2005). This

suggests that β-defensin is the primordial defensin and the common ancestor for all vertebrate defensins. Comparative analysis of the chicken and mammalian β-defensin gene clusters revealed shared synteny, which may indicate that they arose before the divergence of mammals from birds (Xiao *et al.* 2004). At some stage in mammalian evolution, successive rounds of duplication followed by substantial divergence involving positive selection gave rise to a diverse cluster of genes, the α-defensins, in glires and primates about 91 million years ago (Patil *et al.* 2004; Crovella *et al.* 2005). An alpha-defensin cluster is located adjacent to the β-defensin cluster in humans, mouse and rat, but is absent in the canine genome (Patil *et al.* 2005), suggesting that the α-defensin cluster evolved in a species-specific manner during mammalian evolution (Crovella *et al.* 2005).

In primates a third class of defensins, theta-defensins has been suggested to derive from a process involving the fusion and cyclization of two α-defensin related monopeptides, which occurred only after divergence of primates from other mammals around 23 million years ago (Tang *et al.* 1999). The close relationship between theta-defensins and α-defensins has also been shown by Leonova *et al.* (2001), who with the aim to identify the theta-defensin precursors, undertook a cloning strategy and found several α-defensin transcripts whose sequence contained a premature stop codon after the third cysteine residue. This indicates that these circular defensins are products of a process that generates molecular diversity without corresponding genome expansion (Tang *et al.* 1999; Leonova *et al.* 2001).

The theta-defensin genes have been found to be active in orangutans and Old World monkeys but not in humans or New World primates (Nguyen *et al.* 2003). The presence of multiple, divergent subsets of defensin genes in each species may suggest that the evolution of these antimicrobial peptides occurred in response to different environmental pressures reflecting an evolutionary adaptation of the innate immune system to diverse "microbial ecology" (Crovella *et al.* 2005; Selsted and Ouellette 2005).

The chromosomal location of the genes encoding the β-defensins appear to be conserved in mammals, supporting the idea of a pre-mammalian origin of

β-defensins (Crovella *et al.* 2005). Although the majority of β-defensins are conserved across different mammalian species, there are some species-specific gene lineages indicating that some β-defensins evolved after the last common ancestor of mammal (Patil *et al.* 2005). In the rat, mouse and dog, four different β-defensin clusters are present, whereas human and chimpanzee show five different clusters. The five β-defensin clusters in the chimpanzee genome are in conserved synteny with the human β-defensin genes (Patil *et al.* 2005). This suggests that after duplication, β-defensin genes were subjected to positive selection earlier in mammalian evolution (Semple *et al.* 2005). In primates different selective pressures acted on β-defensin genes in different evolutionary lineages, leading either to positive or negative selection, during the divergence of primates (Semple *et al.* 2005; Semple *et al.* 2006).

Widespread copy number variation has been described in the genomes of chimpanzees (Perry *et al.* 2006; Perry *et al.* 2008), rhesus macaques (Lee *et al.* 2008), mice (Li *et al.* 2004; Cutler *et al.* 2007; Egan *et al.* 2007; Graubert *et al.* 2007), rats (Guryev *et al.* 2008) and even in the genomes of *Drosophila melanogaster* (Dopman and Hartl 2007) and *Plasmodium falciparum*, the malaria parasite (Emerson *et al.* 2008). Differences in copy number in the *Drosophila melanogaster* genome have been most notably found to encompass toxin-response genes (Emerson *et al.* 2008). The CNVs described in mammals and non-human primates are particularly interesting and important to understand the functional and evolutionary significance of human CNVs. The studies of CNV in other mammalian species are limited; however, a number of CNV regions have been revealed, many of them overlapping human CNVs. In rats, 113 CNVs were identified in regions orthologous to human CNVs, of which 80 are implicated in human disease (Guryev *et al.* 2008). Analysis of copy number variation in rhesus macaque, using array-based comparative genomic hybridization (array-CGH), revealed that about 20% (25) of the 123 identified macaque CNVs are mapped to regions of the human genome previously found to contain CNVs in 270 HapMap individuals (Lee *et al.* 2008). However, more recently, using an array-CGH platform specifically designed to study copy number variation in the rhesus macaque, Gokcumen *et al.* (2011), identified 1160 CNVs, of which 385 overlapped with 467 human

CNVs. The considerable difference in the number of CNVs identified between studies may reflect the use of a species-specific array platform, instead of human-based arrays, which are inadequate taking in consideration the sequence divergence between species. This suggests the importance of the design of species-specific arrays to realistically assess the CNV in non-human primates. In the chimpanzee genome, CNVs, identified in 30 wild born chimpanzees, were observed for 438 autosomal regions of which 144 overlapped to human CNVRs (Perry *et al.* 2008). This value was soon proved by Gokcumen *et al.* (2011) to under represent the real frequency of CNV in chimpanzee genome, which identified 556 chimpanzee CNVs overlapping with more than one thousand (1387) distinct human CNVs; 170 human CNVs were found to overlap with both chimpanzee and rhesus macaque (Gokcumen *et al.* 2011).

A common feature between the human and other primate genomes is the frequent location of CNVs in regions of segmental duplications with high levels of sequence identity, compared with the genomes of other mammals (Bailey and Eichler 2006). In the human genome, several studies have shown that CNVs regions are enriched by segmental duplications (SDs) (Iafrate *et al.* 2004; Sebat *et al.* 2004; Tuzun *et al.* 2005; Redon *et al.* 2006; Korbel *et al.* 2007; Kidd *et al.* 2008). It was suggested that about 51.6% of CNVRs overlap segmental duplications in the human genome and a similar level of enrichment was observed in the chimpanzee genome (39%) (Perry *et al.* 2008). SDs have also been found to be enriched at CNV regions in the genome of rhesus macaque (Lee *et al.* 2008) and orangutan (Marques-Bonet *et al.* 2009). However, orangutan has fewer rearrangements and segmental duplications than humans and chimpanzees (Locke *et al.* 2011).

Humans and chimpanzees have significantly more SDs than the other primate species, with a large portion being shared between human and chimpanzee (Marques-Bonet *et al.* 2009). SDs have been strongly associated with genomic instability and large-scale chromosomal rearrangements, since their high sequence homology gives rise to recurrent NAHR (Lupski 1998; Sharp *et al.* 2006). This suggests their involvement in evolutionary rearrangements and eventually in the origin of novel genes in the primate

lineage (Samonte and Eichler 2002; Armengol *et al.* 2003; Bailey *et al.* 2004; Bailey and Eichler 2006).

The presence of copy number variation at orthologous genomic regions in both human and chimpanzee or rhesus macaque genomes (shared CNVs) are likely to reflect unstable genomic regions that have been prone to recurrent rearrangements during primate evolution, rather than maintenance of ancestral polymorphisms. This suggests that structural features shared between primate genomes, such as SDs, may predispose certain chromosomal regions to structural instability in these primate species (Kehrer-Sawatzki and Cooper 2008; Perry *et al.* 2008), as previously reported for the human genome (Samonte and Eichler 2002; Armengol *et al.* 2003; Bailey *et al.* 2004). On the other hand, lineage-specific CNVs, which either evolved by positive selection in one species or by negative selection in other contributing to intra- and interspecies phenotypic variation, are important to understand the mechanisms under the origin and evolution of CNV (Kehrer-Sawatzki and Cooper 2008).

The majority of primate CNVs have been shown to overlap functional genomic regions, especially enriched by genes involved in immunity and environmental response (Nguyen *et al.* 2006; Redon *et al.* 2006; Korbel *et al.* 2008; Gokcumen *et al.* 2011). Similar patterns of common copy number diversity between humans and other primates were found for *FCGR3A/B*, *CCL3L1* and the β-defensin cluster. In human and chimpanzee genomes, *FCGR3A/B* genes commonly show 4 copies per diploid genome, but fewer individuals can have three or five copies, in humans (Perry *et al.* 2008). *CCL3L* CNV was observed in chimpanzee and rhesus macaque (Lim *et al.* 2010). Consistent with later findings in humans (Shao *et al.* 2007; Field *et al.* 2009; Urban *et al.* 2009), suggesting that *CCL3L1* CNV is not associated with HIV-1, in rhesus macaque the CNV of *CCL3L* is not associated with susceptibility to AIDS in individuals infected with simian immunodeficiency virus (SIV) (Lim *et al.* 2010). Copy number variation of α- and β-defensin genes has been described in humans and other primates but not in other mammals. In humans β-defensin genes at 8p23.1 are highly polymorphic in copy number. Variation in copy number of these genes was reported in chimpanzee (Perry *et al.* 2008)

and in rhesus macaque the *DEFB4* gene was observed to be copy number variable as well (Lee *et al.* 2008).

In a study carried out within the *Drosophila melanogaster* genome, genes overlapped by CNVs were reported to have increased evolutionary rates (Dopman and Hartl 2007). In humans, elevated evolutionary rates were also observed for genes located within CNV regions (Nguyen *et al.* 2008). Moreover, it has been suggested that CNV regions are enriched by genes implicated in "environmental" functions, possibly essential for the adaptation to the rapid changing environments during human evolution (Gibbs *et al.* 2004; Feuk *et al.* 2006; Nguyen *et al.* 2006; Sharp *et al.* 2006). The β-defensin genes are key components of the innate immune system with important roles in immunity, reproduction and pigmentation. Their multifunctional role, and increasing evidence that β-defensin CNV is implicated with disease phenotypes in humans, suggests that this CNV could be a hotspot for evolution. Therefore, it is essential that the study of CNV will be extended to other great ape, such as gorilla and orangutan, to shed light into the origin and evolution of CNVs in primate lineage

1.3 AIMS OF THE PhD

The proposed PhD project will examine the variation in human β-defensin gene number and its effects in disease, human diversity and human evolution. The research will include investigation of novel methodology for gene copy number measurement for analysis of variable defensin genes, analysis of association between gene copy number and disease, and exploration of mechanisms of gene copy number variation in human populations.

The project is divided into three major objectives for the proposed research. Firstly, new methods will be developed to accurately but conveniently assess copy number variation, making use of the paralogue ratio test (PRT), to build an accurate and high-throughput multiplex assay for β-defensin copy number measurement (chapter 3). Secondly, these newly developed methods will be used in collaborations to evaluate copy number in collections of clinical samples, to look for association between the copy number of β-defensin genes and disease states in psoriasis (Section 4.1). Finally, examination of copy number in human populations (Section 4.2) and non-human primates (chapter 6) will be used to investigate human inter-population diversity and investigate when variation in copy number has arisen in human populations. This work hopes to contribute to the increased effort to understand the consequences of DNA variation for health and disease.

CHAPTER 2: MATERIALS AND METHODS

2.1 MATERIALS

2.1.1 DNA samples

2.1.1.1 ECACC Human random control (HRC) samples

The study of unrelated UK Caucasians used control samples from the ECACC Human Random Control (HRC) panels 1 and 2 (http://www.hpacultures.org.uk/products/dna/hrcdna/hrcdna.jsp). These samples were from unrelated UK Caucasian blood donors and were prepared at a 10ng/μl DNA concentration. The genomic DNA was extracted from lymphoblastoid cell lines derived by Epstein Barr Virus (EBV) transformation of peripheral blood lymphocytes from single donor blood samples. Then, the DNA samples were subjected to standard techniques of DNA extraction and purification and were provided at a standard concentration of 100ng/μl in 10mM Tris-HCl buffer (pH 8.0) with 1mM EDTA. Due to its standard concentration and consistently high quality provided these DNA panels were frequently used as internal reference controls in the PRT assays.

2.1.1.2 HapMap CEU/YRI/CHB/JPT

In this study DNA samples from the International HapMap Project collections (http://hapmap.ncbi.nlm.nih.gov/index.html.en) were used to study the β-defensin copy number variation in different human populations. These samples were collected from blood donors of four different Human populations: Han Chinese from Beijing, China (CHB); Japanese from Tokyo, Japan (JPT); Yoruba from Ibadan, Nigeria (YRI) and US residents from Utah (CEPH/CEU). A set of 30 US trios (90 samples) with northern and western European ancestry were collected by the Centre d'Etude du Polymorphisme Humain (CEPH) and included in the CEPH/CEU HapMap collection. Another set of 30 parent-adult child trios (90 samples) with Nigerian ancestry, with all parents described to have four Yoruba grandparents, was included in the YRI collection. In the JPT collection 45 unrelated individual samples from people whose grandparents were all Japanese were included while in the CHB collection, equally composed by 45 unrelated individuals, all samples came from individuals with at least three Han Chinese grandparents. The DNA concentration provided for all samples was 250ng/μl with 50μg of DNA per well.

2.1.1.3 Disease status samples

A Dutch cohort (from Nijmegen) was used with the aim of studying the β-defensin copy number variation in psoriasis disease and testing the applicability of the HSPD5.8 PRT and the Triplex system in large association studies. Both controls and cases from the Dutch cohort were from native European Dutch origin and were supplied by our collaborator, Joost Schalkwijk from the Department of Dermatology at the Radboud University Nijmegen Medical Centre, Netherlands. 202 psoriatic DNA samples were obtained from patients diagnosed with psoriasis vulgaris from the outpatient clinic of the Radboud University Medical Centre. This patient group included individuals diagnosed with moderate to severe psoriasis aged between 39 and 69 years old (male:female ratio of 70:30). Sixty-four healthy control samples were obtained from the Nijmegen Biomedical study (NBS). The mean age of the control cohort was 58 (±13) years with a male:female ratio of 56:44. The genomic DNA from all blood donors was isolated through standard methods (Qiagen

column or salting-out procedures). All samples from the cases and about half of the controls were prepared by the same method (Qiagen) (Hollox *et al.* 2008b).

2.1.1.4 Non-human primates samples

In this study ten non-human primates were used, namely five gorilla (*Gorilla gorilla*), three chimpanzees (*Pan troglodytes*), one bonobo (*Pan paniscus*) and one orangutan (*Pongo* sp.) from unknown subspecies. The chimpanzee EB176 JC and the gorilla EB JC DNA samples are from ECACC (http://www.hpacultures.org.uk/products/dna/primatedna.jsp). The exact origin of DNA samples of Candy and Violet chimpanzees, orangutan, bonobo and gorillas, Sylvia, Tomoka, Guy and J79 were not possible to track, though they were supplied by Professor Alec Jeffreys's laboratory from the University of Leicester. The DNA concentration provided was 0.5ng/ml for bonobo, 200ng/µl for orangutan, chimpanzee EB176 JC and gorilla EB JC and 50ng/µl for the remaining samples.

2.1.2 Reagents

2.1.2.1 10X PCR MIX

The "10X PCR buffer", with high concentrations of $MgCl_2$ and dNTPs, was essential to perform the HSPD5.8 PRT. This buffer consists of 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.

2.1.2.2 10X Ld PCR Mix

The "10X Ld" ("Low dNTP") PCR mix was used for the Triplex system, microsatellites and RFLP (restriction fragment length polymorphism) assays. This buffer contained a final concentration of 50mM Tris-HCl (pH 8.8), 12.5mM ammonium sulphate, 1.4mM magnesium chloride, 125µg/ml BSA, 7.5mM 2-mercaptoethanol and 200µM of each dNTP.

2.1.2.3 TE buffer

TE buffer was used to dissolve and dilute DNA. The 1xTE buffer was composed of 10 mM Tris-HCl (pH 8), which maintains the solution at a specific pH and 1mM EDTA, a chelator of divalent metals ions, particularly Mg^{2+} and Ca^{2+} which are co-factors for many enzymes including nucleases. As such, the EDTA helps to protect DNA and RNA from enzymatic degradation.

2.1.2.4 0.5X TBE Buffer

This buffer was used in agarose gel electrophoresis and was made from 10X TBE buffer, which contained a final concentration of 1M Tris-HCl, 1M boric acid and 10mM EDTA pH 8.0.

2.1.3 Primer design

PCR primers for the β-defensin region at 8p23.1 were designed from the human reference sequences available on the UCSC Genome Brower March 2006 assembly (http://genome.ucsc.edu/). The study of non-human primates also used the orangutan and chimpanzee reference sequences available on the UCSC Genome Brower on Orangutan July 2007 (WUGSC 2.0.2/ponAbe2) assembly and Chimpanzee Mar. 2006 (CGSC 2.1/panTro2) assembly, respectively. The Ensembl genome browser was used to look for gorilla reference sequences in order to design PCR primers specific for this species at the region of interest (http://www.ensembl.org/index.html).

The thermodynamic properties of PCR primers were confirmed and established by the Primer3 software (http://frodo.wi.mit.edu/). A Basic Local Alignment Search Tool (BLAST) (http://blast.ncbi.nlm.nih.gov) was used to compare the PCR primer sequences (query sequence) with a library of sequences (Trace Archive, NCBI) and identify library sequences that were similar to the query sequences above a certain threshold. These bioinformatics tools were essential to check for common sequence variants underneath the primers.

2.2 METHODS

2.2.1 Polymerase chain reaction (PCR)

2.2.1.1 PCR in 10X PCR MIX and 10X Ld PCR MIX

PCR reactions were set up in a master mix of 10µl or 20µl final volume using the 10X PCR or 10X Ld PCR buffer as described in sections 2.1.2.1 and 2.1.2.2. The 10X PCR mix was mainly used in the PCR for HSPD5.8 PRT (section 2.2.3.1), while all the other PCR reactions used the 10X Ld PCR buffer. In addition to the 10X PCR or 10X Ld PCR buffer, PCR master mix reactions contained final concentrations of 0.5µM of each primer, 0.05U *Taq* DNA polymerase and 10ng input DNA.

Subsequently, PCR products were amplified using 22-37 standard cycles that included denaturation, annealing and extension/elongation steps (e.g. 95°C for 30 seconds, 58°C for 30 seconds and 70°C for 1 minute). Cycle temperatures, duration and number varied according to the properties of the primers used and the products expected. The PCR cycle was frequently followed by a single "chase" phase of annealing for 1 minute and extension for 20 or 40 minutes to reduce levels of single-stranded DNA and complete terminal 3' dA addition. In the presence of GC-rich regions the PCR cycle was preceded by an initial denaturation of 95°C for 5 minutes.

2.2.2 DNA electrophoresis

2.2.2.1 Agarose gel electrophoresis

Gel electrophoresis, using an electric field applied to an agarose gel matrix, was used for the separation of DNA PCR products. The sizes of PCR products were determined by comparison with a DNA ladder (100 bp DNA Ladder, New England BioLabs) containing DNA fragments of known size, which were loaded alongside with the PCR products. The gels were prepared with an agarose concentration of 2-3% (w/v) and the appropriate volume of 0.5X TBE containing 0.5g/ml ethidium bromide.

In order to run the DNA samples, these were prepared in a 10% solution of 5X loading buffer (0.02% bromophenol blue dye, 40% sucrose and 2.5x TBE buffer) and loaded after into the wells. The electrophoresis was normally carried out at 100-130V for 1 to 2 hours and bands were visualized using a Gel doc apparatus, which uses illumination under UV light. The results were record by a CCD camera incorporated in the instrument.

2.2.2.2 Capillary electrophoresis

To perform capillary electrophoresis, a multi-colour fluorescence-based DNA analysis system with 16 capillaries, the 3100 Genetic Analyser, was utilized (Applied Biosystems, UK). The fluorescently-labelled PCR products were prepared to run on the capillary electrophoresis instrument by mixing 1 to 2μl of PCR products with 10μl HiDi formamide and 2μl of ROX, an internal size standard (GeneScan-ROX500, Applied Biosystems, UK). After this procedure, the PCR products were subjected to denaturation for 3 min at 96°C and were separated on POP-4 polymer (Applied Biosystems, UK). Finally, GeneMapper version 3.0 Software (Applied Biosystems, UK) was used for the genotyping analysis and to collect peak area/height of the fluorescent-dye-labelled PCR products for analysis.

2.2.3 Paralogue ratio test (PRT)

The Paralogue ratio test (PRT) is a comparative PCR based approach that uses a precisely-designed pair of primers to simultaneously amplify a variable repeat unit (test locus) and an additional unlinked reference locus that does not vary in copy number (Figure 6) (Armour et al. 2007). The products are then discriminated by size according to internal sequence differences. Paralogue sequences were firstly used in a PCR based method, Paralogue sequence quantification (PSQ), for the detection of chromosomal aneuploidies (Deutsch et al., 2004). While the PSQ used paralogous genes the PRT uses paralogue sequences, not necessarily from paralogue genes.

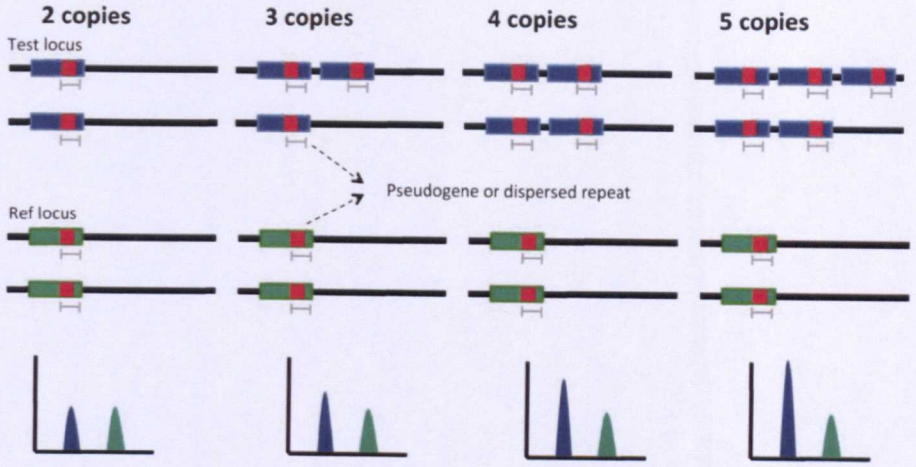


Figure 6: Schematic representation of paralogue ratio test (PRT). The PRT primers (grey segments) simultaneously amplify each copy of test and reference locus. This is only possible due to the presence of a paralogue sequence (pseudogene and dispersed repeat region) within the repeat unit, which is also present elsewhere in the genome but not in a copy number variable region (reference locus). Therefore, after capillary electrophoresis the ratio between the 2 copy reference locus (green) and the test locus (blue) is used to measure the relative number of copies of a test locus (bottom panel).

In order to measure the diploid copy number of β-defensins, three PRT systems were developed showing slight modifications from the first developed PRT, HSPD5.8 (Armour *et al.*, 2007). The PRT primers were designed to amplify from different locations within the β-defensin repeat at 8p23.1. All PRT primers used are shown in Table 2.

Table 2: PCR primer sequences for the three PRT systems developed. The three PRTs were used to amplify different locations in the β-defensin copy number variable repeat.

PRT primer name	Primer sequence (5'-3')	Size resolved in ABI capillary electrophoresis
HSPD5.8	Forward: CCAGATGAGACCAGTGTCC Labelled Reverse: TTTTAAGTTCAGCAATTACAGC	After *Hae*III digestion: Chr8: 302 bp Chr5: 315 bp
PRT107A	Labelled Forward: AGCCTCATTTAACTTTGGTGC Reverse: GGCTATGAAGCAATGGCCTA	Chr8: 157 bp Chr11: 155 bp
HSPD21	Forward (human specific): GAGGTCACTGTGATCAAAGAT Forward (gorilla specific): GAGGTCGCTGTGATCAAAGAT Labelled Reverse: AACCTTCAGCACAGCTACTC	Chr8: 172 bp Chr21: 180 bp

2.2.3.1 HSPD5.8

At the β-defensin cluster, the HSPD5.8 was the first PRT to be developed. This method utilises sequence from a processed pseudogene for the heat-shock protein HSPD, located 2 kb upstream of the *DEFB4* gene and in 10 locations elsewhere in the genome. The designed primers match the copy near *DEFB4* and just one other copy on chromosome 5, with multiple mismatches at the other locations in the genome (Armour *et al.* 2007). The primers therefore amplify only the copy near *DEFB4* (test locus) and the copy on chromosome 5 (reference locus), resulting in product fragment sizes of 443 bp and 447 bp, respectively. However, these products are similar in size, differing just by 4 bp, which makes it difficult to distinguish them by ABI capillary electrophoresis. Therefore, a *Hae*III digestion was performed, producing smaller and more distinguishable products from chromosome 8 (302 bp) and chromosome 5 (315 bp) that can easily be resolved by ABI capillary electrophoresis.

The products were amplified by PCR, performed in 10x PCR mix as described in section 2.2.1.1, for 30 cycles of 95°C for 30 seconds, 53°C for 30 seconds and 70°C for 30 seconds. To reduce the occurrence of single-stranded DNA and allow the complete terminal addition of 3'dA this PCR was followed by a single "chase" phase of 53°C for 1 minute and 70°C for 20 minutes. In order to increase the accuracy and precision of the measurement assay each sample was amplified twice by a duplicate PCR, each performed with a different labelled primer, one with FAM-reverse labelled primer and the other with HEX-reverse labelled primer. 1μl of each PCR product was then digested with *Hae*III, in a 10μl digestion reaction containing 10x ReAct 2 buffer (New England BioLabs) and 5U of *Hae*III and incubated at 37°C for 4 to 16 hours. The use of two distinct labels allows the multiplexing of the duplicate PCR for each sample in the same capillary lane. After *Hae*III digestion, 2μl of digest product was added to 10μl Hi-Di formamide with R-500 marker (Applied Biosystems) and analysed by capillary electrophoresis as described in section 2.2.2.2.

2.2.3.2 HSPD21

The HSPD21 PRT is a modified version of HSPD5.8 PRT and was designed in a way that allows the differentiation between test and reference

locus without the need of a restriction digestion step. This system uses sequence from the same processed pseudogene for the heat-shock protein HSPD used for HSPD5.8 PRT, and a reference locus on chromosome 21. The precisely designed pair of primers is very specific, resulting in products exclusively from test (172 bp) and reference locus (180 bp). Although the HSPD21 primers have counterparts in 8 other locations in the genome, none of these locations has enough sequence similarity to compromise the specificity of the primers to uniquely amplify the target sequences.

The PCR was performed in a 10x Ld PCR mix (section 2.2.1.1) in a duplicate reaction for each sample, with either FAM-reverse labelled primer or NED-reverse labelled primer. The PCR amplification was preceded by a pre-denaturation at 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. The amplification was finalised with a single "chase" phase of 58°C for 30 seconds and 70°C for 40 minutes; 1 µl of each reaction was added to 10µl Hi-Di formamide with Rox-500 marker (Applied Biosystems) and analysed by capillary electrophoresis as described before (section 2.2.2.2).

2.2.3.3 PRT107A

The variable region includes not only the *DEFB4* gene but also other defensin genes such as *DEFB107*. A PRT system was designed from a dispersed repeat region near the *DEFB107* gene that is also present in many other locations in the genome. The primers have counterparts in seven other locations in the genome but match perfectly only with two loci, a region just upstream of the *DEFB107* gene on chromosome 8 and a reference locus on chromosome 11. Although the resultant products have just 2 bp difference in length, 155 bp and 153 bp for test and reference locus respectively, capillary electrophoresis can resolve and distinguish the two products.

The PCR amplification and analysis of PCR products by capillary electrophoresis on the ABI was carried out under the same conditions described for HSPD21 and in section 2.2.1.1 and 2.2.2.2. However, in this case the forward primers were labelled with FAM and HEX dyes.

2.2.3.4 Data analysis

For analysis of PRT data either peak area or peak height of the resultant test and reference locus for each system was used. For HSPD5.8, peak areas of test and reference locus were recorded by the Genescan/Genotyper software (Applied Biosystems) and GeneMapper software (Applied Biosystems), respectively. In the case of PRT107A and HSPD21, GeneMapper software (Applied Biosystems) records peak heights. The peak heights were chosen instead of peak areas for the PRT107A and HSPD21 systems as the ratios given by peak areas, for HSPD21 in particular, were repeatedly higher than the ratios given by peak heights. This was mainly due to the presence of a NED dye peak co-migrating with the test peak of HSPD21, which increased the peak area and consequently the ratios. Thus, peak heights gave more accurate measurements than peak areas.

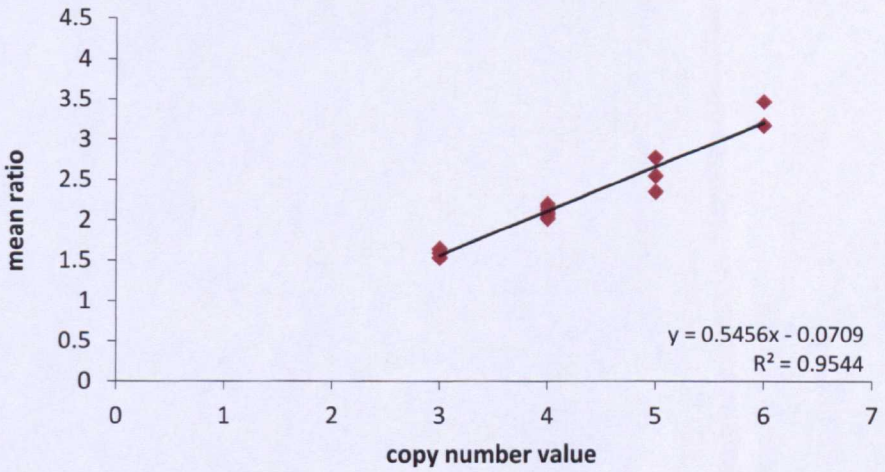
A criterion for peak selection was applied to all peaks of the 4 different PRT systems. Peaks that showed saturation, merged test and reference peaks (only seen in the PRT107A system), split peaks, tails or shoulder peaks that change the peak height or a peak height lower than 50-100, were rejected and not included on the analysis of the copy number.

For the first developed PRT, HSPD5.8, the peak areas corresponding to the 302 bp *Hae*III fragment from the test locus on chromosome 8 and the 315 bp fragment from the reference locus on chromosome 5 were recorded for both FAM- and HEX-labelled products and the ratio between test and reference locus calculated. The ratios of the two duplicates (FAM- and HEX-) were compared and the results were accepted if the difference between the ratios was <15% of their mean. The mean ratios of the accepted tests, which corresponded to about 90% of those attempted, were then used for further analysis.

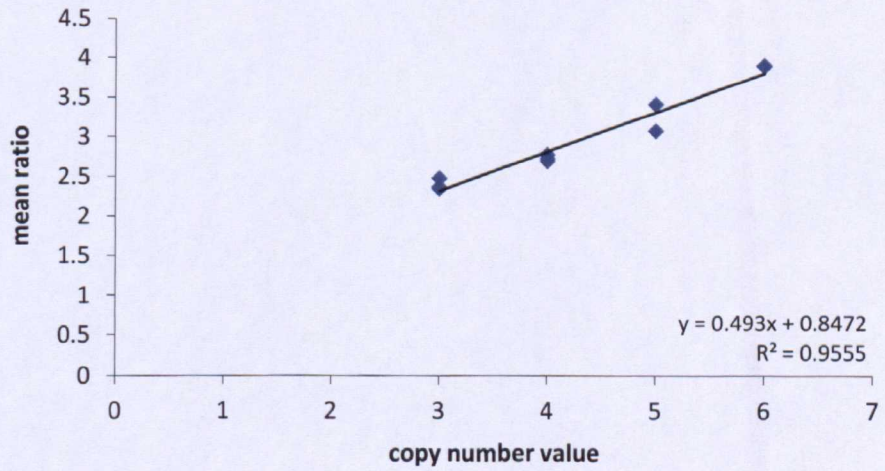
The peak heights of PRT107A and HSPD21 test and reference products were recorded for both dyes and the ratios between the corresponding test/reference pairs of each system were calculated. The mean ratio of HEX-/FAM-labelled products for PRT107A and FAM-/NED-labelled products for HSPD21 were calculated and used for further analysis.

In each experiment using HSPD5.8, PRT107A and HSPD21 selected reference samples with known copy number were amplified simultaneously with the samples for which the copy number was unknown. The mean ratios of the reference samples were then used to calibrate each experiment and the resultant linear regression (least-square) used to infer the copy number of unknown samples. The reference samples were selected based on their reproducibility and agreement of results from numerous experiments either performed with PRT or MAPH/REDVR (Figure 7, a, b and c).

a) Calibration of HSPD5.8 with reference samples



b) Calibration of PRT107A with reference samples



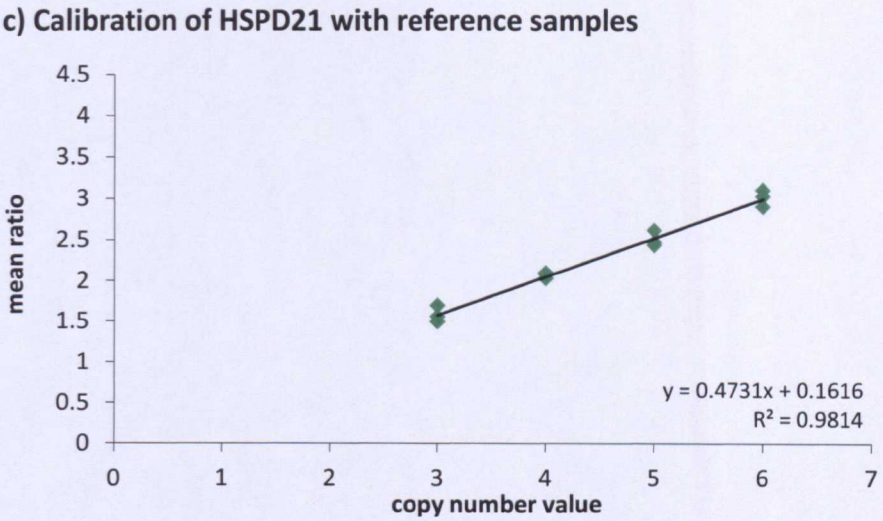


Figure 7: The scatter-plots above show an example of the calibration performed in each experiment with selected reference DNA samples with known copy number for HSPD5.8 (a), PRT107A (b) and HSPD21 (c). The unrounded mean ratio of each reference sample was plotted against the corresponding copy number. The linear regression obtained in each experiment was used to infer the copy numbers of unknown samples.

2.2.4 Indel ratio measurements

Insertion deletion (indel) ratio assays were developed to confirm the β-defensin copy number measured from the PRT assays. The indel assays measure the ratio between the different alleles to predict the copy number. Two indel assays, 5DEL and 9 bp indel, were used to investigate the copy number of β-defensin (Table 3).

Table 3: PCR primer sequences of the two indel assays used. The two indel assays were used to amplify different polymorphisms in the β-defensin copy number variable repeat.

Primer Name	Primer sequence (5'- 3')	Size resolved in ABI capillary electrophoresis
5DEL IV	Labelled Forward: AAACCAATACCCTTTCCAAG Reverse: TTCTCTTTTGTTTCAGATTCAGATG	123 bp to 128 bp
9 bp indel	Labelled Forward: CCAAATGGAAGAATGGCGTA Reverse: GTCCATTGGGTTCTCAAACT	298 bp and 308 bp

2.2.4.1 rs5889219 (5DEL)

Within the β-defensin gene cluster at 8p23.1, about 10 kb upstream of *DEFB107* (chr8: 7,363,859-7,364,008), there are closely adjacent 2 bp and 5 bp

deletions. Taking advantage of the proximity of the two deletions, a pair of primers was designed to span the 2 bp / 5 bp deletions. The ratio between the 3 observed alleles was then used to predict the copy number. This assay was performed together with the two PRT assays to form the Triplex assay.

Genomic DNA (10ng) was amplified by PCR in 10X Ld PCR mix (section 2.2.1.1) for 22 cycles of 95°C for 30 seconds, 58°C for 30 seconds and 70°C for 1 minute. These PCR cycles were preceded by a pre-denaturation at 95°C for 5 minutes and followed by single "chase" phase of 58°C for 30 seconds and 70°C for 40 minutes to complete 3' dA addition. Each sample was amplified in duplicate using either FAM or HEX-reverse labelled primers. Finally, the PCR products were analysed by capillary electrophoresis (section 2.2.2.2) by adding 1µl of each reaction to 10µl Hi-Di formamide with R-500 marker (Applied Biosystems).

2.2.4.2 9 bp indel

A 9 bp insertion/deletion found in the *DEFB4* gene by sequencing (Abu Bakar 2010), was used with the aim to clarify the segregation of copy numbers from parents to children in an 8p23.1 inversion family case study. Two possible alleles, corresponding to deletion or insertion of the 9 bp sequence, were obtained and the ratio measured between them used to predict the β-defensin copy number.

Using the primers listed in Table 3, genomic DNA was amplified with a FAM labelled dye only by PCR with 10X Ld PCR mix (section 2.2.1.1) for 27 cycles of 95°C for 30 seconds, 50°C for 30 seconds and 72°C for 1 minute and then followed by a single "chase" phase of 50°C for 1 minute and 72°C for 40 minutes to avoid the formation of single-stranded DNA. 1µl of PCR product was added to 10µl HiDi formamide with R-500 marker (Applied Biosystems) and examined by capillary electrophoresis as described in section 2.2.2.2.

2.2.4.3 Data analysis

The peak heights corresponding to both FAM- and HEX-labelled PCR products from 5DEL and Hex-labelled PCR products from 9 bp indel were recorded by GeneMapper software (Applied Biosystems).

For the analysis of three or more alleles, such as the three alleles obtained from 5DEL, a program was developed by John Armour (Institute of Genetics, University of Nottingham) to evaluate the copy numbers. To calculate the copy number, the program uses a "squared difference score" between the measured ratio and the expected ratio, as described in more detail in section 2.2.7.2.

2.2.5 Microsatellite measurements

Microsatellites are highly polymorphic and abundant throughout the human genome and due to their co-dominant and easy amplification by polymerase chain reaction (PCR); they have been widely used as genetic markers to identify particular sequences of DNA in the human genome.

Microsatellites within the β-defensin region have also been used to validate copy number measured by PRT assays. Since the number of microsatellite repeats is variable between alleles, these markers can be very informative and be used to study duplications and deletions, but also to dissect the haploid copy numbers and to look at the segregation of those copy number from parents to children. So, to clarify the β-defensin copy numbers, pairs of primers have been designed to genotype three different microsatellites within the β-defensin cluster, EPEV1, 3 and 5, producing diverse product lengths corresponding to different alleles (Table 4). After amplification, the ratio between the peaks corresponding to different alleles is used to predict the copy number.

Table 4: PCR primer sequences for three microsatellite assays used to amplify different regions of the β-defensin copy number variable repeat.

Primer name	Primer sequence (5'-3')	Size resolved in ABI capillary electrophoresis
EPEV1	Labelled Forward: GGCAGTATTCCAGGATACGG Reverse: GAACAATTAGATATCCCTATGC	167 bp to 191 bp
EPEV3	Labelled Forward: GATACTGTGAACTACAGATCAC Reverse: CTGCCCTGATTCAGTATTGAAC	127 bp to 151 bp
EPEV5	Labelled Forward: ACCATTGTTGGTCATTTGTTCTT Reverse: TCAGGCAACTGGACAATCAG	150 bp to 160 bp

2.2.5.1 EPEV1, 3 and 5

Three microsatellites were identified within the β-defensin locus, EPEV1, a variable CT repeat located just downstream of *DEFB106* (chr8: 7725964+7726151) and the two other microsatellites, EPEV3 and EPEV5, both located within the *DEFB107* gene (chr8: 7707791+7707933 and chr8: 7707822-7707975, respectively). EPEV3 is characterized by a variable number of CA repeats while EPEV5 has variable numbers of TA and TG repeats.

The microsatellites were amplified by PCR with a HEX-labelled dye in a 10X Ld PCR mix (section 2.2.1.1). Genomic DNA (10ng) was amplified for 25 cycles of 95°C for 1 minute, 56 to 60°C for 1 minute and 72°C for 1 minute, followed by a single "chase" phase of 72°C for 40 minutes (EPEV3) or 20 minutes (EPEV1 and EPEV5). Annealing temperatures were variable according to the microsatellite to be amplified (56, 58 and 60°C for EPEV3, 1 and 5, respectively). As described earlier, PCR products were analysed on capillary electrophoresis (section 2.2.2.2) by adding 1.5μl PCR product to 10μl HiDi formamide with R-500 marker (Applied Biosystems).

EPEV1 and EPEV3 were resolved in the same capillary electrophoresis, as they have distinct product size ranges of alleles. The EPEV5 analysis was carried out only in the 8p23.1 inversion family case study (section 4.3) and performed in a single capillary electrophoresis, since the PCR product sizes were very close to those of the EPEV3 assay.

2.2.5.2 Data analysis

Microsatellites are widely used as genetic markers and its analysis is normally easy and straightforward. However, due to small differences in allele size, the appearance of non-specific peaks adjacent to the main allele peaks is quite frequent, posing a problem for correct allele-calling that could lead to genotyping error. Sources of error include poor or non-specific amplification, incomplete 3'-dA nucleotide addition and appearance of minor slippage products ("stutter" peaks). Typically 1 to 10 repeat units smaller than the main allele product can appear due to slippage synthesis errors of *Taq* DNA polymerase during PCR elongation. This phenomenon is quite frequent during amplification of di-, tri-

and tetranucleotide microsatellites, generating a characteristic and complex migration profile, which nevertheless can be visually detected if using fluorescent capillary electrophoresis. As such, all microsatellite traces were visually checked to manually correct the allele calls given by the software.

PCR amplifications of EPEV1, EPEV3 and EPEV5 microsatellites were efficient and showed the expected product sizes. In the absence of PCR artefacts the intensities of PCR products from microsatellites could be used directly to infer their likely numbers. However, the formation of minor products was observed in these microsatellites through the presence of "stutter" peaks, usually with about 10-20% of the adjacent main allele peak area/height (Figure 8). For EPEV3 and EPEV5, three main allele peaks were obtained, the other ones being recognized as "stutter" peaks. The EPEV1 can show up to five primary microsatellite alleles and usually two minor products ("stutter" peaks) resultant from PCR slippage.

To analyse the effect of slippage, is important to first recognize if there are any overlap between "stutter" peaks of different alleles. For the microsatellites in question, the alleles overlapped, meaning that PCR slippage created products which were the same size for more than one allele, making it difficult to recognize the original allele and the corresponding slippage product. The commonest case observed is when two adjacent peaks (A and B, in Figure 9a) show only one slippage product (C in Figure 9a). The peak C represents a slippage product from B, but the slippage product from A is not visible because it is "under" B. Therefore, to analyse the effect of slippage (degree of influence) in the two main adjacent alleles a quadratic equation for three peaks was applied: $X=A+1/2(B-\sqrt{(B^2-4AC)})$, $Y=C+1/2(B+\sqrt{(B^2-4AC)})$ where A, B and C are the observed peak areas, and X and Y the corresponding original peak areas, without slippage (Figure 9). The analysis described here only took into account "first-level" slippage products and assumes that all alleles are affected by slippage to the same (proportionate) extent. Since this quadratic equation only analysed the effect of "stutter" peaks in the two adjacent peaks, for microsatellites with more than three peaks, such as the EPEV1, an approximate solution was applied to simulate the original peak areas when 3, 4 or 5 adjacent peaks were present.

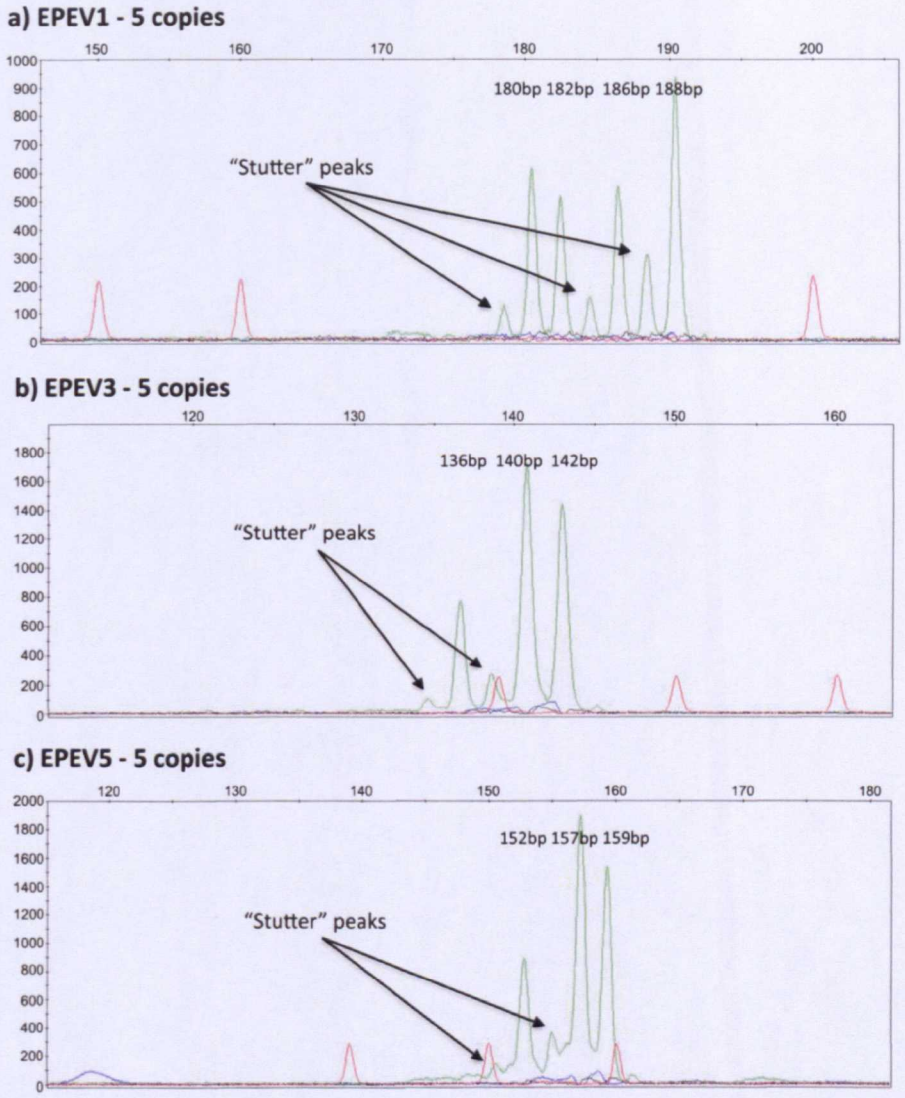


Figure 8: Examples of the GeneMapper electropherogram for EPEV1 (a), EPEV3 (b) and EPEV5 (c) after capillary electrophoresis. In each electropherogram the real alleles and the "stutter" peaks are shown for a sample with 5 copies. The copy number was previously measured by the Triplex assay and confirmed by microsatellite assays. The EPEV1 shows three "stutter peaks" derived from the four primary microsatellite alleles (180 bp, 182 bp, 186 bp and 188 bp). On the EPEV3 and EPEV5 electropherogram three main microsatellite alleles are visible, 136 bp, 140 bp and 142 bp and 152 bp, 157 bp and 159 bp, respectively with two secondary "stutter" peaks for each microsatellite assay.

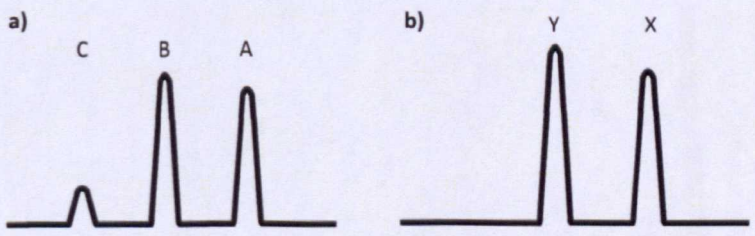


Figure 9: a) Schematic representation of a microsatellite marker trace showing "stutter" peaks. Peaks A and B represent the real alleles and C the "stutter" peak for alleles A and B. b) Hypothetical reconstruction of areas of the original two peaks X and Y after slippage correction, assuming that the two alleles slipped equally.

2.2.6 Restriction fragment length polymorphism (RFLP)

Restriction fragment length polymorphisms (RFLPs) have been used as genetic markers to detect base-substitution variation by differences in the length of restriction fragments. Since restriction enzymes cleave specific sequences, mutations in a sequence can change the specific restriction site of an enzyme creating a new sequence not recognized by the enzyme or, on the other hand, creating a new restriction site. These polymorphisms are identified by changes in the lengths of restriction fragments detected by DNA probes, RFLPs.

One SNP was found in initial analysis to be in linkage disequilibrium with the proximal site (REPP) of β-defensin copy number repeat at the 8p23.1 locus, and therefore it could potentially be used to "tag" the copy number. To investigate this correlation, rs12548700 was genotyped in different samples from ECACC and HapMap cohorts (Table 5).

Table 5: PCR primer sequences designed to genotype one SNP found to be in LD with the β-defensin copy number variable repeat by the Haploview software.

Primer name	Primer sequence (5'-3')	Size resolved in agarose gel electrophoresis	Products sizes after digestion
rs12548700	Forward: GCTTGCCAATTCCAAAGAAG Reverse: CCATGCTTGAAAATGTCTGAATGGtA	228 bp	*Rsa*I: 171, 38 and 19 bp or 146, 25, 38 and 19 bp

2.2.6.1 rs12548700

To genotype the rs12548700 SNP a mismatch in reverse primer was deliberately produced to create a context for *Rsa*I (GT^AC) RFLP, allowing the discrimination between the two alleles (G/C). PCR using these two primers produced a sequence different from the one present on UCSC due to the mismatched reverse primer, creating a third restriction enzyme site if the polymorphic base G was present.

Genomic DNA (10ng) was amplified by PCR in 10X Ld PCR mix (section 2.2.1.1) for 37 cycles of 95°C for 1 minute, 60°C for 1 minute and 70°C for 1 minute. PCR was finalized with a single "chase" phase of 72°C for 20 minutes.

To allow the reverse primer to anneal efficiently, the PCR was performed at 60°C annealing temperature even though the reverse primer had a higher predicted melting temperature. The extra 5 to 6°C of "leniency" allowed the reverse primer to perform efficiently, despite the mismatch. PCR products of 228 bp were digested with *Rsa*I restriction enzyme. On digestion with *Rsa*I, the 228 bp product was cut into 171 bp, 38 bp and 19 bp in every sample. However, if the polymorphic base was a G instead of C, there would be an extra cut to split the 171 bp into smaller fragments of 146 bp and 25 bp. PCR products were separated and visualized in 2% agarose gel electrophoresis (section 2.2.2.1). The smaller fragments, such as the 19 bp, 25 bp and 38 bp were not visible in the agarose gel, so in practice the 171 bp fragment indicated the presence of the C allele (uncut), while the 146 bp fragment corresponded to the G allele (cut).

2.2.7 Triplex system

The Triplex system resulted from the combination of two PRT systems (PRT107A and HSPD21) and the 5DEL system, developed to measure the β-defensin copy number. For each sample, two parallel amplifications were performed using two different dyes, one with a FAM or NED label and the other with HEX or NED label. Since the PCR products, resulting from the DNA amplification using each of these systems, were sufficiently different to be distinguished by capillary electrophoresis, this makes possible to run the three systems together in one capillary. The PRT systems were amplified together in one PCR reaction, while the 5DEL system was amplified in a separate PCR reaction. All PCR reactions were carried out in 10X Ld PCR mix (section 2.2.1.1) using 10ng of genomic DNA following the PCR cycles previously described for PRT systems in section 2.2.3 and for 5DEL system in section 2.2.4.1. Finally, 0.5µl to 1µl of PCR products, resulting from each reaction, were added to 10µl HiDi formamide with Rox-500 marker (Applied Biosystems) and separated by capillary electrophoresis (section 2.2.2.2).

2.2.7.1 Data analysis

For data analysis the peak heights of each labelled product for PRT107A, HSPD21 and 5DEL (155 bp and 157 bp, 172 bp and 180 bp, and 123 bp to 128 bp, respectively) were recorded using GeneMapper software (Applied Biosystems).

Mean copy number of both PRT systems was predicted from the mean ratio of the two labelled products obtained for each PRT, using the calibration equation taken from reference samples. To calibrate each experiment reference samples were used as described in section 2.2.3.4. The unrounded copy number of each PRT and the peak heights of the 5DEL alleles were introduced in a likelihood program (section 2.2.7.2), which inferred the most likely copy number for each sample.

2.2.7.2 Maximum likelihood analysis

A maximum likelihood program, written in C++ was developed by John Armour to evaluate the copy number given by the Triplex system using a likelihood analysis. An input file, including the unrounded mean copy numbers of PRT107A and HSPD21 and the peak heights/areas of the 5DEL system was used to run the program. The program evaluates the probability of the data for each possible copy number from 2 to 9 for each individual PRT and 5DEL systems, scoring the relative probability of the data for each copy number class relative to the highest probability across all copy numbers (to 1) (Table 6). In addition, the program also gives a combined probability, resulting from the combination of the three systems together, with an overall probability that indicates the maximum likelihood copy number (MLCN) for each given sample. Simultaneously, a "minimum ratio score" which indicates the probability of the data for the MLCN relative to the next most likely copy number accompanied each MLCN. This "minimum ratio" value is therefore a good indicator of confidence in the copy number, with best results represented by a higher minimum ratio.

The probabilities, obtained from PRT data for each copy number group, were calculated taking in consideration a normal distribution with a mean on

integer values and a standard deviation of each copy number class for each PRT system based on empirical observations. The 5DEL probabilities, on the other hand, were calculated using a "difference score" $(R\text{-}X)^2$, between the measured ratio (R) and the expected ratio being tested (X). If more than two peaks applied, the program uses the sum of two "difference scores" obtained from the two ratios and takes the reciprocal of this value, so that the "difference scores" are taken as proportional to the probability (the lower the difference score, the higher the probability). Maximum likelihood analysis as described here, relies not just on the accuracy of the standard deviation provided for each copy number class of each PRT system but also assumes that each copy number is equally likely before any test is performed. Despite knowing that the prior probabilities of different copy numbers are not equal in reality, here a conservative assumption of "flat priors" was followed.

Table 6: Output data from the likelihood analysis showing the relative probability of the data at each copy number from 2 to 9 for PRT107A, HSPD21 and 5DEL data. The best copy number for this sample is also shown, as well as the minimum ratio associated with the result observed. The minimum ratio indicates that the copy number of 5 (with the highest probability) is 399.16 times more likely than a copy number of 6 (second highest probability).

Sample C65	Area of peak			Best copy no.
	1	2	3	
5DEL	2474	2341	1330	5
PRT107A	5.27			
HSPD21	5.02			
minimum ratio	399.16			

Test analysis	Relative likelihood values for N=:								Ratio peaks 1:2	Ratio peaks 1:3
	2	3	4	5	6	7	8	9		
Combined	0	8.15E-50	3.74E-08	1	0.002505	3.79E-10	1.01E-11	1.06E-16		
PRT107A	1.71E-135	4.84E-17	0.002397	1	0.372553	0.000406	0.000166	2.89E-06		
HSPD21	9.46E-99	2.24E-32	2.59E-32	1	0.09001	5.78E-06	1.70E-07	4.89E.10		
5DEL	0	0.0747095	0.060208	1	0.07471	0.161496	0.358414	0.07471	1.05681	1.86015

2.2.8 DNA sequencing

To follow the investigation of the β-defensin copy number variation in non-human primates, sequencing analysis was performed for the loci of interest, since very little sequencing information was available for some of the species included in this study, particularly for gorilla. Sequencing analysis was performed for different loci (chromosome 21 and *DEFB103* gene), in order to clarify the primates' DNA sequence or confirm the results given by PRT. PCR and sequencing primers were designed in regions conserved in human, chimpanzee and orangutan in order to increase the chance of amplification of the gorilla and bonobo genomes (unknown sequence).

2.2.8.1 Chromosome 21 locus

In order to investigate the presence of the hspd3 pseudogene (reference locus on human chromosome 21 for HSPD21 PRT) in the gorilla and bonobo genomes and confirm the sequence information in chimpanzee and orangutan for this locus, sequence analysis was carried out in all non-human primates; five gorillas, three chimpanzees, one bonobo and one orangutan.

PCR amplification was performed using forward Chr21F and reverse Chr21R (Table 7) in a 10X Ld PCR mix reaction (section 2.2.1.1). Amplification was carried out for 37 cycles of 95°C for 1 minute, 63°C for 1 minute and 72°C for 1 minute. The amplification was finalised with a single "chase" phase of 72°C for 20 minutes.

Table 7: PCR and sequencing primer sequences used for sequencing of the chromosome 21 locus.

Primer name	Primer sequence (5'–3')	Size resolved in agarose gel electrophoresis
Chr21	Forward: CCAATGCTCACCGTAAGCTTTTGG Reverse: AATCTGACCGTGGCATCACAACC	734 bp (human and chimpanzee) 726 bp (orangutan)

2.2.8.2 *DEFB103* locus

In order to sequence the whole region of the *DEFB103* locus, a pair of primers was used to amplify a 1.8 kb product. Additional internal primers were also designed to complete the sequence analysis of this locus.

Gorilla DNA was initially amplified by PCR using forward DEFB103F and reverse DEFB103R primers (Table 8) in a 10X Ld PCR mix reaction (section 2.2.1.1) to produce a 1.8 kb product. Amplification was carried out for 37 cycles of 95°C for 1 minute, 59°C for 1 minute and 72°C for 1 minute. The amplification was finalised with a single "chase" phase of 72°C for 20 minutes. The PCR products generated by the primary PCR were subsequently amplified in a secondary PCR using two pairs of internal primers, forward DEFB103F2 and reverse DEFB103R3 or DEFB103R4, as described in Table 8. PCR amplification was carried out for 37 cycles of 95°C for 1 minute, 53°C and 57°C for 1 minute and 72°C for 1 minute; and finalised with a single "chase" phase of 72°C for 20 minutes.

Table 8: Primer sequences used for PCR and sequencing of the *DEFB103* locus.

Primer name	Primer sequence (5'–3')	Size resolved in agarose gel electrophoresis
DEFB103	Forward: CCAAGAGAGTGAAGAGTCCAACTT Reverse: GCAAAGTACGGACAAGTCAGC	1832 bp (human)
DEFB103F2/R3	Forward: TTTCTTCGGCAGCATT Reverse: CTTTCCCCAACTCTTCAAGG	591 bp (human)
DEFB103F2/R4	Forward: TTTCTTCGGCAGCATT Reverse: TTGGTCCAAAGCACTCTG	1007 bp (human)

2.2.8.3 Allele-specific PCR at *DEFB103* locus

Allele-specific primers were designed for mixed allelic positions of interest (Table 9) found by sequence analysis at the gorilla *DEFB103* locus.

PCR amplification of the target locus was performed using the appropriate allele specific primers, as described in Table 9, in a 10X Ld PCR mix (section 2.2.1.1). Amplification was carried out for 37 cycles of 95°C for 1 minute, 54°C to 62°C for 1 minute and 70°C for 1 minute, finishing with a single "chase" phase of 72°C for 20 minutes.

Table 9: Allele-specific primer sequences used for sequencing variant positions found at the *DEFB103* locus in J79, Tomoka and Sylvia.

Primer name	Primer sequence (5'–3')
DEFB103R_J79_A	CATTGGAATGATGCATCA
DEFB103R_J79_G	CATTGGAATGATGCATCG
DEFB103F_To.Sy_C	GCATTTTCGGCCACGC
DEFB103F_To.Sy_T	GCATTTTCGGCCACGT

2.2.8.4 Sequencing strategy

Before sequencing, PCR products were purified using the Agencourt AMPure XP PCR Purification system (Beckman Coulter), to remove dNTPs and unincorporated primers. To check that the correct products were amplified, products were separated on a 1.5% agarose gel electrophoresis (section 2.2.2.1). Sequencing reactions were set up in a 10µl total volume with the purified DNA products in a mix with final concentration of 0.5µM primer, 5X sequencing buffer (250mM Tris (pH 9.0) and 10mM MgCl2), standard BigDye® Terminator v3.1 (Applied Biosystem, UK). For each DNA product two sequencing reactions were set up, each with one of the two primers described for each locus, in a high-profile 96 well microplate. DNA products were amplified for 25 cycles at 96°C for 30 seconds, 50 to 58°C for 15 seconds and 60°C for 4 minutes. The conditions applied for amplification, as cycle temperatures and times, were adjusted according to the primers used and the expected products. Finally, products were subject to purification using Agencourt CleanSEQ (Beckman Coulter). Capillary electrophoresis of sequence products was performed at DBS Genomic (Durham University, School of Biological and Biomedical Sciences) using Applied Biosystems 3730 instrument.

CHAPTER 3: DEVELOPMENT OF A MULTIPLEX PRT BASED SYSTEM TO MEASURE THE β-DEFENSIN MULTIALLELIC CNV

3.1 BACKGROUND

In human genetics the identification of genetic variants that contribute to disease has long been a primary aim. The genetic methods used until recently, such as standard cytogenetic methods, Southern blotting and PCR based approaches, were able to identify large heterochromatin polymorphisms (large deletions and duplications) and single nucleotide polymorphisms. Even though many genetic determinants have been identified for common and complex disorders, little was known about intermediate variants, their frequency in the human genome and their importance for disease and human evolution.

With the development of microarray technology, SNP genotyping arrays and next generation sequencing in the last five years, our knowledge of human genetic variation was extensively improved with the discovery of several copy-number variants including some with important roles in disease phenotypes (Iafrate *et al.* 2004; Redon *et al.* 2006; Korbel *et al.* 2007; Kidd *et*

al. 2008; Conrad *et al.* 2010; Handsaker *et al.* 2011). These discoveries presented a completely new perspective on the structural organization of the human genome, highlighting copy number variation as a frequent type of genetic variation with phenotypic consequences, considered as important as single nucleotide polymorphisms (SNPs) (Iafrate *et al.* 2004; Sebat *et al.* 2004; Feuk *et al.* 2006). Copy number variation has been recognized to have an important role in human evolution, genetic diversity between different human populations and in conferring susceptibility to complex disease traits (genomic disorders) (McCarroll and Altshuler 2007; Wain *et al.* 2009; Lee and Scherer 2010; Stankiewicz and Lupski 2010).

Established technologies for copy number typing, such as FISH, array-CGH (Comparative genome hybridization) and QMPSF (Quantitative Multiplex PCR of Short Fluorescent fragments) have been extensively used for CNV detection encompassing between 0 and 3 copies, showing a good level of accuracy (Armour *et al.* 2002; Vaurs-Barriere *et al.* 2006; Carter 2007). However loci highly polymorphic in copy number, such as the β-defensin locus, pose a technical challenge for accurate copy number typing. Other methodologies, such as MLPA and MAPH/REDVR, have been frequently used and proved to be able to determine the copy number of highly polymorphic loci, but they use large amounts of genomic DNA (between 1μg and 100-250ng of DNA, respectively) and are expensive and/or laborious (Hollox *et al.* 2003; Hollox *et al.* 2005; Groth *et al.* 2008). On the other hand, real-time PCR, which has been one of the most extensively used methods for copy number determination in large scale studies, did not demonstrate enough accuracy to distinguish between high copy number values. This led to contrasting results in two independent association studies where the β-defensin copy number was investigated in Crohn's Disease patients (Fellermann *et al.* 2006; Bentley *et al.* 2010). So far, none of the technologies available seems to be an inexpensive, accurate and high-throughput method to measure copy number in large case-control association studies. In order to address copy number variation and the biological importance of β-defensin copy number variation, a new methodology was developed taking into consideration accuracy, cost and applicability to large cohorts.

In this study, a new comparative PCR based approach was developed to measure the copy number of β-defensin at the 8p23.1 locus. Initially adapted from paralogue sequence quantification (PSQ), used for the detection of chromosomal aneuploidies (Deutsch *et al.* 2004), the paralogue ratio test (PRT) uses paralogue sequences located both in the reference and test locus to comparatively measure the copy number of the test locus (Armour *et al.* 2007). The PRT is a quantitative multiplex PCR-based approach that uses a unique pair of primers to amplify from a copy of the paralogue sequence within the variable repeat unit and from a copy on the reference locus, which does not vary in copy number. The resultant PCR products are subsequently discriminated according to size differences in internal sequence. Using a unique pair of primers to amplify test and reference locus reduces problems of accuracy and reproducibility frequently observed in multiplex PCR. Due to the different thermodynamic properties of primers and amplicons, differences in amplification rate between test and reference locus can be observed in multiplex PCR, as for example in real-time PCR, and compromise the accuracy of the measurement. As such, the PRT system should create a great advantage over any multiplex PCR-based method, in particular real-time PCR. Moreover, the use of only a small amount of genomic DNA (10-20ng) and the simple and inexpensive methodology design of PRT, should allow an accurate, rapid and high-throughput copy number typing of β-defensin in large-scale studies.

In the first developed PRT method, HSPD5.8 PRT, the primers were designed to amplify from a heat-shock protein pseudogene (HSPDP3), just ~2 kb upstream of the *DEFB4* gene on chromosome 8, and from a reference copy on chromosome 5 (section 2.2.3.1) (Armour et al. 2007). Additionally, this heat-shock protein pseudogene of ~2 kb is found at 10 other locations elsewhere in the genome. However, HSPD5.8 primers show several mismatches with the other copies of the HSPDP3 pseudogene as shown in Figure 10. Therefore, following careful and precise primer design, a single pair of primers was designed to amplify exclusively the test and reference loci. The method was validated for the copy number measurement of β-defensin by comparison of its results with measurements previously obtained from MAPH/REDVR, MLPA and array-CGH for the same samples

(Armour *et al.* 2007). The HSPD5.8 PRT was comparable in accuracy to these alternative methods and showed a high reproducibility in determining the copy number (~89-95% correct calling) (Armour *et al.* 2007). Moreover, due to its simple format, PRT is a suitable method to apply in large cohorts, showing clear advantages relative to other methodologies frequently used to measure copy number. However, for highly polymorphic copy number loci, as the β-defensin cluster at 8p23.1, the development of further assays that include the multiplex of several systems will improve the accuracy of the copy number measurements.

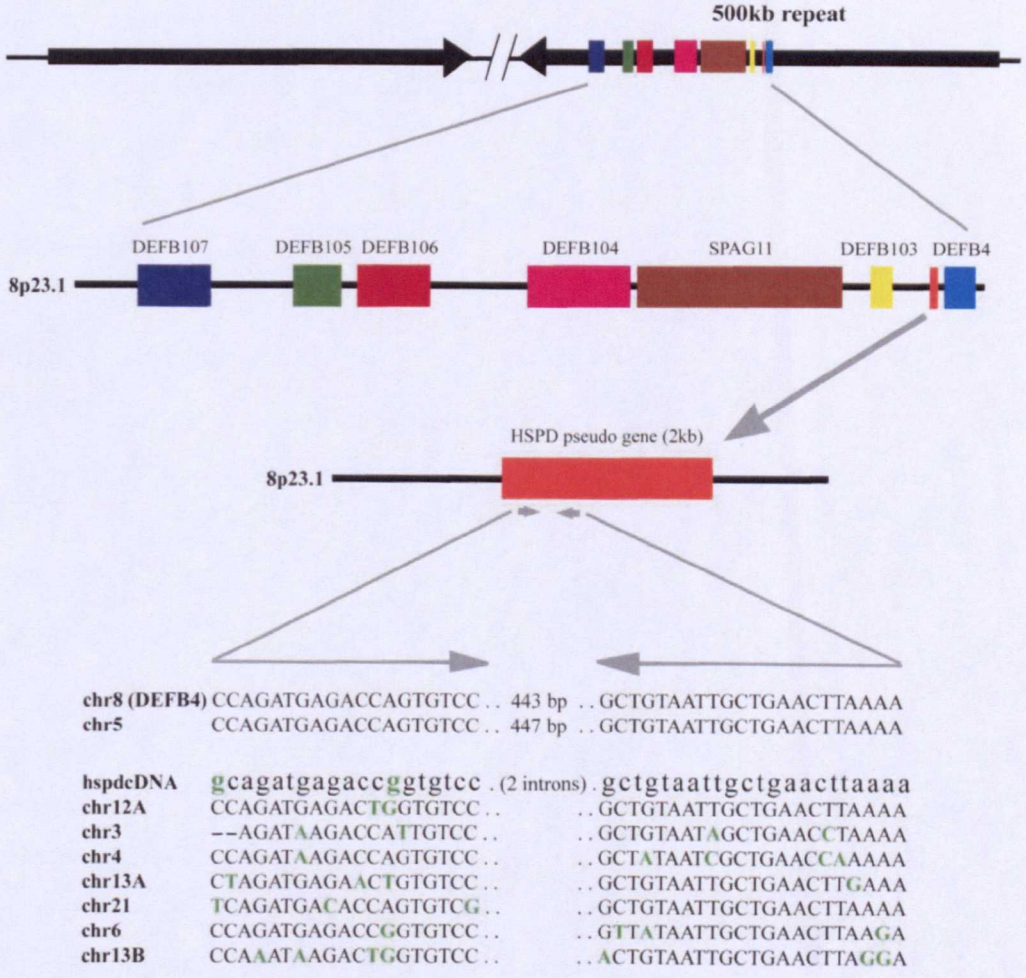
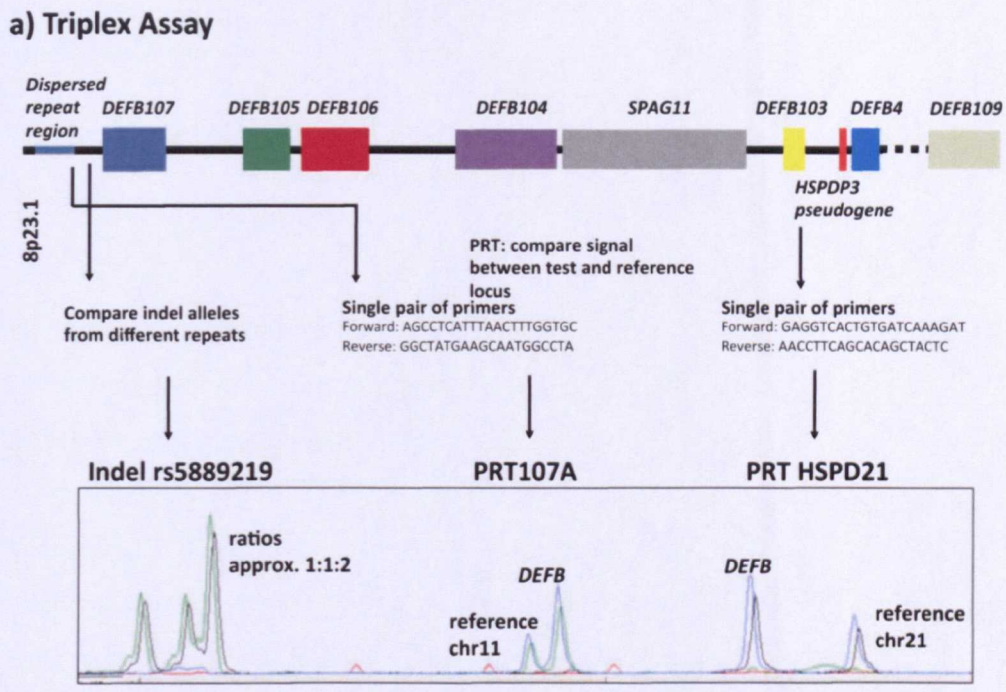


Figure 10: Principle of the HSPD5.8 PRT assay at the 8p23.1 locus. The top line shows the structure of the repeat unit with the black arrows representing two inverted repeats (March 2006 assembly). The middle panel shows the location of the seven different β-defensin genes (*SPAG11*, *DEFB4* and *DEFB103-107*) in each of the inverted repeat units. The bottom line shows the location of the primers used for this assay, just 2-3 kb upstream of *DEFB4* gene, and the amplified PCR products on chromosome 8 (test locus) and chromosome 5 (reference locus), as well as the multiple mismatches with other copies of the HSPDP3 pseudogene in other locations in the genome.

To improve the accuracy and efficiency of the copy number measurement, a Triplex assay composed of two paralogue ratio tests, PRT107A (section 2.2.3.3) and HSPD21 (section 2.2.3.2), and an indel system, 5DEL (section 2.2.4.1) was developed to measure the copy number of the β-defensin cluster at 8p23.1. The idea behind the development of a multiplex system relies on the assumption that typing the copy number of a given sample by more than one method at the same time should improve the results without compromising the time, cost and efficiency of the method. The Triplex assay should consequently increase the accuracy and precision of the copy number measurements and represents a new high-throughput method to apply in large-scale studies. Similar to the HSPD5.8 PRT described earlier, the PRT107A and HSPD21 PRT followed the same principle, which relies in the simultaneous amplification of test and reference locus with just one pair of primers. At ~20.6 kb upstream of *DEFB107* gene lies a dispersed repeat region, which can also be found in other locations in the genome. The PRT107A primers were designed to uniquely amplify from the dispersed repeat region of the test locus on chromosome 8 and from a reference locus on chromosome 11, showing several mismatches with other loci (Figure 11b). Similarly, the design of the HSPD21 PRT primers relied on the presence of the heat-shock protein pseudogene (HSPDP3) near *DEFB4*. As such, primers were designed to amplify each copy of the pseudogene present in the test locus on chromosome 8 and in only one other locus on chromosome 21 (reference locus) (Figure 11a). The primers showed mismatches with other locations in the genome which were sufficiently different in sequence to not interfere with the expected products (Figure 11b). The third component of the triplex system, 5DEL, compares indel alleles from different repeats and works as a verification assay for copy number measured from PRT analysis (section 2.2.4.1.). Taking advantage of a 2 bp and 5 bp deletion only a few base pairs apart and located within the copy number variable region on 8p23.1 10 kb upstream of *DEFB107* gene, the primers were designed to include these two variants. Therefore, three possible alleles can be obtained corresponding to the full repeat (128 bp), the repeat with a 2 bp deletion (126 bp) and the 5 bp deletion sequence (123 bp), and the ratio between the 3 alleles can be used to predict the copy number (Figure 11a). As a result, the triplex trace obtained for each sample showed the PCR products of each

assay as shown on the bottom panel of Figure 11a. Since the products from each assay have different length sizes, the three assays can be multiplexed in the same capillary and analysed without any interference between each other. In the example shown in Figure 11a, an individual with 4 copies of the β-defensin repeat unit is represented. The PRTs show a ratio between reference and test locus of 2:4 and the 5DEL indicates a ratio of 1:1:2. All three assays agree between each other indicating a copy number of 4. Moreover, the results for each sample provided by the three assays were combined and analysed by a likelihood analysis that determines the most likely copy number from 2 to 9 (section 2.2.7.2). The Triplex assay described here was applied to a large cohort of control samples (ECACC panel 1) and the results were compared with previous results obtained with other methodologies for the same samples, to test the accuracy of the method.



b) PRT107A

```
Primers     AGCCTCATTTAACTTTGGTGC.....................TAGGCCATTGCTTCATAGCC
DEFB107A    AGCCTCATTTAACTTTGGTGCctatgctctga...157bp...tgagtaTAGGCCATTGCTTCATAGCC
DEFB107B    AGCCTCATTTAACTTTGGTGCctatgctctga...157bp...tgagtaTAGGCCATTGCTTCATAGCC
Chr11       AGCCTCATTTAACTTTGGTGCctatgctctga...155bp...ggattgTAGGCCATTGCTTCATAGCC

Chr2        AGCCTCATTTAACTTCGGTGCctgtgctctga...157bp...ggattgTAGGCCATTGCTTCATAGCC
Chr4B       AGCCTCATTTAACTTCTGTGTctgtgctctga..........ggattaTAGGCCATTGCTTCATAGCC
Chr13       AGCCTCATTCAACTTTGGTGCctgtgctttga..........ggattgCAGGTCACTGATTCATAGCT
Chr15       AGCCTCATTTAACTTCGGTGCctgttctctga..........ggactgTAGACCAT-GCTTCATATCC
ChrX        AGCCTCATTTAACTTCGCGGTgcctgggctctga.........ggattgTAGGCCATTGCTTCATAGCC
Chr4A       AGCCTCATTTAACTTCAGTACctgtgctctaa..........ggattgTAGGCCATTGCTCCATAGCC
Chr4C       AGCCTCATTTAACTTCGGTGTctgtgctctga..........ggatcaTAGGCCATTGCATTGTAGCC
```

c) HSPD21

```
Primers     GAGGTCACTGTGATCAAAGAT.....................GAGTAGCTGTGCTGAAGGTT
Chr8        GAGGTCACTGTGATCAAAGATtatgcc...172bp...cagatgGAGTAGCTGTGCTGAAGGTT
Chr21       GAGGTCACTGTGATCAAAGATgatgct...180bp...cagatgGAGTAGCTGTGCTGAAGGTT

Chr13A      GAGGTCATTGTGACTAAAGATtatgct..........cagatgGAGTAGCTGTGCTGAAGGTT
Chr4A       GAGGTCATTGTGACCAAAGACgatgcc..........cagatgGAGTAGCTGTGCTGAAGGTT
Chr5B       GAGGTCATTGTGGACCAAAGGTgatgct..........aaaatgGAGTAGCTGTGCTGAAGGTT
Chr12       CAAGTC-CACCGCGCCCAGCCgatgcc..........cagatgGAGTAGCTGTGCTGAAGGTT
Chr3        GAGGTCATTGTGACCAAAGATgatgcc..........cagatgGGATAGCTATGCCAAAGGTT
Chr5        GAGGTCATTGTGACCAAAGACgatgcc..........cagatgGAGTAGTTGTGCTGAAGTTT
Chr2        GAGGTCATTGTGACCAAAGACgatgcc..........cagatgGAGTGGCTGTGCTGAAGGTA
Chr6        GAGGTCATTTTGACCAAATATgatgct..........cagatgGAGTAGCAGTGCTGAAGGCT
```

Figure 11: Principle of the Triplex assay at the 8p23.1 locus. Schematic representation of the Triplex assay (a), showing on the top line the structure of the repeat unit (March 2006 UCSC assembly) with the seven β-defensin genes (*SPAG11*, *DEFB4* and *DEFB103-107*), the dispersed repeat region just upstream of *DEFB107* and the heat-shock protein pseudogene (HSPDP3) ~2 kb upstream of *DEFB4*. The bottom panel represents a typical capillary electrophoresis trace for Triplex assay. Each trace shows the 5DEL, PRT107A and HSPD21 PCR products in two labelled dyes. PRT107A (b) and HSPD21 (c) amplification details showing the amplified PCR products on test (chromosome 8) and reference locus (chromosome 11 and 21), as well as the multiple mismatches with other copies of the dispersed repeat region and the HSPDP3 pseudogene, respectively, elsewhere in the genome.

3.2 RESULTS

3.2.1 Quality control, accuracy and validation

The Triplex assay was firstly applied to a cohort of 96 control samples from ECACC HRC panel 1, which have been previously genotyped by other methodologies, such as MAPH (multiplex amplifiable probe hybridization), REDVR (restriction enzyme digest variant ratio), MLPA (multiplex ligation probe amplification) and array-CGH (comparative genomic hybridization) (Armour *et al.* 2007). The use of samples with known copy number allowed a comparison between the results produced by the Triplex assay and the copy number values established by the previous methodologies. By using this comparison it was possible to test the accuracy of the Triplex assay for copy number measurement. The Triplex assay showed a miscalling rate of just 5.26%, meaning that in ~ 95% of the samples the Triplex yielded the correct integer copy number value. To undertake the genotyping of this cohort, four reference samples selected from the ECACC HRC panel 1 (C11, C18, C62 and C66) were used as internal controls to calibrate each experiment. The reference samples were selected not only because they showed reproducible results for multiple assays, including PRT assay, but also they represent copy numbers of 3 (C11), 4 (C18), 5 (C62) and 6 (C66). For these samples, which we had prior knowledge of the copy number using other methods, such as MAPH, and concordance of PRT data with other results demonstrated that the reference locus was not variable in copy number, showing a constant diploid copy number of 2 (Figure 7, section 2.2.3.4). The remaining ECACC HRC panel 1 samples (92 samples) were genotyped several times, to test the reproducibility of the method, in 96 well plates that included the reference samples and blanks. For each sample/repeat the peak area of each allele was used to calculate the PRT ratios obtained for FAM- and HEX/NED-labelled products. PRT ratios were then calibrated against the ratios of reference samples to predict the copy number of those samples.

To test the concordance and accuracy of the Triplex assay, the unrounded copy number values from PRT107A and HSPD21 were plotted in a scatter plot and the level of clustering was used to evaluate the performance of the PRT

systems. As shown in Figure 12, different copy number clusters were formed around integer values. These clusters correspond to different copy number groups and are clearly separated from each other, showing the ability of this assay to accurately measure copy numbers up to 10 copies. On the other hand, the clusters centred at integers also demonstrate the concordance in the copy number measured by each PRT for the same sample, giving a clear perspective of the extent of variation between independent measurements of the same sample with different PRT systems.

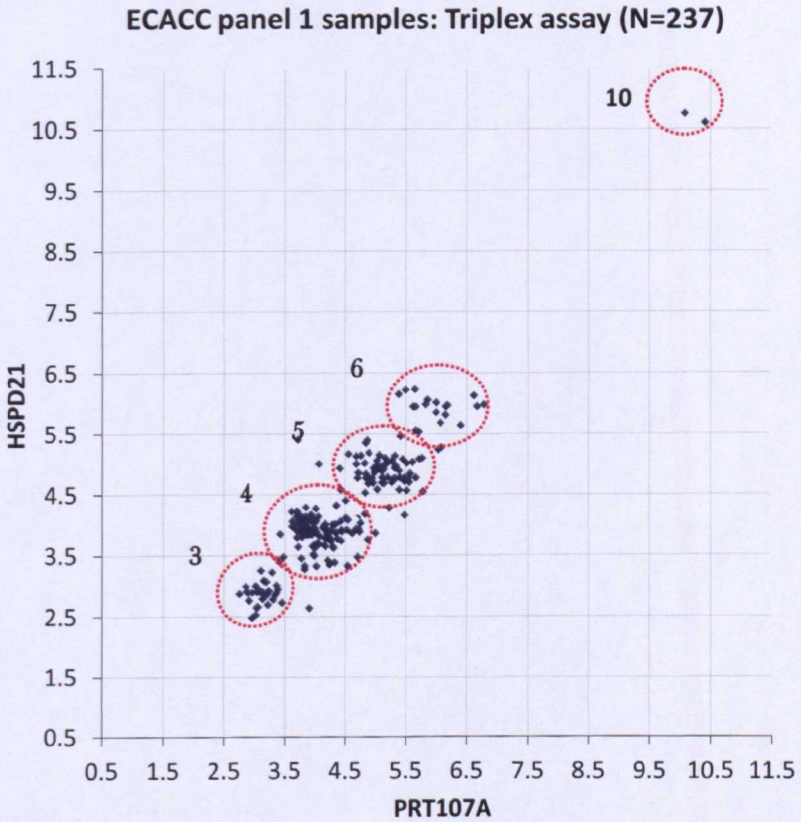


Figure 12: Unrounded copy number values of PRT107A and HSPD21 for ECACC panel 1 samples typed by the triplex assay. The scatter-plot shows clear clusters centred in integer values, highlighted by the red circles (drawn by hand), that correspond to different copy numbers classes. In this plot, the PRT107A values are more spread than the HSPD21 copy number values. (The 96 ECACC panel 1 samples were typed several times with the Triplex assay making a total of 237 different repeats).

The same degree of clustering is also demonstrated by the mean copy number distribution of PRTs and the HSPD21 PRT alone (Figure 13b and c). In the histograms b) and c) of Figure 13, the unrounded copy numbers are distributed around integer values showing either a gap or a very low frequency between the different copy number classes. The degree of clustering observed

for HSPD21 and the mean of PRTs was not reproduced for PRT107A (Figure 13a). Although the unrounded copy number values of PRT107A tend to approximate to integer numbers, the clustering for 4, 5 and 6 copies is not so defined, with higher than expected frequencies of intermediate values. The mean copy number data (c) shows the lower standard deviation (0.046), calculated from the normalized mean unrounded copy number values, which support the accuracy of the measurements obtained from the Triplex assay and shows that the combination in Triplex of different systems is the most accurate way to measure the beta-defensin copy number variation.

The analysis of copy number values is based in the assumption that only integer values represent the real biological copy number. This assumption disregards the existence of any somatic mosaicism, and therefore the differences observed between the measured values and the closest integer values (ML CN), denoted copy number residuals, should represent the error of the method and can be used to evaluate the accuracy of the Triplex assay. To test this hypothesis the copy number residuals for PRT107A and HSPD21 were calculated from the final copy number value given by the likelihood analysis (ML CN) and plotted in a scatter plot as shown in Figure 14. The residuals of both PRTs tend to be close to zero, which indicates that the majority of the copy number values obtained from the Triplex assay were correctly called and the lack of correlation suggests that no mosaicism was observed. The closer the copy number residuals are to zero, the more accurate the measurements and the system. In the presence of somatic mosaicism, the residuals would have correlated distributions, instead of independent distributions, indicating a correlation between the PRT107A and HSPD21 residuals. In this case no significant correlation was found between the PRT107A and HSPD21 residuals (p-value=0.6719, Pearson's correlation test), which provides no evidence for somatic mosaicism. The PRT107A residuals vary between -0.51 and 1 (average residual 0.126) from the rounded copy number given by the ML CN, while the residuals from HSPD21 vary between -0.91 and 0.75 (average residual -0.15). The range of variation for the HSPD21 residuals is wider than for PRT107A, but the HSPD21 frequently gave more copy number residuals closer to zero as shown in the histogram (Figure 15).

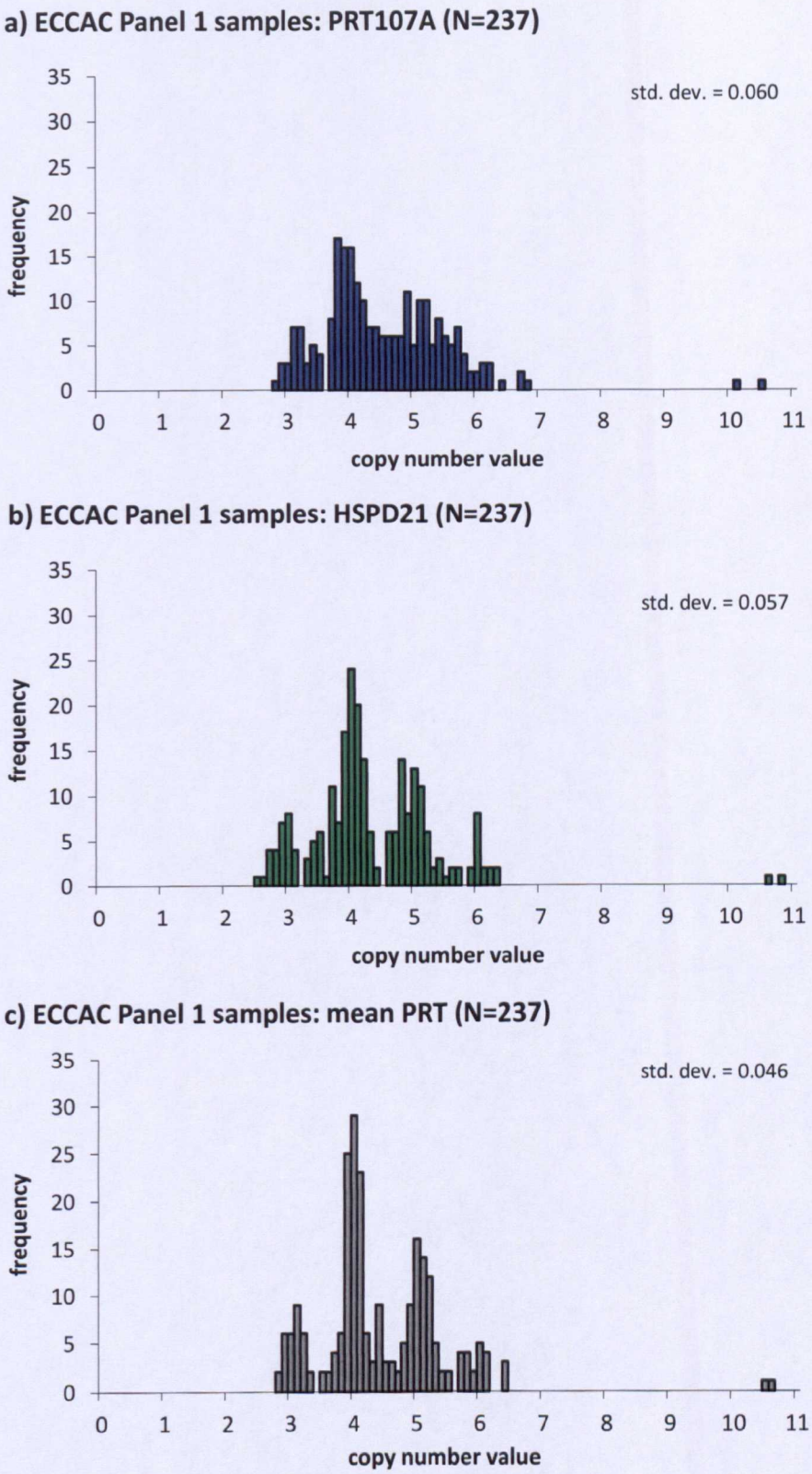


Figure 13:Histogram of unrounded copy number distribution of PRT107A (a), HSPD21 (b) and the mean of PRT107A and HSPD21 (c), showing a clustered distribution centred in integer values (the 96 ECACC panel 1 samples were typed several times with the Triplex assay making a total of 237 different repeats). The standard deviation shown for each assay was calculated using normalized copy number values from PRT107A, HSPD21 and the mean PRT, respectively.

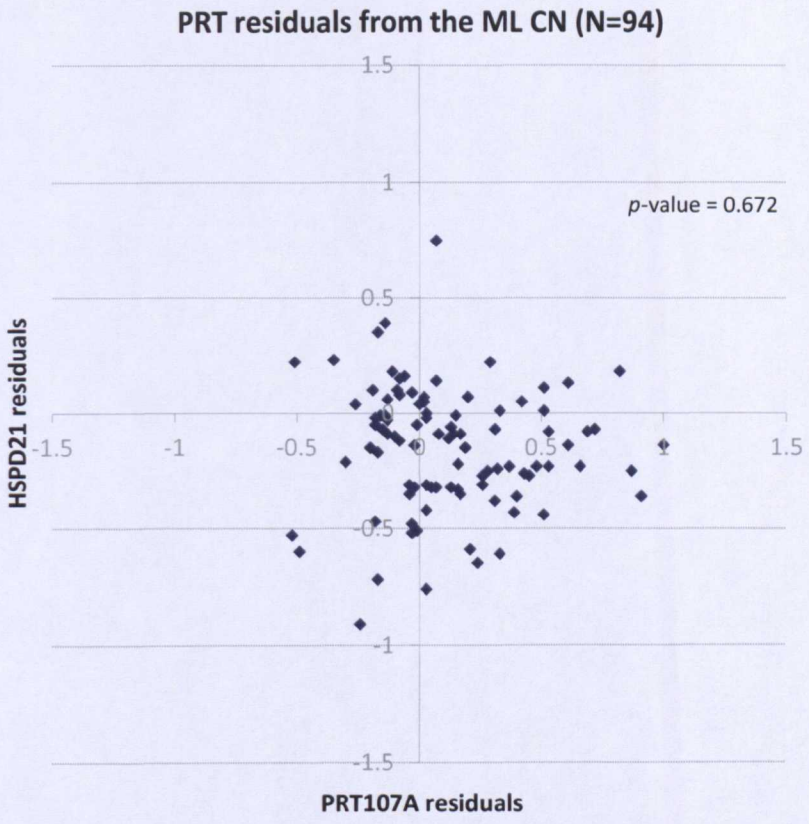


Figure 14: Comparison of copy number residuals of PRT107A and HSPD21 from ECACC HRC panel 1 data. The residuals were calculated from the difference between the measured unrounded copy number value of each PRT and the integer copy number (ML CN) given by the maximum likelihood program for 94 samples.

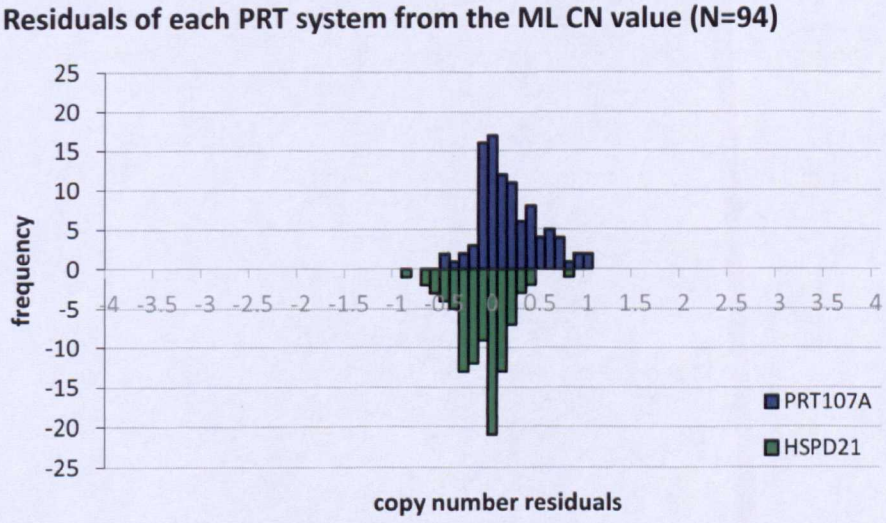


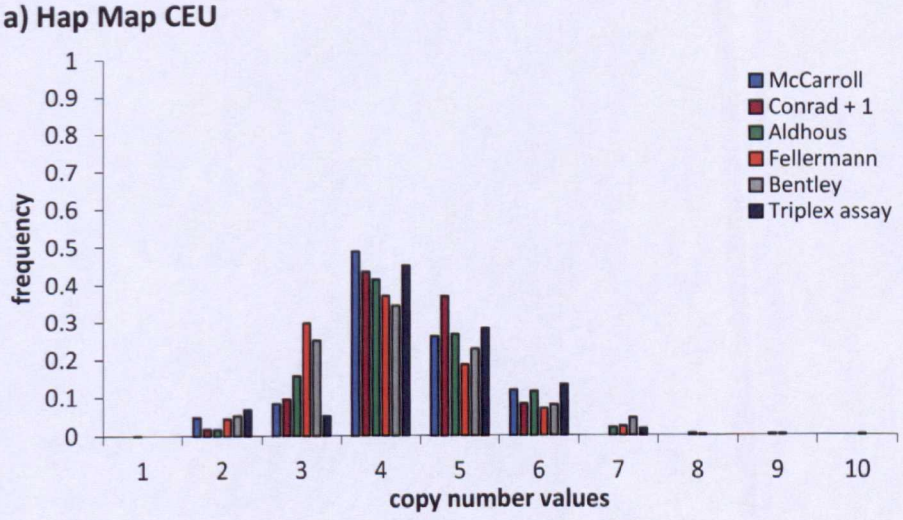
Figure 15: Histogram showing the frequency distribution of copy number residuals from PRT107A and HSPD21 performed in Triplex assay for the ECACC panel 1 HRC panel 1 samples. The residuals were calculated from the difference between the measured unrounded copy number value of each PRT and the integer copy number (ML CN) given by the maximum likelihood program.

Further validation of the Triplex assay was carried out through an extensive comparison with the publicly-available data generated from a hybrid genotyping array featuring more than 1.8 million genetic markers (Affymetrix SNP 6.0) and tiling oligonucleotide microarrays comprising 42 million probes, by McCarroll *et al.* (2008) and Conrad *et al.* (2010), respectively. This data comprises three HapMap collection cohorts, CEU (CEPH individuals with northern and western European ancestry from Utah, USA), YRB (Yoruba from Ibadan, Nigeria) and CHB/JPT (Chinese from Beijing, China and Japanese from Tokyo, Japan). The table below shows the number of samples typed and the percentage of agreement between the copy number values generated for the three HapMap cohorts by the Triplex assay and by the two other studies (Table 10). From a initial comparison of either Triplex assay or McCarroll *et al.* (2008) data with Conrad *et al.* (2010) copy number values, a very low degree of agreement was observed for any of the populations (between 0% and 3.80%). This suggested that copy number values were systematically called with a one copy number value offset due to an original copy number miscalling of the reference genome used in the array-CGH based approach by Conrad *et al.* (2010). Therefore, further comparisons were carried out based in this assumption and Conrad *et al.* (2010) data referred as Conrad +1. As such, the Triplex assay showed a high percentage of agreement for any of the three populations (between 86.08% and 93.51%) with both of the studies, McCarroll *et al.* (2008) and Conrad +1 (Conrad *et al.* 2010).

Table 10: Comparison of copy number data from Triplex assay with data generated by McCarroll et al. (2008) and Conrad et al. (2010) for three HapMap collection cohorts: CEU (CEPH individuals with northern and western European ancestry from Utah, USA), YRB (Yoruba from Ibadan, Nigeria) and CHB/JPT (Chinese from Beijing, China and Japanese from Tokyo, Japan). Number of samples typed in common between studies and percentage (%) of agreement is indicated for each population. Conrad +1 refers to the Conrad et al. (2010) data with the increment of one copy number value for each sample.

	CEU		**YRB**		**CHB/JPT**	
	Total typed	% of agreement	Total typed	% of agreement	Total typed	% of agreement
McCarroll vs. Conrad	80	0	74	0	69	1.45
McCarroll vs. Conrad +1	80	100	75	93.33	70	97.14
McCarroll vs. Triplex	84	92.86	84	89.29	77	93.51
Conrad vs. Triplex	86	1.16	79	3.80	82	3.66
Conrad + 1 vs. Triplex	86	93.02	79	86.08	82	90.24

The frequencies obtained for each copy number class for these HapMap cohorts were also compared with frequencies previous published by different studies. The first histogram (Figure 16a) reports the β-defensin copy number frequency distribution from seven independent studies carried out in the European HapMap cohort, CEU. β-defensin copy number frequency distributions were very similar between studies, with the exception of the frequencies observed for 3 copies, in which Bentley *et al.* (2010), Fellermann *et al.* (2006) and Aldhous *et al.* (2010) showed higher frequencies than the other studies. Regarding the Yoruba and Chinese/Japanese HapMap cohorts (Figure 16b and c), copy number frequency distribution given by Triplex assay was plotted together with frequency distributions obtained from McCarroll *et al.* (2008) and Conrad *et al.* (2010). In Yoruba, similar frequency distributions were shared between the three studies for copy numbers of 3 and 4 and between Triplex assay and McCarroll *et al.* (2008) for copy numbers of 5 and 6. Comparable frequencies were also observed between all studies for copy numbers of 2, 3 and 4 in the CHB/JPT cohort, but only Triplex assay showed frequencies for high copy number values up to 7 copies.



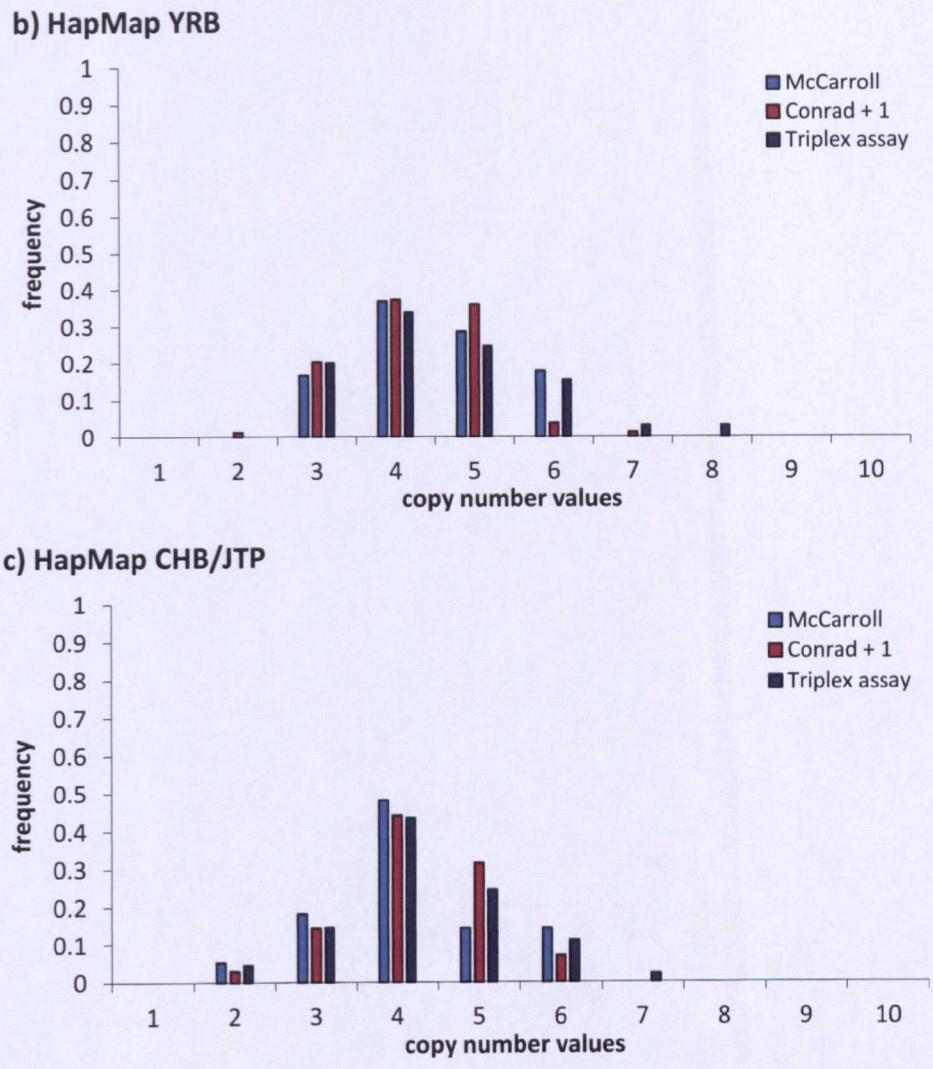


Figure 16: β-defensin copy number frequency distribution in CEU (CEPH individuals with northern and western European ancestry from Utah, USA) (a), YRB (Yoruba from Ibadan, Nigeria) (b) and CHB/JPT (Chinese from Beijing, China and Japanese from Tokyo, Japan) (c) HapMap collection samples. All histograms show frequency copy number distribution given by McCarroll *et al.* (2008), Conrad + 1 (Conrad *et al.* 2010) and the Triplex assay. Additionally, in the CEU histogram copy numbers frequencies from Aldhous *et al.* (2010), Fellermann *et al.* (2006) and Bentley *et al.* (2010) are also represented.

3.2.2 Data characterization

To analyse the data, given by the Triplex assay, a likelihood function was used to calculate the probability of each copy number given the measurements obtained, assuming a normal distribution for PRT data (section 2.2.3.4). Therefore, here it was tested if the copy number values given by the Triplex assay PRT results followed a normal distribution using quantile-quantile plots (Q-Q plots). A Q-Q plot is a probability plot, which allows the comparison of two probability distributions by plotting their quantiles against each other. In

this case, the normalized CN data set was plotted against a standardized normal distributed data, to determine how well the CN data set of values fitted a normal distribution (theoretical model) (Figure 17).

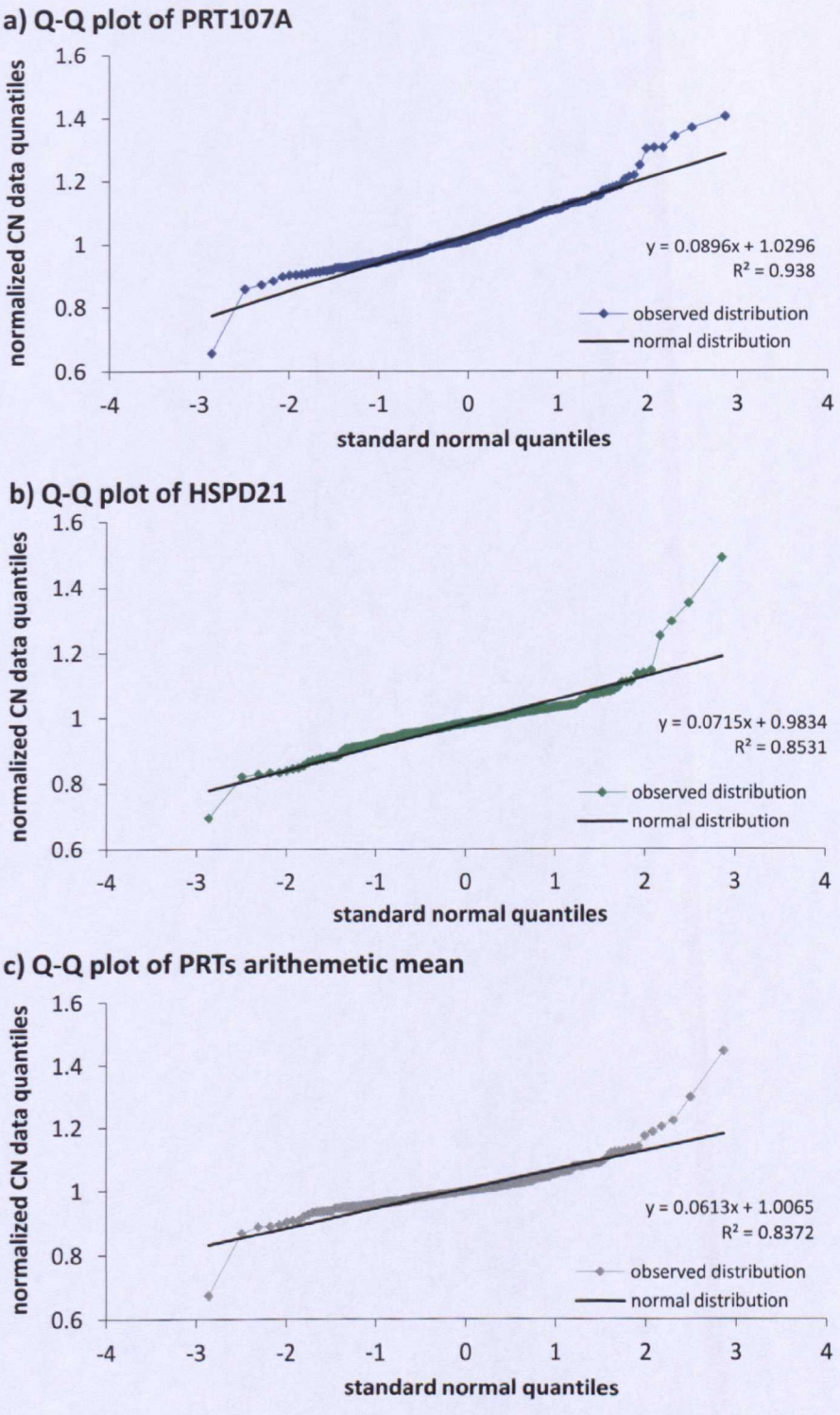


Figure 17: Quantile-quantile plots of PRT107A (a), HSPD21 (b) and arithmetic mean of PRTs (c) comparing normalized CN distribution on the Y-axis to a standard normal distribution on the X-axis (theoretical normal distribution). The linearity of the points suggests that the data are normally distributed.

Therefore, if the two distributions being compared were linearly correlated, the points in the Q-Q plot would approximately lie on the linear regression line, as shown in the plots for PRT107A, HSPD21 and the mean of PRTs.

To further evaluate the performance of each PRT system in measuring different copy numbers, the standard deviation and "integer error" (incorrect classification of an integer) of PRT107A, HSPD21 and the mean of both PRTs were calculated for each copy number value (Table 11). The "integer error" described here represents, for each copy number class, the cumulative probability of an observed value occurring below or above ±0.5 decimal points around each integer value, taking into account its mean and standard deviation. As a result, it gives the predicted probability of the method assigning the wrong copy number. As shown in the table below, the standard deviation and "integer error" of 6 copies is the highest in any system. Moreover, the PRT107A is the system that exhibits the highest standard deviation and "integer error" for almost all the copy number groups. On the other hand, the mean PRTs showed the lowest standard deviation and "integer error" for all copy number values analysed.

Table 11: Standard deviation (std. dev.) and "integer error" of each copy number from 3 to 6 for PRT107A, HSPD21 and the mean PRTs. The values were calculated from the copy number values obtained from Triplex assay for ECACC HRC panel 1 cohort.

CN	PRT107A		HSPD21		Mean PRTs	
	std. dev.	integer error	std. dev.	integer error	std. dev.	integer error
3	0.29	0.16	0.27	0.07	0.24	0.04
4	0.40	0.22	0.35	0.16	0.28	0.08
5	0.34	0.18	0.25	0.06	0.23	0.03
6	0.60	0.41	0.47	0.32	0.48	0.31

Finally, the unrounded copy number of each PRT system (PRT107A and HSPD21) and the peak area of each allele from the 5DEL assay were analysed by a likelihood function which gave the most likely copy number along with a index of confidence (minimum ratio) for each sample (section 2.2.7.2). If samples were typed more than once, the integer copy number was selected based on the higher minimum ratio. The β-defensin copy number for 94 samples from the ECACC HRC panel 1 cohort is shown in Figure 18. As

shown in the histogram, the β-defensin copy number varied between 3 and 10 copies with a mean copy number of 4.43 copies. The copy number of 4 is the most common and is present in about half of the samples typed.

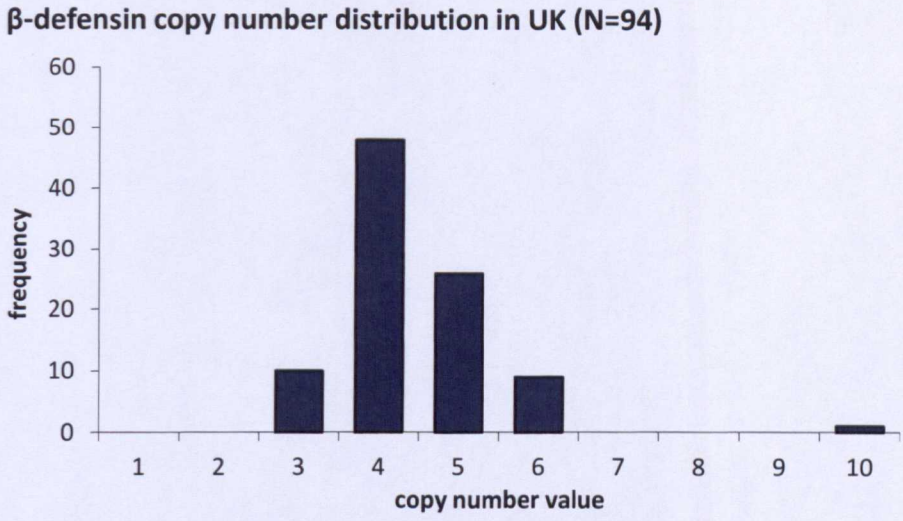


Figure 18: β-defensin frequency copy number distribution in a UK population cohort (ECACC HRC panel 1). The histogram shows integer copy numbers obtained from the Triplex assay. The copy number for each sample was defined by the result with a higher minimum ratio.

3.3 DISCUSSION

The present study aimed to develop an accurate and robust methodology to measure β-defensin copy number variation in large case-control association studies. The PRT-based Triplex assay, developed here, was able to consistently distinguish between different copy number classes, as demonstrated by the analysis of clustering from the repeated typing of reference samples (C11, C18, C62 and C66 from the ECACC HRC panel 1). The formation of clusters around integer values, separated by gaps, was observed both for the reference samples and for all other ECACC HRC panel 1 samples and was a clear indication of the accuracy of the Triplex assay. This clustered distribution of the unrounded copy number values suggested that the copy number values measured by the Triplex assay represented the real copy numbers and the outlying measurements represented the error of the method. The hypothesis presented here was also supported by the analysis of residuals, which showed no significant correlation between PRT107A and HSPD21 residuals suggesting that either mosaicism is not frequent or is not very extensive. As such, if two independent measurements of the same sample give a different outcome, this would be more likely due to a random error than due to somatic mosaicism. Nevertheless, a small degree of mosaicism cannot be ruled out, since the analysis used here is not able to detect this degree of mosaicism. To achieve that, it would be necessary to use mathematical modelling to predict the amount of mosaicism that would be represented by a significant *p*-value. Finally, to overcome any possible difference in accuracy between experiments, the calibration of the copy numbers given for each sample was carried out against internal reference controls, thus guaranteeing an overall uniformity along different experiments.

The Triplex assay was validated by comparisons with results from three other established methodologies for the ECACC panel 1 cohort, showing a low error rate of 5.26%. Further validation, carried out in three HapMap collection cohorts, comparing the Triplex assay with array-CGH based approaches again confirmed the accuracy of the Triplex assay due to the high level of agreement observed on this analysis (>86%). The Triplex assay therefore proved to have comparable accuracy to very well established methods, such as MAPH/REDVR, MLPA and

array-CGH. The Triplex assay combines the ability of a high-throughput method with the advantages of a low cost and simple methodology, a combination of characteristics that is not present in any of the other methodologies.

The calculated standard deviation and the estimated rates of "integer error" of each copy number class indicated that highest copy numbers (6 copies) presented the greatest "integer error"; a result already expected for a method that relies on the measurement of ratios. These parameters were independently calculated for each PRT system (PRT107A and HSPD21) and for the mean of the two PRTs. The values obtained for standard deviation and rates of "integer error" for each system demonstrated that PRT107A was the less accurate. A possible explanation for this is the co-amplification of the reference locus with a further locus on chromosome 2 that has only one mismatch with the forward primer sequence. In this case, the reference locus shows a higher yield resulting in an apparent lower copy number from the PRT107A system. However, this phenomenon did not affect the other two assays (HSPD21 and 5DEL). On the other hand, the combination of two PRTs provided the higher accuracy and precision in the measurement of β-defensin copy number variation, demonstrated by the lowest values of standard deviation and rates of "integer error". Moreover, as it would be expected from the multiplex of PRT systems, the PRT-based Triplex system developed here showed a lower error rate (5.26%) when compared with a single PRT system (HSPD5.8, ~7%) (Armour *et al.* 2007), which supports the initial idea for the development of a multiplex system. The Triplex assay increased the power of copy number measurement, which is especially important in studying multiallelic copy number loci. A Multiplex PRT, combining three different PRT systems was first used to measure the *CCL3L1/CCL4L1* copy number variant genes (Walker *et al.* 2009). In this study the multiplex PRT also showed accurate and reproducible data in large scale genotyping, assigning all samples to discrete copy number classes (Walker *et al.* 2009).

Amplification failure of any of the two copies of the reference locus, due to primer mismatch or a copy number variant, would lead to an apparent increase in copy number of test locus for the system affected (Armour *et al.* 2007). The failure to amplify a variant sequence from either test or reference locus or the

possible variation in copy number of the reference locus would lead to the disagreement of a particular system with the other two assays that compose the Triplex assay. However, this effect was not observed, as neither of the PRT systems showed a copy number difference of exactly twice the value of consensus copy number, so there is no evidence to suggest lack of amplification of one or more copies of the reference locus. The degree of concordance between the systems also suggests that the presence of SNPs and/or copy number variants in the reference loci is not frequent among the European population.

The Triplex assay has now been successfully applied by other groups to investigate the β-defensin genomic copy number variation in different populations (Fode *et al.* 2011). In this study the Triplex assay was used alongside real-time PCR and pyrosequencing-based paralogue ratio test (P-PRT) to type β-defensin copy number. Interestingly, the triplex assay appeared to be the most precise and accurate method shown by the improved clustering of this assay when compared with the other two methods.

The applicability of PRT and of a PRT-based Triplex assay can be expanded to other loci. Though the design of PRT is dependent on the existence of paralogous sequences, these can be often found in the human genome and used to design PRT assays to investigate other copy number variable regions. PRT systems have been used to measure the copy number variation of the *DEFA1/DEFA3* genes (α-defensins) at the 8p23.1 locus (Fayeza Khan, personal communication), the *CCL3L1/CCL4L1* copy number variant at the 17q12 (Walker *et al.* 2009), the complement *C4* gene (in the class III region of the major histocompatibility complex, MHC) on the short arm of chromosome 6 (Fernando *et al.* 2010) and the immunoglobulin-receptor genes *FCGR3A* and *FCGR3B* on chromosome 1q23.3 (Hollox *et al.* 2009). Apart from these loci many others have closely linked paralogous sequences, which are present at constant copy number, that can be used as reference loci in PRT systems to measure for example the *RHD*, *CYP2D6* and *PRB1* copy number variable genes (Armour *et al.* 2007). Another possibility is the use of an unlinked reference locus, as described previously for HSPD5.8 and HSPD21 PRT which rely on the presence of a pseudogene (HSPDP3) near *DEFB4* gene to allow the design of these PRT systems. In addition to *DEFB4,* 18 other copy

variable loci were found to harbour an unlinked reference locus which allows the design of a specific PRT assay, such as for *UGT2B17* or *SMN2* genes (Armour *et al.* 2007). The PRT applicability is relatively broad and it was suggested that at least 50% of copy number variable genes allow the design of PRT systems (Armour *et al.* 2007).

The PRT-based Triplex assay described here showed accurate and reproducible results for about one hundred European control samples (ECACC HRC panel 1) and concordance between independent PRTs confirms that the reference loci are not variable in copy number, supporting the idea that this is a high-throughput system suitable to apply in large-scale case-control association studies. However, the accuracy described here for the PRT-based Triplex system needs to be demonstrated in larger genotyping studies, an essential requirement to evaluate the performance and reliability of a copy number measurement methodology. Recently, this was demonstrated in two studies carried out in large case-control association studies using the PRT-based Triplex assay to measure β-defensin copy number variation in Crohn's Disease (Aldhous et al. 2010) and psoriasis, which will be reported in chapter 4.

CHAPTER 4: APPLICATION OF THE TRIPLEX SYSTEM

4.1 β-DEFENSIN CNV AND PSORIASIS: A CASE-CONTROL ASSOCIATION STUDY

4.1.1 Introduction

The human β-defensins are secreted antimicrobial peptides with cytokine-like properties (Ganz 1999b; Yang *et al.* 1999; Niyonsaba *et al.* 2004). The β-defensins are secreted by leukocytes and a variety of epithelial tissues, such as skin, lung, gastric antrum, oral and nasal mucosa, cornea and urogenital tract (Schutte and McCray 2002; Ganz 2003). They are part of the innate immune system acting as the first line of defence against several pathogens, such as bacteria, virus and fungi (Ganz 2003; Klotman and Chang 2006). Taking into consideration the role of β-defensins in innate immunity, the variation of β-defensin gene copy number suggests variability in gene dosage that ultimately could contribute to the susceptibility to infectious and inflammatory diseases (Fellermann *et al.* 2006; Bentley *et al.* 2010).

In 2005, Hollox *et al.* investigated association between β-defensin copy number variation and severity of cystic fibrosis, an autosomal recessive genetic

disease associated with a frequent lung infection caused by chronic bacterial colonisation of the lungs. The human β-defensin 2 peptide (encoded by the gene *DEFB4*) is expressed in the lung airway and due to its antimicrobial activity, the *DEFB4* gene was identified as a possible candidate gene for modifying cystic fibrosis severity. However, the study revealed no significant association between severity of cystic fibrosis and β-defensin copy number variation in 355 patients with the disease (Hollox *et al.* 2005).

The β-defensin genomic copy number, in particular of the *DEFB4* gene, has been associated with Crohn's disease. However, this is not without controversy (Fellermann *et al.* 2006; Bentley *et al.* 2010). While in the Fellermann *et al.* study in 2006, Crohn's disease of the colon was associated with low *DEFB4* copy number, a more recent study by Bentley *et al.* (2010) found an opposite association with higher *DEFB4* copy number. The first study suggests that the antimicrobial properties of defensins protect the intestinal mucosa against bacteria. As such, a low copy number of these defensins could provide a lower level of protection against microorganisms and cause chronic inflammation, characteristic of Crohn's disease (Fellermann *et al.* 2006). Bentley *et al.* (2010) on the other hand suggest that Crohn's disease has an autoimmune effect and that the high copy number of β-defensins could increase the expression of antimicrobial peptides resulting in inflammation of the intestinal epithelium (Bentley *et al.* 2010). Even though it has been proposed that β-defensin genes play a role in susceptibility to Crohn's disease, a recent study carried out with PRT-based methods failed to replicate any association with β-defensin copy number variation (Aldhous *et al.* 2010; The Wellcome Trust Case Control Consortium 2010).

Although the studies mentioned above did not clearly report any disease associated with β-defensin copy number variation, other inflammatory diseases have also been investigated. Among these was psoriasis, an inflammatory skin disease with a strong genetic aetiology and associated with high β-defensin copy number (Hollox *et al.* 2008b). Many different susceptibility loci have been described to influence and increase the risk of the disease, with the *HLA-Cw6* allele, of MHC locus on chromosome 6p21.3, being described as the major risk allele for psoriasis (Russell *et al.* 1972; Tiilikainen *et al.* 1980; Gudjonsson *et al.*

2003; Karason *et al.* 2003). Some histological evidence acquired from the examination of psoriatic lesions, where considerable amounts of the hBD-2 protein encoded by *DEFB4* gene have been detected, support the hypothesis that β-defensins could be implicated in the susceptibility to psoriasis (Harder *et al.* 1997a; Hollox 2008). Moreover, a study from Hollox *et al.* (2003) correlated the lymphoblastoid mRNA expression level variation of *DEFB4* with β-defensin genomic copy number (Hollox *et al.* 2003). This suggests that mRNA and consequently the protein levels of these peptides may reflect the number of *DEFB4* gene copies, and therefore may be dependent on gene dosage. Finally, taking into consideration the chemokine and cytokine-like properties of hBD-2, which lead to the recruitment of immature dendritic cells to the site of inflammation, it is possible that high amounts of this protein may cause an atypical inflammatory response in the skin, ultimately leading to the formation of psoriatic plaques. Therefore, is appropriate to highlight *DEFB4* copy number as a candidate gene for psoriasis.

To investigate the relationship between β-defensin genomic copy number and susceptibility to psoriasis, two independent case-control association studies were performed. In the first study, a Dutch cohort was genotyped using HSPD5.8, the first developed PRT. These samples were simultaneously genotyped by Hollox *et al.*, using MAPH/REDVR in order to confirm the copy number. Alongside this, a German cohort of psoriasis patients and controls was genotyped using HSPD5.8 by our collaborators in Erlangen (Germany), using patients from Münster (Hollox *et al.* 2008b). In the second (follow up) study a PRT-based triplex assay was used to re-type the Dutch cohort only. The study presented here aims to contribute to the understanding of how copy number variation can influence disease development, focusing in particular on the consequences of the β-defensin copy number variation in psoriasis. Hopefully the clarification of such mechanisms will be a step forward for the understanding of psoriasis and may lead to the development of improved therapies.

4.1.2 Results

4.1.2.1 Quality control and accuracy

To carry out the genotyping of psoriatic and control samples in the two independent case-control association studies, four reference samples, with known β-defensin copy number, have been used as internal controls to calibrate each experiment (Figure 7, section 2.2.3.4). The reference samples C11, C18, C62 and C66 were selected from ECACC HRC panel 1, corresponding respectively to copy numbers of 4, 3, 5 and 6. The samples have not only showed reproducible results with multiple PRT assays but their copy numbers had been also confirmed by other methodologies such as, MAPH (multiplex amplifiable probe hybridization) and REDVR (restriction enzyme digest variant ratio) tests (Armour *et al.* 2007).

In the first study carried out using only one PRT system, HSPD5.8, the peak area of each test and reference product were recorded in order to calculate the relative ratio (section 2.2.3.1). Alternatively a PRT based triplex assay was applied in the second (replication) study and in this instance the peak heights were recorded instead to avoid the presence of a NED "dye peak", that co-migrated together with the test peak of the HSPD21 PRT system (Figure 19) (section 2.2.3.2). FAM and HEX dyes also exhibited dye peaks, but these did not interfere with any of the peaks corresponding to the test or reference locus of any system used in the triplex assay (section 2.2.7). Moreover, it was possible to observe in some samples a shift of "dye peaks" from their original position showing that they have variable mobility (Figure 19). Once the NED "dye peak" overlaps the HSPD21 test peak, the peak area will not correspond to the real area of the test peak because it will include extra material from the "dye peak". As such, data from the peak area of NED dye products will not be accurate, as demonstrated in the histograms below (Figure 20b). However, when using peak height both FAM and NED unrounded ratios are clustered, as shown by the histogram of Figure 20a. In the second histogram (Figure 20b), the values obtained from peak area for NED dye are spread along the Y-axis, indicating the addition of extra material from the "dye peak" and consequently giving inaccurate copy number measurements. Therefore, these results supported the use of peak heights as a better indicator of copy number variation to be used in further analysis (Figure 20a).

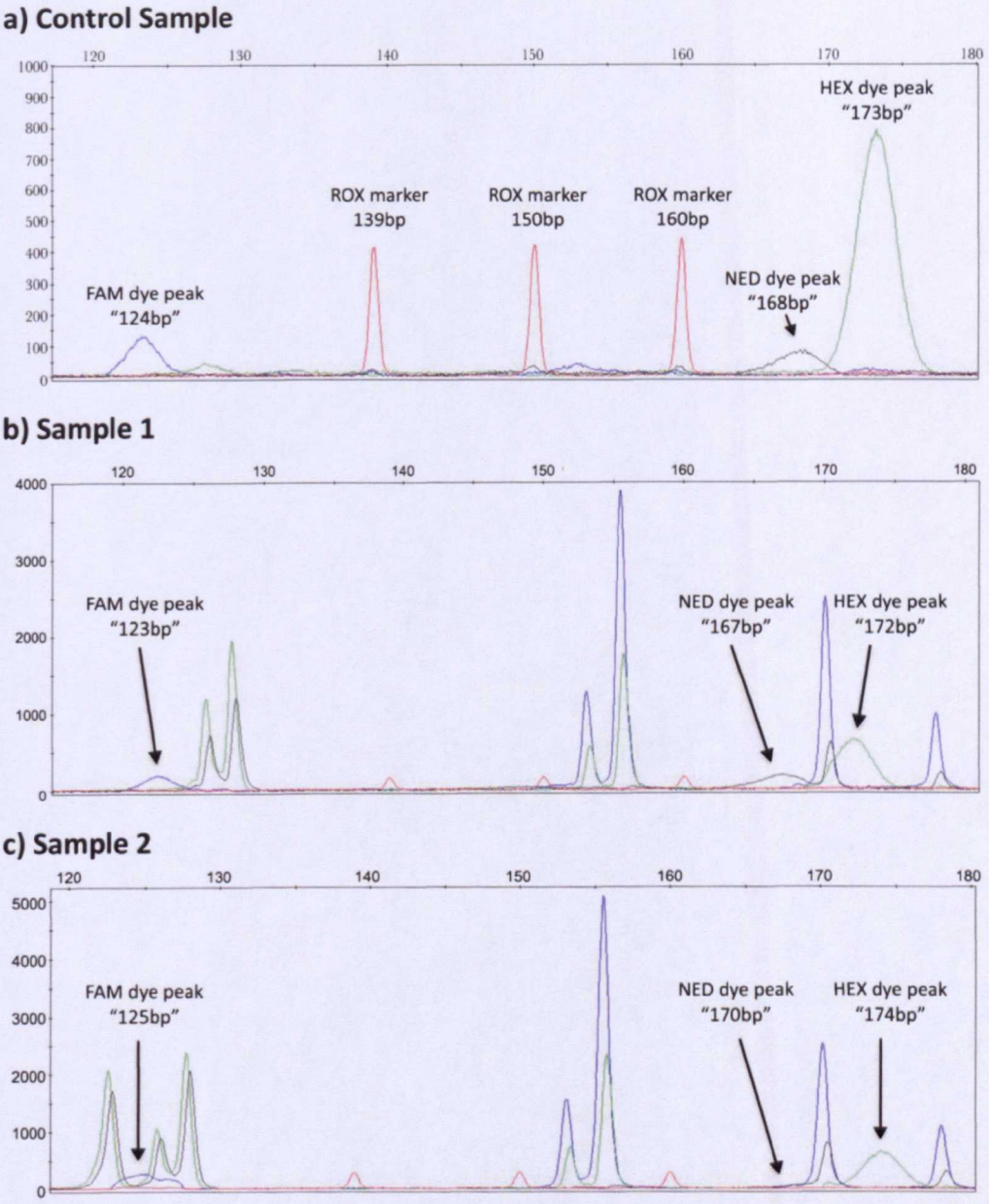


Figure 19: Location of dye peaks generated from FAM, HEX and NED primers on a triplex PRT-based assay. The blank control shows the position of dye peaks from FAM, NED and HEX with apparent sizes of 124 bp, 168 bp and 173 bp, respectively (a). Location of the three dye peaks in sample 1 (b) and sample 2 (c) is slightly shifted, approximately 1 to 2 bp, when compared with the original location in the blank control. On both samples 1 and 2, the variable mobility of the dye peaks is shown.

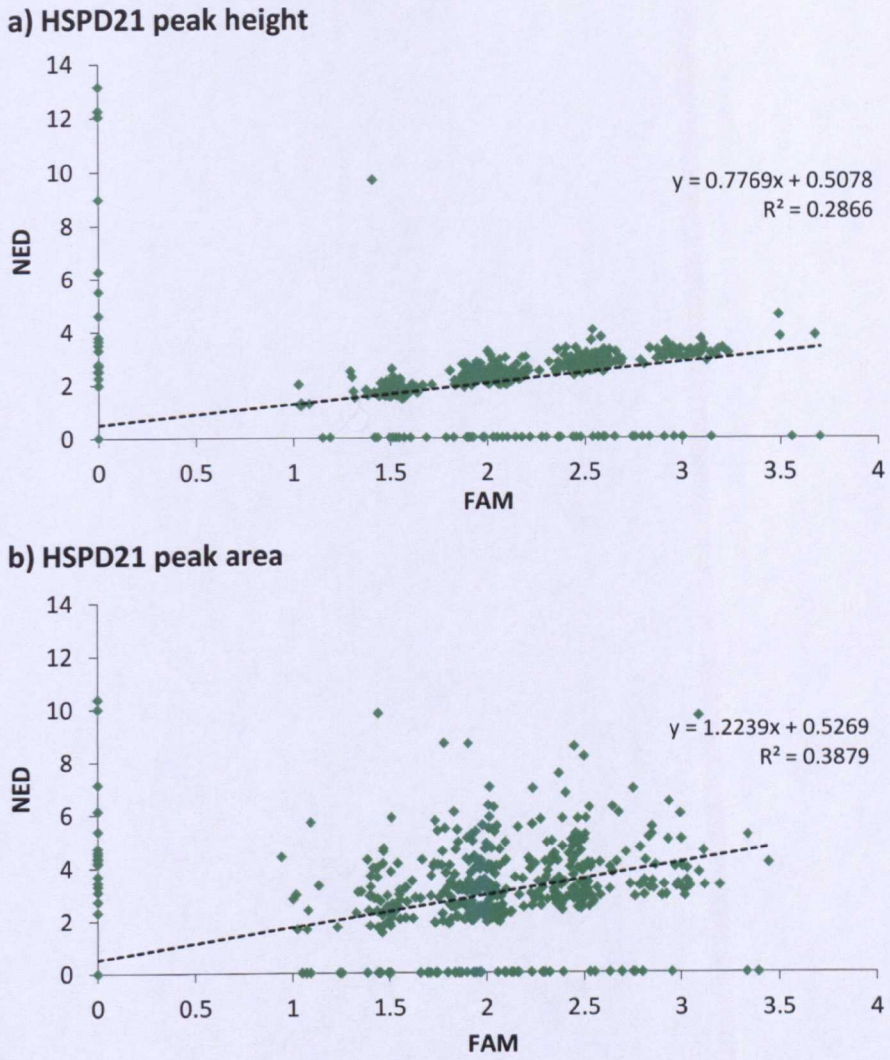


Figure 20: Results produced by HSPD21 PRT system in Triplex assay given by the data analysis of peak areas and peak heights for FAM and NED dyes. a) Scatter-plot of unrounded ratios from peak height for FAM and NED dyes showing a high concordance between the two data sets and clustering of values, presumably corresponding to integer values. b) Scatter-plot of unrounded ratios from peak area for FAM and NED dyes showing a poor concordance between the two data sets. Analysis of peak areas produced consistently higher ratios for NED dye products due to the introduction of extra material from the "dye peak".

The HSPD21 NED peaks frequently showed a lower signal when compared with FAM peaks, and were consequently more prone to error. For this reason, a comparative analysis was carried out to test if the unrounded copy numbers given by FAM and NED dyes were equivalent. The histograms (Figure 21) show the HSPD21 copy number distribution given by FAM only, the mean of FAM and NED and the weighted mean of the two dyes. Clearly, either the use of the FAM dye by itself or the weighted mean of PRTs showed more distinct clusters when compared with the use of the

arithmetic mean of PRTs. To calculate the weighted mean ratio a proportion of 2:1 (FAM, NED dyes, respectively) was used to determine the HSPD21 mean ratio of each sample. Decreasing the influence of the results from NED dye in the mean ratio improves the final copy number results. When compared with the arithmetic mean, the weighted mean improved the copy number measurements, especially for high copy numbers, suggesting that NED dye actually introduced greater error into the measurement. As such, using a weighted mean ratio of FAM and NED gave a more accurate outcome, as shown in the histogram by the formation of clear clusters separated by deeper gaps (Figure 21c). For the PRT107A, an arithmetic mean of the two dyes, FAM and HEX, was calculated since the two dyes gave constantly equivalent outcomes.

To confirm the accuracy of the triplex measurement on the reference samples, the degree of clustering of the unrounded copy numbers for PRT107A and HSPD21 was analysed. As illustrated by the scatter plot (Figure 22), the two PRT systems show clusters around integer numbers corresponding to 3, 4, 5 and 6 diploid copies of the four reference samples used, C18, C11, C62 and C66, respectively. Although the unrounded copy numbers are still clustering around the integer numbers for the majority of the cases, the clusters corresponding to the C11 (4 copies) and C62 (5 copies) are not exactly centred at 4.0 and 5.0 by the PRT107A method. The C11 cluster is shifted towards a slightly lower unrounded copy number (~3.8 copies) while the C62 cluster is shifted towards a higher unrounded copy number (~5.2 copies), possibly due to the mismatch in the primer that allows simultaneous amplification of a third locus on chromosome 2.

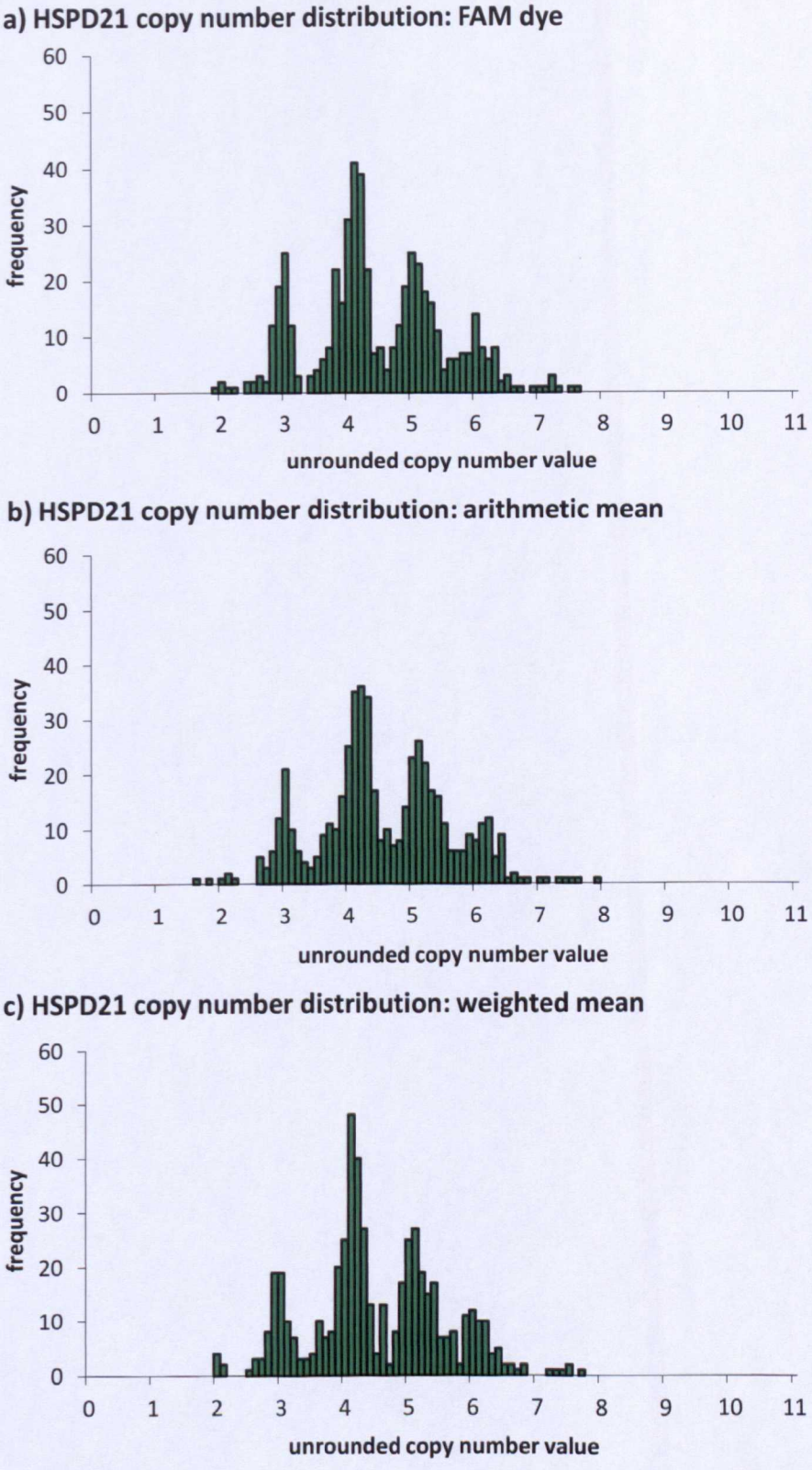


Figure 21: Histogram of HSPD21 unrounded copy number distribution of Nijmegen samples typed in the second case-control association study. a) HSPD21 copy number distribution for FAM dye only. b) HSPD21 copy number distribution for the mean. c) Weighted mean copy number of FAM and NED dyes. The weighted mean ratio was calculated as: (2 x FAM ratio + NED ratio)/3 and then used to calculate the weighted mean copy number of each sample.

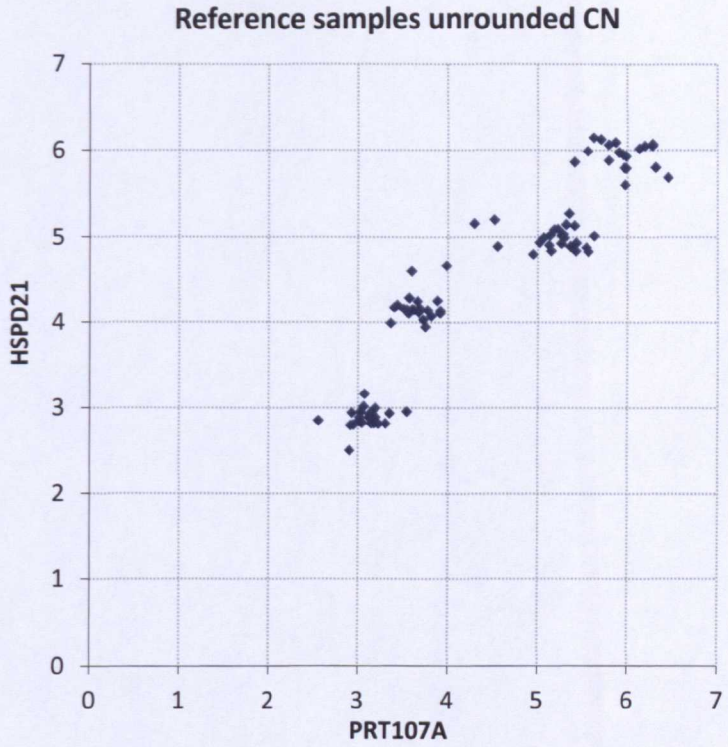


Figure 22: Unrounded copy number given by PRT107A and HSPD21 PRT systems in the triplex assay for multiple typing of four reference samples, C18, C11, C62 and C66. The scatter-plot shows clear clusters around integer numbers 3, 4, 5 and 6, corresponding, respectively to the four reference samples above.

4.1.2.2 Typing details, strategies and results concordance

4.1.2.2.1 FIRST CASE-CONTROL ASSOCIATION STUDY (HOLLOX *ET AL.* 2008B)

The initial study to investigate the relationship between β-defensin genomic copy number and susceptibility to psoriasis disease was carried out on a Dutch cohort consisting of 493 samples, 303 of which were controls and 190 psoriasis cases. However, for 42 Dutch samples no PRT data was available (8.52% failed), lowering the total number of samples actually typed with PRT to 451 (272 controls and 179 cases). These samples were genotyped in 96-well plates, which included a mix of cases, controls, reference samples and blanks allocated at random to a well position in each 96-well plate. This proceeding minimizes any batch-related error that could cause differential bias between cases and controls. Several case and control samples (267 samples in total) were replicated twice or more (maximum four times) between the 10 different 96-well plates used.

The PRT ratios obtained for HEX- and FAM- labelled products were selected and accepted for use in copy number determination if the ratios of the two-labelled products differed by less than 15% of their mean. For the results accepted, the average of the two ratios was calculated and used to determine the copy number of each sample. Unrounded copy number values given by HSPD5.8 PRT were plotted on a histogram and the analysis of integer clustering was carried out to test the accuracy of PRT to measure β-defensin copy number (Figure 23). In the histogram, different clusters were formed corresponding to distinct copy number classes. While lower copy numbers, from 2 to 5 copies, showed good clustering with deep gaps, higher copy numbers were more difficult to distinguish and measure accurately. However, this was already expected from a method based on ratios, which normally determines lower copy numbers more accurately.

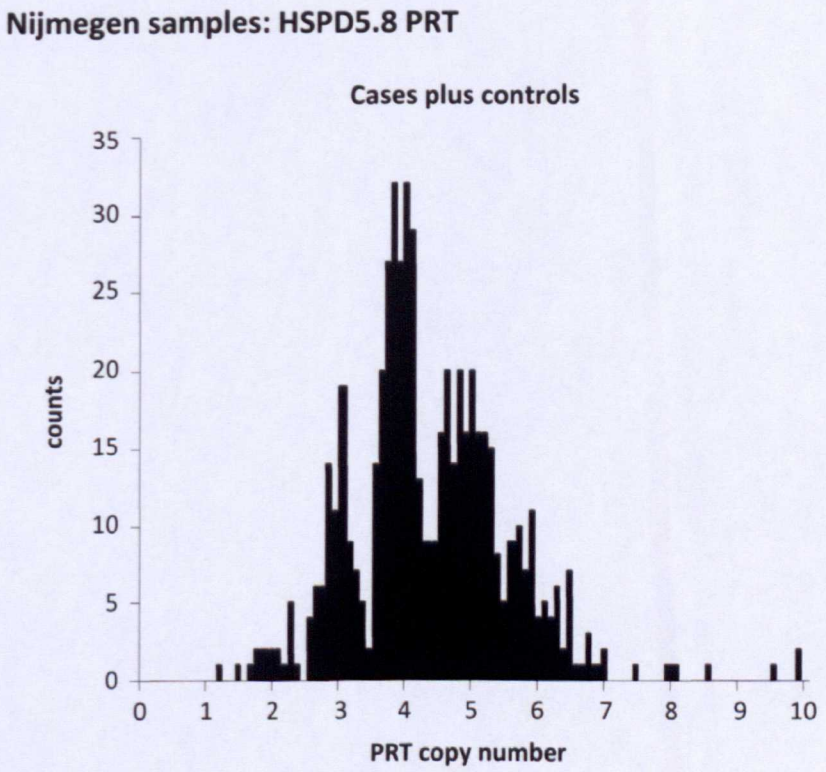


Figure 23: Histograms of unrounded copy number distribution of Nijmegen samples using HSPD5.8 PRT. The copy number distribution shows clusters around integer numbers, with low copy numbers showing an improved clustering when compared with high copy numbers (from Hollox *et al.* 2008).

The cumulative frequencies of residual values from PRT are shown below (Figure 24). Since residual values are the difference between the measured unrounded copy number and the integer value, they can actually demonstrate how well the method performed. In other words, the residuals represent the error of the method, if we assume that the biological meaning of copy number can only be represented by integer values. In the histogram, Figure 24, the frequency distribution of copy number residuals from cases (mean -0.055) and controls (mean 0.038) were comparable. Although the frequency of cases was slightly increased for negative residuals, overall the frequencies for cases and controls were very similar. The standard deviation observed for cases (std. dev.=0.25) and controls (std. dev.=0.24) indicates that the unrounded copy number values given by PRT are normally close to the integer value demonstrating the accuracy of the method.

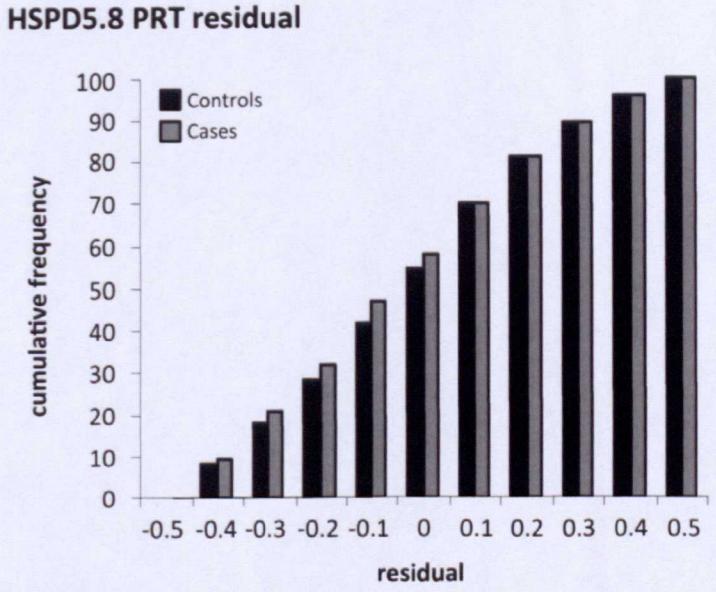


Figure 24: Comparison of cases and controls cumulative frequency distributions of residuals from HSPD5.8 PRT data. The residuals were calculated from the difference between the measured unrounded copy number value of PRT and the integer copy number (from Hollox *et al.* (2008b)).

To increase the accuracy of the copy number estimates, this cohort was also retyped by other methods, MAPH and REDVR, and a consensus integer copy number was established for each sample using the information given by the three methods together, as described in Hollox *et al.* (2008b). The comparison of integer copy numbers showed that 78% of samples

agreed between first-pass PRT and MAPH/REVDR and a further 11% agreed when a repeat typing PRT was performed on the discordant samples (Hollox *et al.* 2008b). Overall, 89% of samples agree in a copy number value between the different methods, which confirms the accuracy of the HSPD5.8 PRT to measure the β-defensin copy number, even in large case-control association studies.

4.1.2.2.2 SECOND CASE-CONTROL ASSOCIATION STUDY

The study described here is an independent case-control association study carried out in the same Dutch cohort from Nijmegen, Netherlands. Here, samples were genotyped using a PRT-based Triplex assay to measure the β-defensin genomic copy number variation. The β-defensin copy number variation was analysed in a total of 475 samples, of which 279 were controls and 196 were psoriatic cases (Table 12). From those samples, 33 controls and 7 cases were excluded either because they failed to give any genotype (20 controls) or because no copy number was determined (13 controls and 7 cases). The genotyping failure was mainly due to the low amount of DNA available while the inability to determine copy number was due to the disagreement between different systems and the low quality of results. Further analysis was carried out on the remaining 246 controls and 189 cases. To genotype the β-defensin copy number variation, cases and controls were arranged at random in 96-well plates, which also included blanks and reference samples. Some of these samples were replicated more than once among the seven 96-well plates used. As described earlier, this procedure reduces any batch-specific error associated with technical handling that could affect cases and controls differently.

Table 12: Numbers of cases and controls successfully typed and failed.

	Successful Data	Tests attempted	No data available
Dutch controls	246	279	33
Dutch cases	189	196	7
Total	435	475	40

The main advantage of the Triplex assay is to provide three independent measurements of the same sample simultaneously, two from PRT systems and one from an indel assay. In comparison to a single PRT, the Triplex assay offers much more information about the copy number of a given sample. However, to accept the results, concordance between the copy number measurements derived from each system that composed the triplex assay was taken in consideration. Such combination of all results increases the confidence of the copy number measurement.

The criteria to accept the results required the analysis of the "minimum ratio", a confidence index of the copy number given for each sample by a likelihood function (maximum likelihood copy number, ML CN) (section 2.2.7.2). From the 435 samples successfully typed, 47 samples showed a minimum ratio less than 20 and were therefore retyped. All samples with a higher minimum ratio were selected for further analysis. Then the copy numbers for every sample given by each system were compared. The ML CN was accepted if at least two systems agreed between themselves and with the ML CN (329 samples). In case of disagreement a careful analysis of the information from 5DELR4 peaks was carried out to investigate if the 5DELR4 system supported any of the PRT systems or the ML CN given by the maximum likelihood program. When the allele ratios obtained from 5DELR4 showed evidence of agreement with any of the PRT systems or the ML CN, the final copy number was dictated by 5DELR4 and if necessary the ML CN would be changed (3 samples). When neither of these conditions applied (56 samples), samples were retyped using either the Triplex assay or microsatellite analysis. From the total of 103 samples that did not pass the criteria or failed to give any outcome, only 98 samples were retyped, as DNA was no longer available for 5 samples. Samples were retyped using the Triplex assay if the initial typing failed or results were only obtained for just one system (59 samples). In this case samples were subjected to the same criteria of selection used in the first Triplex assay. If the ML CN given by each Triplex assay were not in agreement, then the copy number selected was the one that passed all the criteria or the one with higher minimum ratio. However, in case of disagreement between the three systems of the Triplex assay, microsatellite

analysis (EPEV1 and EPEV3) was applied to further clarify the sample copy number (section 2.2.5) (39 samples). The examination of allele ratios from microsatellite analysis allows the determination of the number of different alleles and can be used to support any of the results given by the PRT systems, 5DELR4 or ML CN. In case of strong support by microsatellites to any of the results from the systems mentioned above, the ML CN would be changed to a final copy number that was determined by the microsatellite analysis (6 samples). Finally, if the microsatellite analysis did not add any further information, the ML CN was maintained.

The analysis of integer clustering of PRT unrounded copy number data was carried out to verify the accuracy of the copy number measurements obtained by the Triplex assay. The unrounded copy number values from PRT107A and HSPD21 tend to cluster around integer values corresponding to copy numbers of 2, 3, 4, 5 and 6, as shown in Figure 25a. Although clustering around integer numbers can confirm the accuracy of the Triplex assay to measure the β-defensin copy number variation, clusters from PRT107A were not precisely centred on integer numbers and were shifted slightly towards lower values, possibly due to the amplification of an additional locus on chromosome 2. The chromosome 2 locus had only one mismatch with the reference locus and showed products sizes very similar to the reference locus products. Such similarity could be responsible for variation in the amplification of the reference locus for PRT107A giving rise to the anomalous calibration of PRT107A as described in section 4.1.2.1. As a consequence, the mean copy number of reference samples differs for PRT107A. To avoid the propagation of error, a linear transformation (adjustment) was applied to the entire cohort (cases and controls) to adjust the copy number values and improve the quality of the results (Figure 25b).

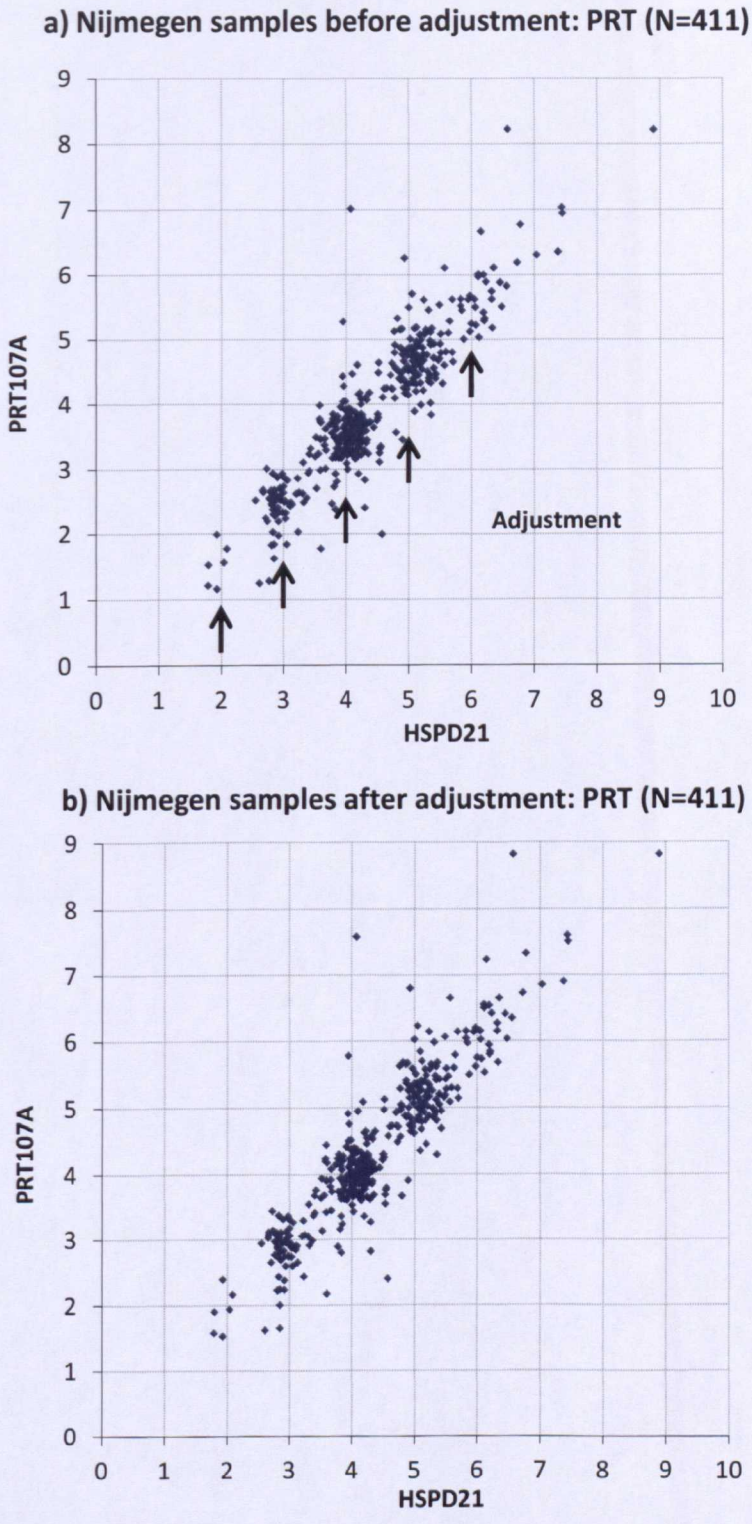


Figure 25: Copy number values from PRT107A and HSPD21 resulting from a triplex assay. a) Scatter-plot of unrounded copy number values of PRT107A and HSPD21 showing clear clusters corresponding to different copy numbers classes. In this plot PRT107A values are shifted down from the integer copy numbers when compared to HSPD21. b) Scatter-plot showing the unrounded copy number values after PRT107A adjustment. Clusters are placed more exactly around integer values.

In a different analysis copy number values have been represented in a histogram (Figure 26) that shows the accuracy of the method but also the frequency distribution of the different copy number classes measured by the Triplex assay. To analyse each PRT independently, a histogram was separately plotted for PRT107A (Figure 26a) and HSPD21 (Figure 26b), while a third histogram was plotted with the weighted mean values from the two PRT systems (Figure 26c). These three histograms show a copy number distribution characterized by the formation of clear clusters, which are placed around integer numbers. Although none of them show clear gaps between each copy number group, the clustering remains evident in the three histograms. The PRT107A copy number values showed a wider distribution when compared with HSPD21 or PRT weighted mean copy number distribution (section 2.2.7). Both HSPD21 and PRT weighted mean copy numbers presented tighter clusters divided by clear gaps showing an improved distribution of the different copy numbers, ultimately confirming the precision of the method. Even if none of the histograms showed a clear separation between the clusters, the weighted mean copy number distribution of the two PRT systems illustrates a remarkable clustering for β-defensin copy number variation. The PRT weighted mean values were then plotted according the final ML CN, which takes in consideration all the Triplex data obtained for each sample (Figure 27). Each copy number was represented in the histogram by a distinct colour, allowing the definition of the different copy number clusters. As shown in Figure 27, the edges of some copy number clusters of PRT values, such as the ones corresponding to 2, 5 and 6 copies overlap between each other, demonstrating the importance of the 5DELR4 analysis in the final copy number determination, mainly when measuring high copy numbers. Finally, to better illustrate the Triplex performance, unrounded copy number values obtained from the weighted mean of PRTs were normalized as shown in Figure 28 by dividing by the nearest integer value, so that a mean value of 1.0 is expected. The data tend to cluster around the mean showing a low standard deviation, which is indicative of the high quality of the results and the accuracy of the measurements obtained by the Triplex assay.

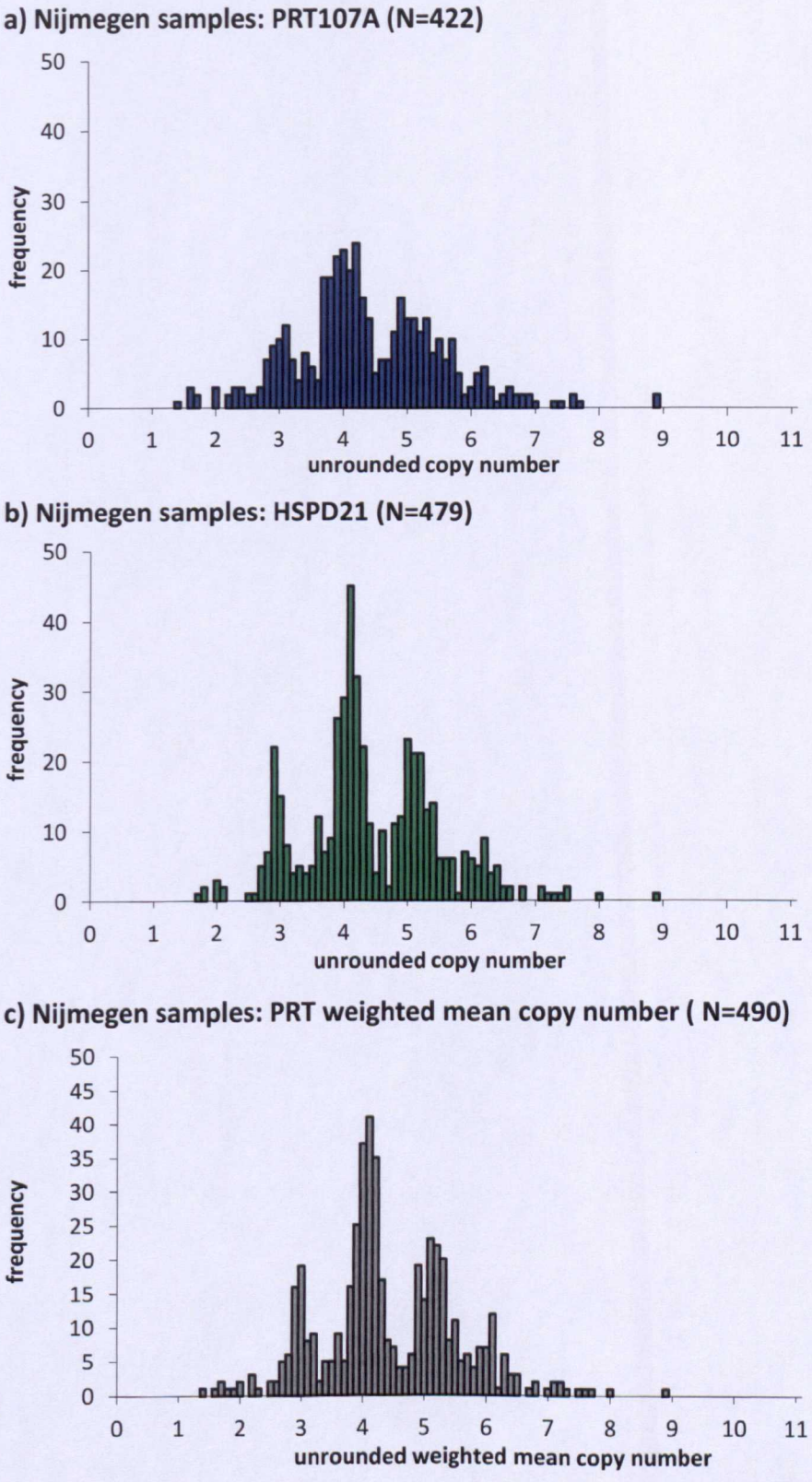


Figure 26: Histograms of unrounded copy number distribution from PRT107A, HSPD21 and weighted mean of PRT systems. PRT107A (a) and HSPD21 (b) copy number distribution showing clusters around integer numbers. The HSPD21 histogram illustrates a copy number distribution where clusters are better separated when compared with PRT107A. PRT weighted mean (c) copy number distribution showing an improved clustering of the copy number values compared with the individual PRT systems.

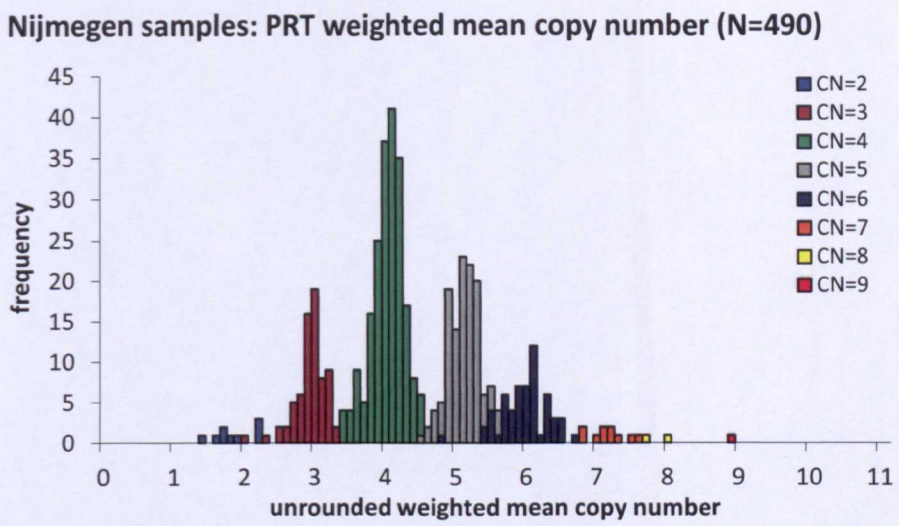


Figure 27: PRT weighted mean copy number distribution according to the final ML CN. Each colour represents a different copy number given by the analysis of the Triplex data on a maximum likelihood program.

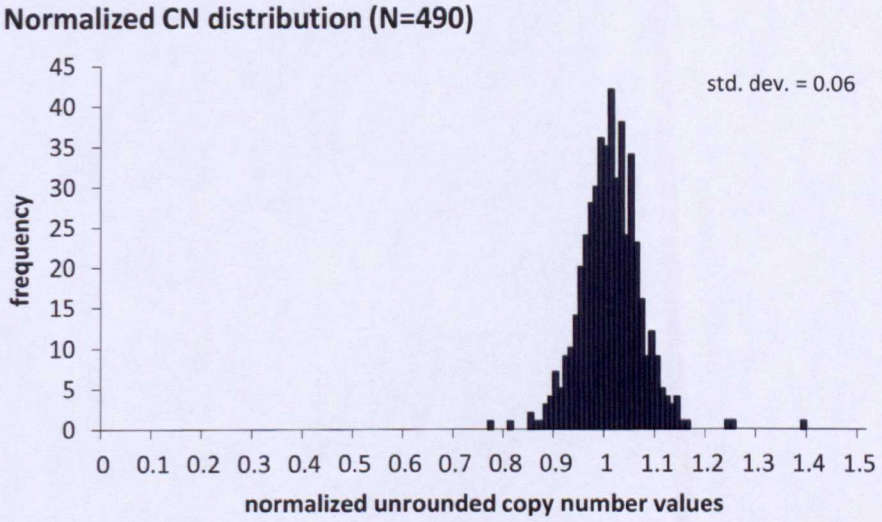


Figure 28: Distribution of the normalized unrounded copy number values obtained from the PRT weighted mean, showing a low standard deviation.

The analysis of clustering allows us to assess the accuracy of the measurement methods, using the assumption that only integer copy numbers have biological meaning. Therefore, differences obtained between the measured values for each individual system and the final integer copy number, ML CN, should represent the error of the measurement, which can be used as an indicator of the method accuracy. Figure 29 shows the copy number residuals for each PRT, calculated from the difference between the measured unrounded copy number and the obtained ML CN for each

sample. The graph shows how PRT107A residuals have a wider distribution than HSPD21 residuals. PRT107A residuals vary between approximately -1.6 and 1.6 from the rounded copy number given by the ML CN, while residuals from HSPD21 are more frequently close to zero, varying between approximately -2 and 0.9, indicating a higher accuracy. Overall, PRT residuals tend to congregate around zero (average residual: PRT107A=-0.0048 and HSPD21=0.0069) forming a spherical distribution indicating that the majority of the results from either of the PRT systems gave a correct outcome that was very close to the integer copy number value. A non-significant correlation (*p*-value=0.1758, correlation test) was found between PRT107A and HSPD21 residuals indicating the absence of somatic mosaicism or if present infrequently or at low level.

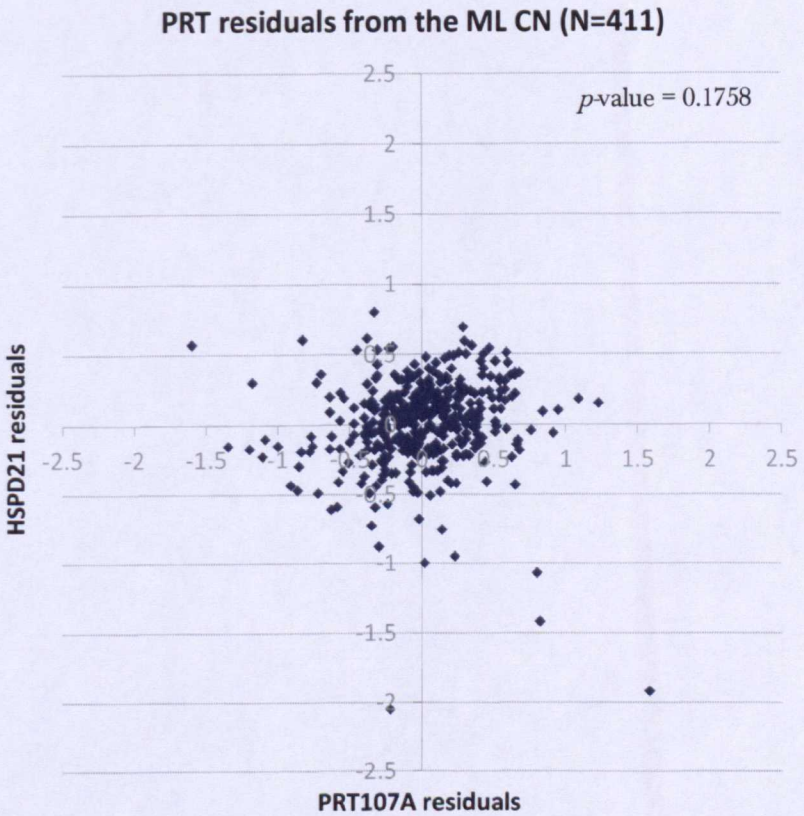


Figure 29: Comparison of copy number residuals from PRT107A and HSPD21 data. The residuals were calculated from the difference between the measured unrounded copy number value of each PRT and the integer copy number (ML CN) given by the maximum likelihood program.

4.1.2.2.3 CONCORDANCE BETWEEN THE TWO STUDIES

Considering the first and second case-control association study together, 516 samples from Nijmegen, Netherlands, were typed in total. From these, 446 samples were typed in both studies that after removing the samples that failed become a total of 407 samples successfully typed, from which 240 and 167 were controls and cases, respectively. The copy number values described here for the first case-control association study, result from a consensus copy number of PRT (HSPD5.8), MAPH and REDVR, as described in Hollox *et al.* (2008b). The agreement between the copy numbers given by each association study was analysed for the samples replicated on both studies (407 samples). For 87.96% of samples (358 samples) the integer copy number agreed between studies, while for 12.04% of samples (49 samples) the copy number value differed. From those 49 samples that disagreed, the copy number values given by the second study were higher for the majority of the samples, either for cases or controls (29 samples), when compared with the first study (Table 13). Since cases and controls showed concordantly higher copy number values in the second study, the disagreement observed between the two studies was not biased for cases or controls. Overall, the net difference in copy number repeat units between the two studies differed by just 11 repeat units. The control samples were the ones that varied the most between the two studies, contributing 10 repeat unit differences. The disagreements observed in the copy number measurements between the studies (44 samples) were mainly in the order of one copy number repeat unit with the exception of 5 samples, in which the outcome differed by 2 (4 samples) or 3 copies (one sample).

Table 13: Dutch samples for which the final copy number value differed between the two association studies. The first column shows the number of cases, controls and total of samples that disagreed. The second and third columns indicate the number of samples in which the copy number was higher (UP) or lower (DOWN) in the second study compared with the first association study. In the last column the net difference in terms of copy number repeat units is represented.

	Number of samples that disagree	Up	Down	Net difference of repeat units
Controls	31	19	12	10
Cases	18	10	8	1
Total	49	29	20	11

4.1.2.3 Psoriasis disease association results

To investigate the relationship between β-defensin genomic copy number and susceptibility to psoriasis disease, the copy number variation was analysed and the association with psoriasis was tested in two independent case-control association studies using PRT based assays. As reported previously, the copy number variation at the 8p23.1 β-defensin cluster commonly varies between 2 and 7 copies.

In the first case-control association study the β-defensin genomic copy number was analysed in 272 controls and 179 psoriasis cases from Nijmegen, Netherlands. The β-defensin copy number reaches higher values in cases than in control samples, with controls varying between 2 and 7 copies and cases varying between 2 and 8 copies. Comparison of PRT results from cases and controls suggested an association of high β-defensin copy number with psoriasis, confirmed by a *p*-value of 0.01 (*t*-test) (Hollox *et al.* 2008b). To improve the overall accuracy of the copy number measurements the data from PRT was combined with the data from MAPH and ratios of multisite variants (MSVs) in a consensus copy number. This consensus copy number showed, once again, a significant difference (p-value=7.8×10^{-5}, *t*-test) between cases and controls, confirming the association between high β-defensin copy number and susceptibility to psoriasis (Figure 30) (Hollox *et al.* 2008b). Alongside this cohort our collaborators in Erlangen, Germany, typed a German cohort with 624 samples, 305 controls and 319 psoriasis cases from Münster using the HSPD5.8 PRT alone. In this independent test of association they also found a significant association between β-defensin copy number and psoriasis (p-value=2.95×10^{-5}, *t*-test) (Hollox *et al.* 2008b).

In the second association study carried out, 246 controls and 189 psoriatic cases were analysed for β-defensin copy number variation by a PRT-based Triplex assay. Overall, the β-defensin copy number varied between two and nine copies per diploid genome with a mean copy number of 4 (Figure 31). Considering each group separately, the copy number varied between two and seven copies for controls and between two and nine copies for cases. The cases showed a wider range of copy number variation with a higher mean copy

number (mean CN cases=4.52381) compared with controls (mean CN control=4.252033). The difference in the mean copy number observed between cases and controls proved to be significantly different using a *t*-test showing a *p*-value of 0.0058.

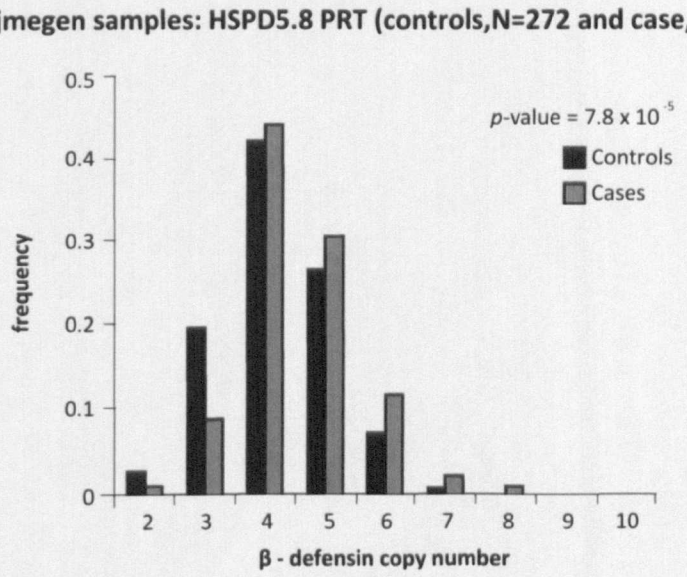


Figure 30: Frequency distributions of β-defensin genomic copy number in the Dutch cohort from Nijmegen. Copy number distribution of 272 controls and 179 psoriasis cases, where cases showed on average higher copy numbers (Hollox *et al.* 2008b).

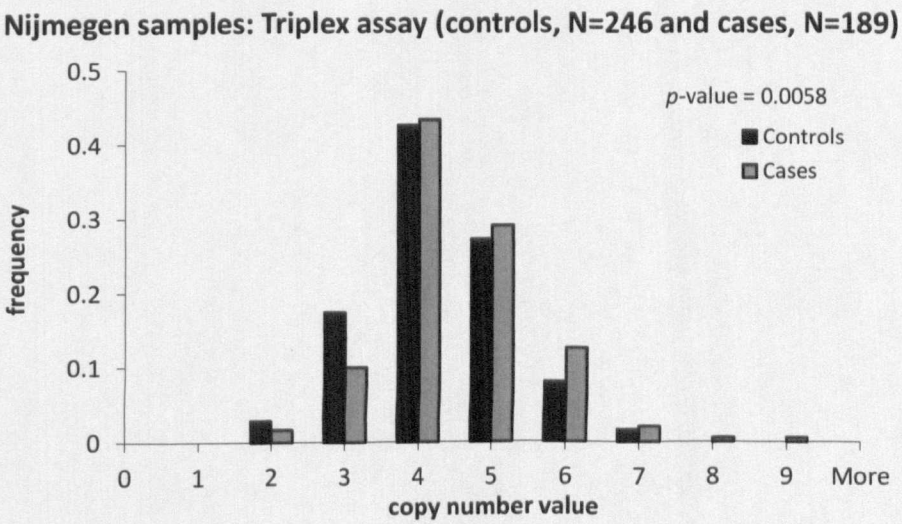


Figure 31: Frequency distributions of β-defensin genomic copy number from 435 samples, 246 controls and 189 psoriatic cases, separately. The histogram illustrates the β-defensin copy number variation for cases and controls in a European population from Nijmegen, Netherlands with cases showing clearly higher copy numbers (8 and 9 copies) compared with controls.

In a case-control association study is important to assure that no differential bias in genotyping is observed between cases and controls, which could increase the false-positive rate and lead to spurious associations (Clayton *et al.* 2005; Plagnol *et al.* 2007). If DNA preparation methods are sufficiently different between cases and controls this could be a source for genotyping bias which ultimately could lead to an apparent association (Clayton *et al.* 2005). When measuring copy number, particularly high copy numbers, this matter is especially important. Since higher copy numbers are more difficult to measure, they are more likely to produce unsuccessful results that will consequently be rejected. Thus, differential drop-out of samples is more likely to affect samples with high copy numbers. To avoid this, the DNA preparation methods used in these studies were the same for all Dutch samples. In addition, cases and controls were randomly interspersed throughout the 96-well plates used for PRT analysis, thus minimizing any genotyping error favouring either cases or controls.

However, to test for differential biases, cumulative frequencies of the set of copy number values obtained for cases and controls were plotted in a graph that shows the cumulative totals of the set values. In the case of differential bias caused by systematic experimental error affecting just one of the sample sets, the effect should be detected through the analysis of cumulative frequencies of each copy number group separately. The cumulative frequency plots for cases and controls with 3 (N=73), 4 (N=207) and 5 (N=124) copies were very similar, with *p*-values (*t*-test; 0.995, 0.985 and 0.948, respectively) showing no significant difference between cases and controls (Figure 32). Thus, the significant correlation reported here cannot be attributed to differential bias between the two groups studied but instead to real frequency differences of integer values.

Combining the two studies together, using all samples for which reliable measures are available, the association between β-defensin copy number and psoriasis susceptibility was once again tested for a total of 516 samples, 314 controls and 202 psoriasis cases. The copy number distribution was very similar to the one previously reported, with mean copy number for controls at 4.24 and for cases at 4.55. The results confirmed the association of β-defensin copy number with psoriasis susceptibility, but at a higher significance (p-value=9.44×10^{-4}, *t*-test) (Figure 33), which strongly confirms the association previously reported.

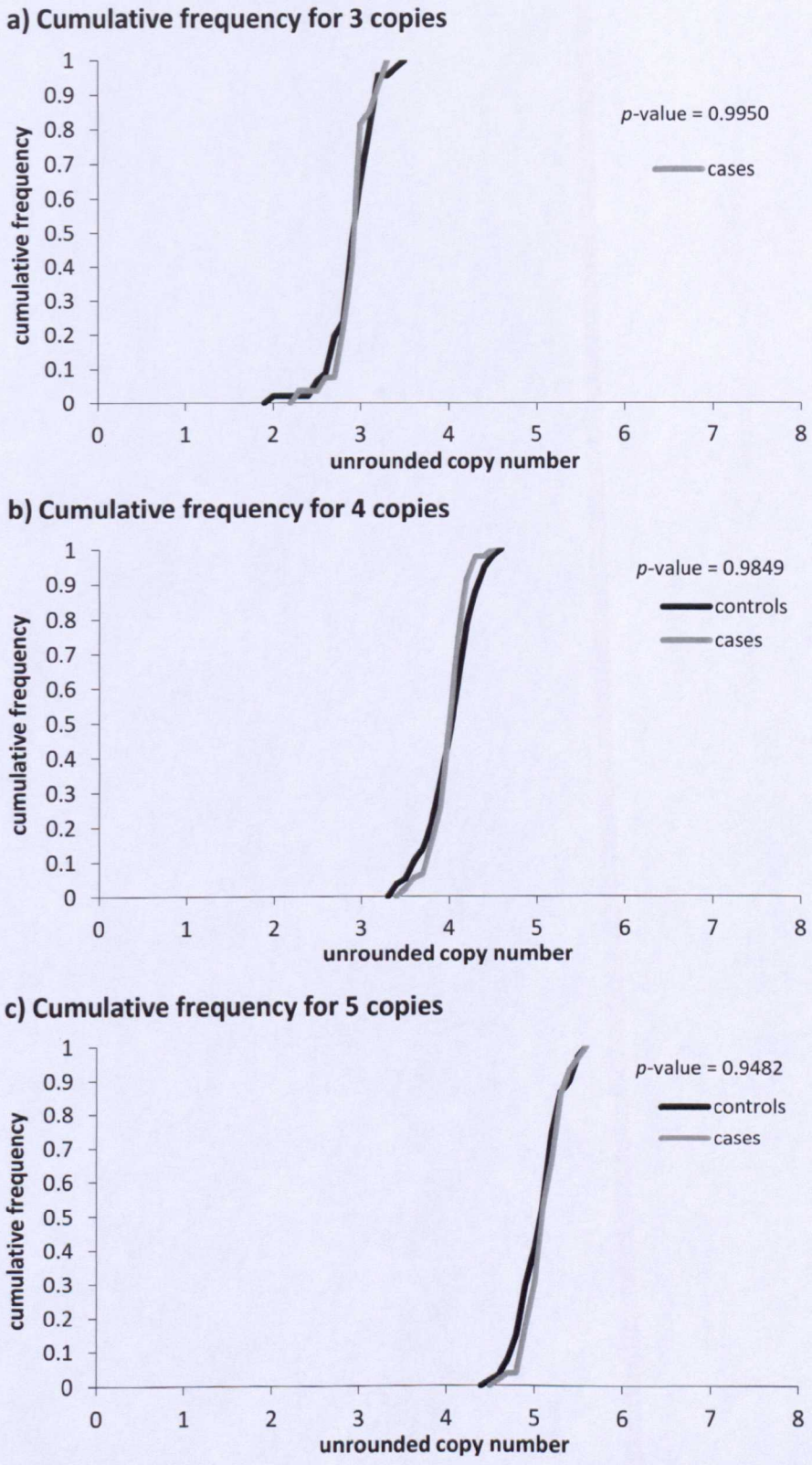


Figure 32: Cumulative frequency graphs of cases and controls for 3 (a), 4 (b) and 5 (c) β-defensin copies, representing the cumulative totals of the set values. The cumulative frequency plots of cases and controls do not show much difference for any of the copy number groups, confirming the absence of large differential bias between cases and controls.

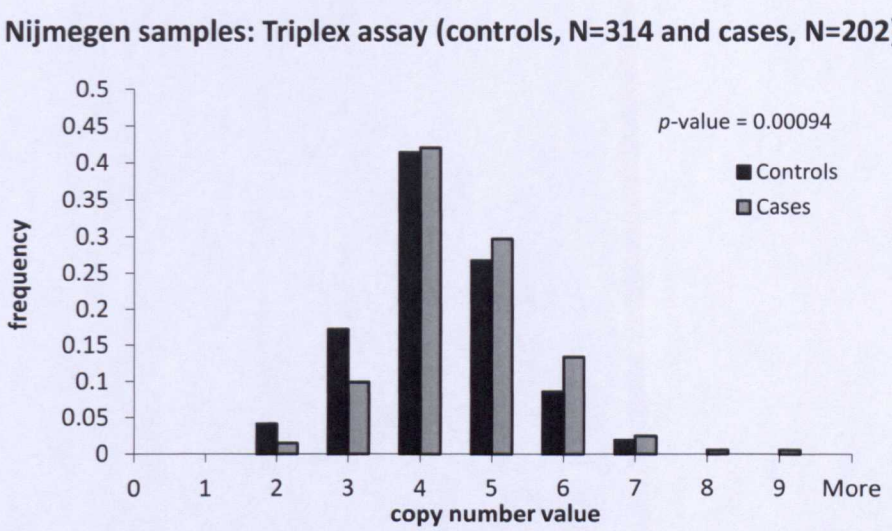


Figure 33: Frequency distributions of β-defensin copy number for 314 controls and 202 psoriatic cases, separately. The histogram illustrates the β-defensin copy number variation in a European population from Nijmegen, Netherlands, where cases show higher copy numbers (8 and 9 copies) than controls.

4.1.3 Discussion

Variants such as SNPs have long been studied and associated with disease (The International HapMap 3 Consortium 2010). On the other hand, little is known about the consequences of copy number variation in disease, but recent studies have highlighted the importance of CNV in disease susceptibility (Wain *et al.* 2009; Conrad *et al.* 2010; Sudmant *et al.* 2010). The studies presented here aimed to investigate the consequences of β-defensin copy number in disease, particularly in psoriasis. To investigate the association of copy number with disease it is essential to measure the copy number accurately, as power to detect a real association will be reduced with error-prone methods. Here, new PRT-based measurements methods were used and their accuracy tested in a large case-control association study to investigate the β-defensin copy number variation in psoriasis.

The PRT-based methods used in our studies accurately measure the β-defensin copy number variation in association studies. Moreover, the results have a high degree of reproducibility shown by the repeat typing of internal reference samples with known copy number (data not shown). Furthermore, the internal reference samples were also previously typed by other measurement methods, such as MAPH and REDVR, which confirmed our results. From the clustering analysis of internal reference samples, a very good

clustering around integer values (3, 4, 5 and 6) was obtained. Such clustering indicates the accuracy and reproducibility of the PRT assays that is emphasised by a miscalling error rate of ~8.1% for a single PRT test (HSPD5.8) (Armour *et al.* 2007) and 4.85% for the Triplex assay. A similar degree of clustering was also observed for the Dutch cohort in which five distinct clusters centred on integer values of 2, 3, 4, 5 and 6 were formed. Considering that only an accurate measurement method can distinguish different copy numbers, clustering analysis is without doubt a powerful tool to evaluate the performance of a CNV measurement method, working as a simple "barometer" for accuracy. Furthermore, a high concordance from the detailed comparison of independent typing platforms was shown between the first and second studies for cases and controls. Once again the reproducibly of the PRT was tested and confirmed in large case-control association studies.

A multiplex PRT was first described by Walker *et al.* (2009) to measure the *CCL3L1/CCL4L1* copy number variation. They demonstrated that multiplex PRT is an accurate and robust method to measure multiallelic copy number and suitable to use in large scale case-control association studies (Field *et al.* 2009; Walker *et al.* 2009). Taking this into consideration, the combination of three different systems in our study, two PRTs and an indel assay, provides more accurate information than a single PRT. Even knowing that the reference locus of PRT107A has one mismatch with a locus on chromosome 2, which could lead to a degree of co-amplification and result in a lower copy number in this system, the other two systems are not affected and consequently the triplex assay still improves the power of copy number measurement. Comparing the data from the Triplex assay versus each of the individual PRTs demonstrates the higher accuracy of the Triplex assay and its improved clustering when compared with HSPD21 or PRT107A alone. Although it is clear that the Triplex assay have an improved degree of clustering than any individual system, the quality of clustering decreases for more than 6 copies, even for the Triplex assay. It is then worth thinking if it is meaningful to separate the copy number classes greater than 6. Previous reports, which measure the beta-defensin copy number variation using real time PCR, followed a very conservative analysis separating the copy number values only in three different

classes, such as: less, equal or higher than 3 or 4 copies (Fellermann *et al.* 2006; Bentley *et al.* 2010). This kind of approach gives little information about the real copy number of beta-defensins and was one of the motivations for the development of a method that could actually distinguish the true integer states of the beta-defensin copy number variation. Moreover, the data also shows a higher standard deviation and integer error for the 6 copy number class, in all systems, when compared with lower copy number classes (3, 4 or 5 copies), indicating that the Triplex assay is probably less accurate in distinguishing copy numbers higher than 6. However it is still possible, even with a higher degree of error, to distinguish between copy numbers higher than 6. In the Nijmegen cohort there are some examples of samples showing a copy number of 8 and 9. For each of them, the copy number measured between repeats for the same sample never showed a copy number difference of more or less than one integer value. The approach used here can, with some degree of error, measure high copy numbers, giving an accurate or estimated value closer to the real copy number than previous methods have done for the beta-defensin locus. This is nevertheless better than pooling all the copy numbers higher than 6 in the same copy number class. For the Nijmegen cohort only 17 samples showed a copy number higher than 6. Certainly, with access to more samples with high copy number would be possible to improve the copy number clustering seen here for 6, 7, 8 and 9 copies.

Recent studies also compared the accuracy of the Triplex assay with other methodologies. In a side-by-side typing, Triplex assay and real time PCR were used to measure β-defensin copy number variation in a large case-control association study of Crohn's disease (Aldhous *et al.* 2010). This study strongly supports the Triplex assay as a reliable and robust method for copy number typing. The real time PCR showed very poor clustering which suggests reduced accuracy and reproducibility when compared with the Triplex. Therefore, the Triplex assay appears to be the most accurate measurement system, so far, to be used in measuring multiallelic copy numbers.

Our findings cannot be attributed to differential genotyping bias, since different genotyping parameters were closely matched between cases and controls. Important parameters such as DNA preparation, integer clustering

and independent replication were carefully taken in consideration in these case-control association studies. As such, the difference between the means of cases and controls can only be attributed to the real difference of integer values, which indicates an association between high β-defensin copy number and susceptibility to psoriasis.

The results found in the replication study strongly supports the association previously found using HSPD5.8 PRT alone (Hollox *et al.* 2008b). Even given that the differences observed in copy number values were few between the studies for cases or controls, and that both showed a significant association between β-defensin copy number and susceptibility to psoriasis, the *p*-value obtained in the replication study was higher than the one obtained in Hollox *et al.* (2008b). However, combining the two studies decreases the resultant *p*-value when compared with the *p*-value obtained for the replication study alone. This increased level of significance could possibly reflect the importance of sample size in case-control association studies, since the combined analysis of the two studies included more samples.

Another comparison between cases and controls for the same Dutch cohort (179 cases and 272 controls) was performed using a consensus integer copy number given by PRT (HSPD5.8) and MAPH/REDVR. This study also showed a significant higher copy number among cases (p-value=7.8×10^{-5}, *t*-test) (Hollox *et al.* 2008b). Finally, the association between β-defensin copy number and psoriasis was once again tested and confirmed (p-value=2.95×10^{-5}) in an independent association study carried out in a German population using PRT alone (HSPD5.8) (Hollox *et al.* 2008b). These studies once again confirmed the association previously described, and together with our findings give a strong argument that high copy number of β-defensin increases the susceptibility for psoriasis.

The statistical analysis proposed here to test for significant differences between the means of two groups (cases and controls) was carried out using the t-test for independent samples. The t-test has traditionally been used when only two groups are involved, while the analysis of variance (ANOVA) can be used when two or more groups are being tested. A t-test shows the differences in the means between two groups assuming that the variances in the populations from which the two sample groups were taken are identical. Similarly, analysis of

variance (ANOVA) measures the overall difference among the means of two or more groups comparing the variability within the groups to the variability between the groups to determine if there are any significant differences among the means. In the study presented here, as the variances between the groups are similar we used the t-test to determine if the difference between means was significant. ANOVA could be expected to detect differences in distribution between copy number classes if analysed separately, but in testing the mean of the entire distribution, the use of ANOVA in this case is mathematically identical and would produce the same conclusions as the t-test for independent samples.

In the replication study described here, performed with the Triplex assay, a 0.272 mean increase in copy number among individuals affected with psoriasis vulgaris was observed. These findings support earlier work showing higher concentrations of β-defensin proteins, particularly hBD-2 (*DEFB4* gene), in psoriatic lesions (Hollox *et al.* 2003). Previous studies had indicated hBD-2 as the main beta-defensin with a role in psoriasis susceptibility (Harder *et al.* 1997a; Hollox *et al.* 2003). However, is important to not rule out other beta-defensin genes included in the copy number variable region at 8p23.1. hBD-3, was also isolated from human psoriatic lesions and is expressed in different tissues, including skin (Harder *et al.* 2001). The particular contribution of each beta-defensin gene of this copy variable region in psoriasis risk is not known, but is likely that the ones expressed in skin, such as *DEFB103* and *DEFB4,* have a key role in skin immunity and therefore in the psoriasis aetiology. Moreover, the increase in copy number is also concordant with the fact that β-defensin shows cytokine-like properties that activate different immune and inflammatory cells and promote keratinocyte migration and proliferation that could lead to psoriasis (Niyonsaba *et al.* 2007). The involvement of β-defensin in skin immunity (cutaneous inflammation) and wound healing suggests that a higher copy number of these defensins may cause an atypical inflammatory response in the skin, ultimately leading to the formation of psoriatic plaques. Even though the role of β-defensin in the psoriasis molecular pathway is not completely understood, our studies identified a new susceptibility locus for psoriasis that has not been previously found by linkage analysis, highlighting for the first time the importance of β-defensin copy number and their influence in the aetiology of psoriasis.

4.2 CHARACTERIZATION OF β-DEFENSIN CNV IN HUMAN POPULATIONS

4.2.1 Background

Our knowledge of human genetic variation has been to some extent limited to heterochromatin polymorphisms, sufficiently large to be visualized in the light microscope, and to genetic markers, such as SNPs (single nucleotide polymorphisms), microsatellites and minisatellites, conventionally detected by PCR-based or Southern blot typing. In the last few years new methodologies with greater resolution, such us microarray technologies, SNP genotyping arrays and next-generation sequencing, have been able to detect other types of genetic variation of intermediate scale, such as copy number variation. Among the regions enriched for CNV are genes involved in chemosensation and immune responses, which are especially important in adaptation to new environmental niches, but also genes involved in fertility and reproduction, which were subject to a rapid evolution in primates due to the expansion of sexual competition (Nguyen *et al.* 2006; Voight *et al.* 2006). This suggests the role of CNVs in human evolution and adaptation, conferring genetic plasticity for organisms to evolve in response to external selective pressures (de Smith *et al.* 2008).

Although purifying selection has been suggested to overcome positive selection in human copy number variation evolution (Nguyen *et al.* 2008), some studies reported that certain human CNV genes are under positive selection, such as the amylase locus (*AMY1* gene) (Perry *et al.* 2007) and *UGT2B17* gene (Xue *et al.* 2008). Apart from this, neutral evolution has also been suggested to underlie certain CNVs, mainly those including chemosensory genes (Nozawa *et al.* 2007; Young *et al.* 2008). A positive correlation was described between the copy number of the salivary amylase gene (*AMY1*) and the starch content of diet. Perry *et al.* (2007) found that populations with a diet rich in starch have a higher mean copy number than populations with "low-starch" diets, which consisted mainly of meat, fruit, honey and milk. This suggests that higher copy numbers of the *AMY1* gene had been selected in response to adaptation to "high-starch" diets, but in the absence of such selective pressure the copy number has evolved neutrally

(Perry *et al.* 2007). The human *UGT2B17* gene also varies according to the geographic region, with an increased frequency of the deletion polymorphism from Africa to East Asia and intermediate frequency in Europe (Xue *et al.* 2008; Yang *et al.* 2008; Campbell *et al.* 2011). *UGT2B17* gene encodes an enzyme that catabolises steroid hormones, androgen and estrogen, which have important roles in maintaining cancellous bone mass and integrity. Thus, increased expression of *UGT2B17*, as a consequence of the increased copy number of this gene, might lead to an imbalance between bone formation and bone resorption and enhance the risk of osteoporosis. Their findings suggest that higher *UGT2B17* copy number is associated with increased risk of osteoporosis in both Han Chinese and Europeans. Moreover, CNV of *UGT2B17* also showed a strong ethnic difference, with the deletion polymorphism of *UGT2B17* being much more common in Chinese than in Europeans. This phenomenon is in agreement with the lower levels of osteoporosis observed nowadays in the Chinese compared with western populations (Yang *et al.* 2008). In the past, these populations might have been exposed to different selective pressures and selection might have favoured different levels of *UGT2B17*, suggesting positive selection for deletion in East Asia (Xue *et al.* 2008).

Another example of notable geographical difference among human populations, especially between European (0 to 4 copies) and African (2 to 14 copies) populations occurs in the copy number of the *CCL3L1* gene (Gonzalez *et al.* 2005; Walker *et al.* 2009; Campbell *et al.* 2011). *CCL3L1* is a chemokine gene encoding for a potent ligand that binds to several pro-inflammatory cytokine receptors, such as the CCR5, the principal co-receptor for HIV. Some studies have suggested a close interaction between low *CCL3L1* copy number and increased risk for HIV infection (Townson *et al.* 2002; Gonzalez *et al.* 2005; Mackay 2005). However, this finding has not been independently replicated (Shao *et al.* 2007; Urban *et al.* 2009) and cannot explain the differences in copy number of this locus observed between populations. Although, the reason that accounts for this population diversity is not clear, the role of *CCL3L1* in the CCR5 pathway in blocking its signal, suggests the importance of the chemokine system in modulating immune responses (McKinney *et al.* 2008).

The existence of geographical differences in CNV between populations highlights the fact that certain genes may be under selection, mainly those that confer a clear adaptive advantage for human populations, such as the ones with an important role in autoimmunity in humans (Perry *et al.* 2007; Xue *et al.* 2008; Schaschl *et al.* 2009). β-defensins have an important role in the innate immune system, acting directly against pathogens or as signalling molecules mediating inflammatory responses. The role of β-defensins suggests that their variation in copy number could affect the susceptibility to inflammatory or autoimmune diseases, as demonstrated by the association of high copy number with psoriasis (Hollox *et al.* 2008b), or conversely by the association of low copy number and an increase risk for infectious diseases. The β-defensin copy number can commonly vary between 1 and 9 copies per diploid genome with a mean copy number of 4, but copy numbers up to 12 copies can be found in some individuals (Hollox *et al.* 2003; Armour *et al.* 2007; Groth *et al.* 2008; Hollox *et al.* 2008a). So far it has been demonstrated that mean copy number of β-defensin is higher among individuals with psoriasis than in a healthy population (Hollox *et al.* 2008b), although there is no evidence for any geographical variation. The mean copy number of β-defensins has been reported to be 4 copies in all human populations typed by a single PRT assay (Hollox 2008); however, the triplex assay with its increased power in copy number measurement should provide a better evaluation of the copy number variation. Consequently, the triplex assay was used to type β-defensin copy number in 6 different human populations to better understand CNV between populations.

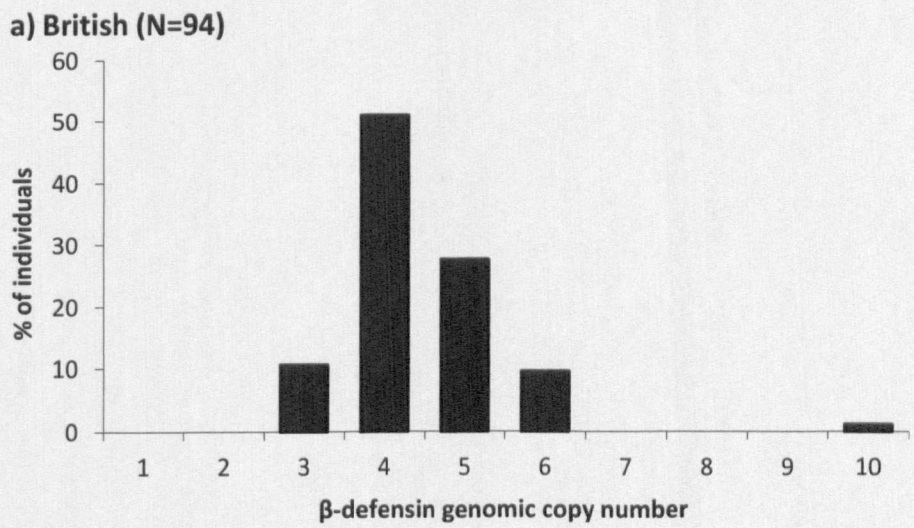
4.2.2 Results

To investigate the geographical variation of the β-defensin copy number among different human populations, the Triplex assay was used to type copy number in 6 different populations: Yoruba from Ibadan, Nigeria (YRB); Dutch from Nijmegen, Netherlands; British, UK; Utah residents with Northern and Western European ancestry from the CEPH collection, USA (CEU); Japanese from Tokyo, Japan (JPT); Han Chinese from Beijing, China (CHB).

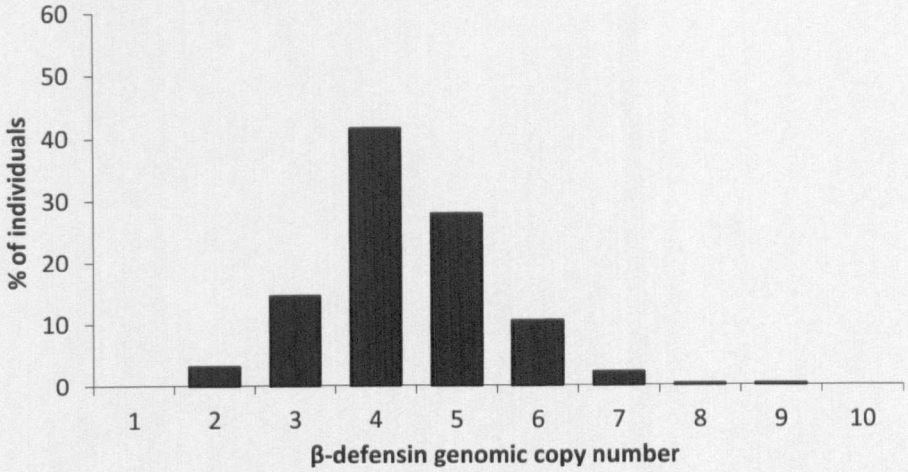
Clustering analysis of the unrounded copy numbers from PRT107A and HSPD21 for each population indicated the accuracy of both PRT systems with

the distribution of values centred in integer values (scatter-plots not shown). Moreover, as explained in section 3.2.1, the copy number values obtained from the Triplex assay for the HapMap CEU, YRB, CHB and JTP cohorts were validated with data previously published by different studies (McCarroll *et al.* 2008; Abu Bakar *et al.* 2009; Conrad *et al.* 2010). The agreement observed between copy number values obtained from different studies, range from 86.08% to 93.51% and confirms the accuracy of the Triplex assay.

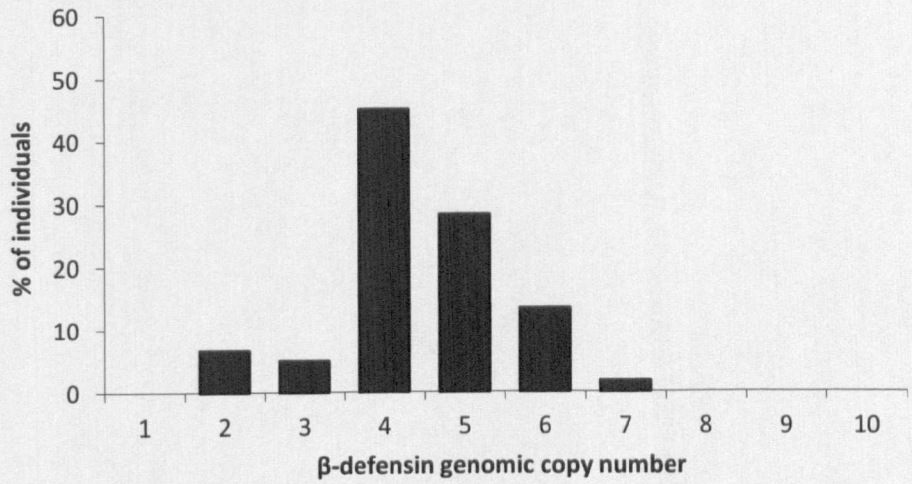
As shown in the histograms below (Figure 34), the β-defensin copy number varies generally between 2 and 10 copies per diploid genome, with copy numbers up to 7 being the most common. The modal diploid copy number in all populations is 4 copies (mean for UK=4.43, Dutch=4.36, JPT=4.51, CHB=4.09, YRI=4.57, CEU=4.42) however, slight differences in copy number frequencies appear to move the copy number distribution towards higher values in Yoruba and Japanese populations as compared with European populations. A one-way analysis of variance (ANOVA) was used to test the amount of variation that occurs between populations. Despite the slight differences observed in the copy number distribution, β-defensin copy number variation was not significantly different between the human populations studied (*p*-value=0.261, ANOVA).



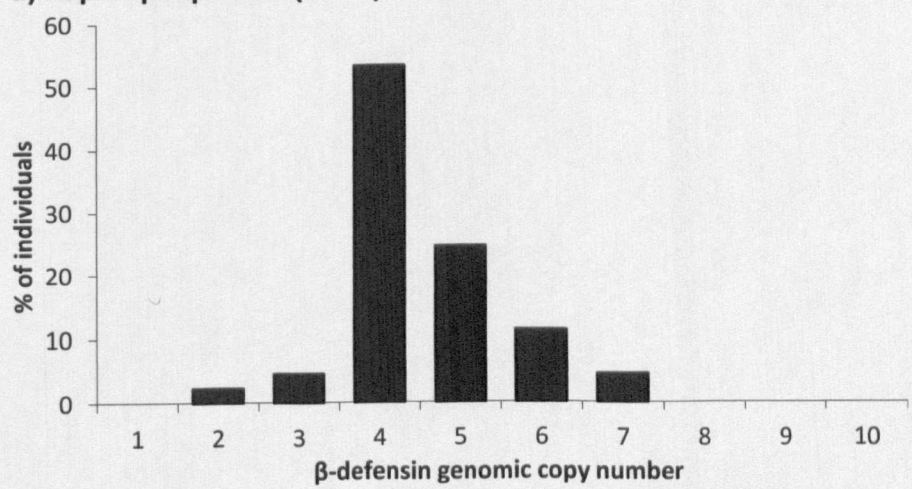
b) Dutch (N= 516)



c) HapMap CEU (N=60)



d) HapMap Japanese (N=45)



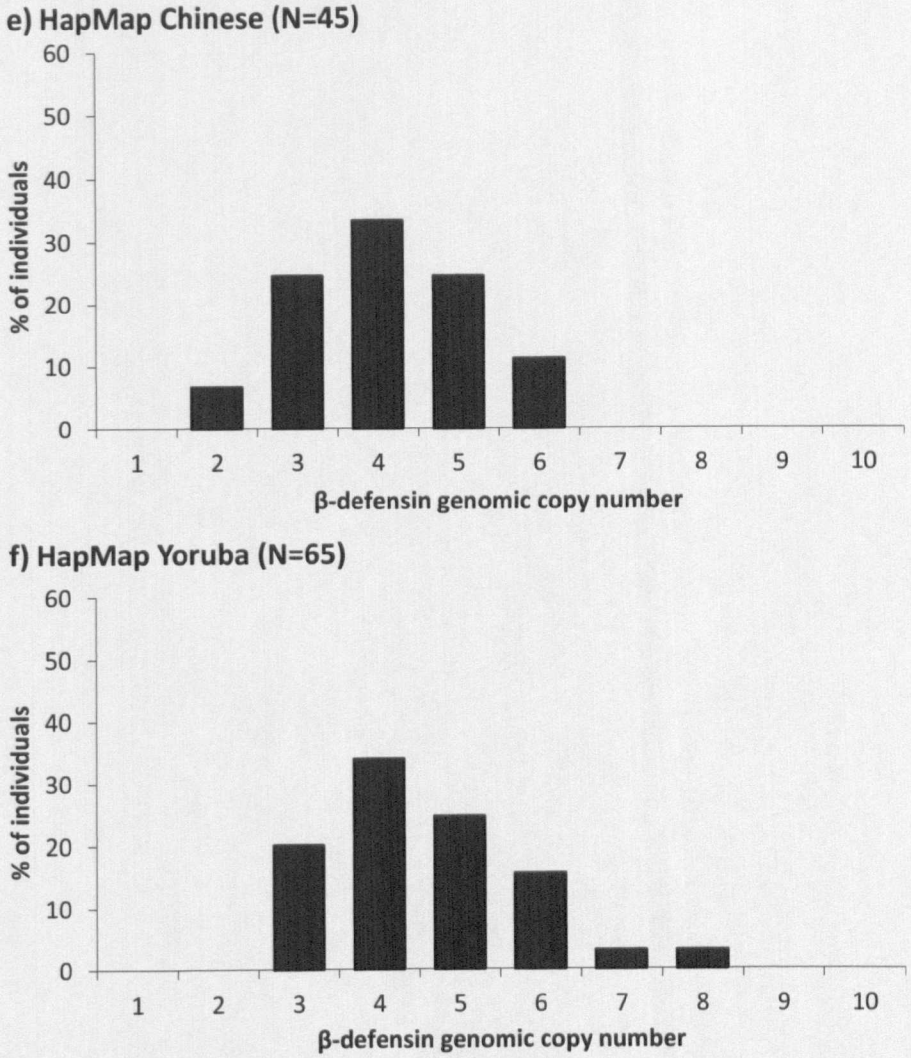


Figure 34: β-defensin diploid copy number distribution in six different populations: British, UK (a); Dutch from Nijmegen, Netherlands (b); CEPH individuals with northern and western European ancestry from Utah, USA (c); Japanese from Tokyo, Japan (d); Chinese from Beijing, China (e) and Yoruba from Ibadan, Nigeria (f). The copy numbers were typed by the Triplex assay.

4.2.3 Discussion

Several CNVs show large differences in diploid frequencies and distribution between human populations, but this was not observed for the β-defensin locus. The data obtained from the genotyping of β-defensin showed no significant differences between the 6 populations studied, although the mean copy number in Japanese and Yoruba was slightly higher than other populations. The slight increase in copy number observed for Japanese and Yoruba populations could be due to genetic drift, different mutation rates or selective pressures or an artefact due to small sample size. The similarity of

copy number distribution between the European populations clearly reflects the geographical proximity between them; nevertheless this does not explain the similarity between Europeans and Yoruba. Japanese and Chinese populations showed the most differentiated distributions. In the first case there is a clear shift towards high copy numbers, for example the frequency for 7 copies was higher than for 2 and even 3 copies. Conversely, the high frequency of lower copy numbers, such as 2 and 3 copies, makes the Chinese copy number distribution unusual.

The observation of the same average copy number in all populations indicates that this locus is not under different strong selective pressures. This suggests that average copy number of 4 might be a good balance in all populations. Since the β-defensin copy number has been correlated with the *DEFB4* mRNA levels (Hollox *et al.* 2003), 4 copies represents an intermediate level of expressed β-defensin, not too low or too high, which may protect against the risk of infectious and/or inflammatory diseases. These findings agree with the previous analysis of the β-defensin cluster at 8p23.1 by Hollox and Armour (2008), describing the absence of positive selection at this locus in the last 25 million years.

Finally, the data reported here was carefully analysed and the accuracy of the Triplex assay was, once again, demonstrated by the direct comparison with results previously published for the HapMap collection cohorts (section 3.2.1). The results obtained in this study for those samples showed a high agreement (>86%) with the published data, thus demonstrating the high performance of the PRT-based Triplex assay for copy number measurement of β-defensin. Moreover, this also demonstrates the vast applicability of the multiplex system to different human populations, showing the assay is not population dependent.

4.3 8p23.1 INVERSION DUPLICATION: A FAMILY CASE

4.3.1 Background

Rearrangements in genome structure are an important source of variation in the human genome. This variation can include deletions and duplications, and thus variation in copy number, but also orientation and positional polymorphisms. The implications of such structural variation are known to induce effects on gene and protein expression, leading to phenotypic and disease susceptibility variations. A classical example of a region showing several types of variation is the chromosomal band 8p23.1, the region in which the defensin genes reside. The olfactory repeat regions (OR), REPP and REPD, that flank the 8p23.1 locus, predispose to an increased number of chromosomal rearrangements due to recurrent recombination between the two olfactory regions, giving rise to large imbalances in the 8p chromosome arm (Giglio *et al.* 2001). Non-allelic homologous recombination (NAHR) occurs frequently between REPD and REPP and is the mechanism that leads to the deletion or duplication of the sequence that separates these two olfactory repeats, commonly referred as 8p23.1 deletion syndrome (Devriendt *et al.* 1995; Wat *et al.* 2009) and the reciprocal duplication syndrome (Barber 2005; Barber *et al.* 2007), respectively. Both syndromes are characterized by phenotypic manifestations, including developmental delay, behavioural problems and congenital heart disease associated with the deletion syndrome, while mild dysmorphism and developmental and speech delay are associated with the duplication syndrome (Barber 2005; Barber *et al.* 2007). In addition to these rearrangements, high copy number variants of β-defensin can easily be misinterpreted as duplications of the REPD/REPP interval region as observed by microscopy, but they can be clearly distinguished at the molecular level (Barber *et al.* 2005).

A common large structural inversion polymorphism (4.7 Mb) involves the region flanked by the olfactory repeat gene clusters (REPD and REPP) at the 8p23.1 region. This polymorphism is present in 25% of Europeans and approximately 33% of Japanese are heterozygous for the inversion polymorphism (Giglio *et al.* 2001; Giglio *et al.* 2002). The presence of these

polymorphisms can lead to unequal recombination between the olfactory receptor regions and generate a variety of recurrent chromosomal rearrangements including inverted duplications and deletions of 8p (inv. dup. del. (8p)) and its reciprocal rearrangement. This results in a supernumerary marker chromosome (+der(8) (8p23.1pter)) that contains the distal 8p region duplicated (Giglio *et al.* 2002; Sugawara *et al.* 2003). The phenotype of patients with such large chromosomal imbalances is characterized by facial dysmorphism, agenesis of the corpus callosum, development delay, mental retardation and congenital heart diseases (Guo *et al.* 1995).

This chapter describes a study carried out in a Cologne family trio, kindly supplied by Raoul Heller from Utrecht (Netherlands), in which the child (patient) has a terminal hemizygous deletion extending into REPD and a duplication starting at REPP extending up to roughly 20 Mb from pter. The origin of this chromosomal rearrangement is unknown. The mother is a healthy individual (karyotype: 46, XX) but the father has a supernumerary marker chromosome at 8p (karyotype: 47, XY) showing evidence of mental retardation and recorded epilepsy episodes during adolescence (Figure 36). With the aim of identifying the parental origin of the β-defensin repeats found in the child, the measurement of β-defensin copy number was carried out in the three members of this family with the Triplex assay and a series of multiallelic markers (microsatellite analysis). This analysis aimed to shed further light on the structural origins of 8p23 rearrangements and if possible to reconstruct the pattern of β-defensin segregation in this family.

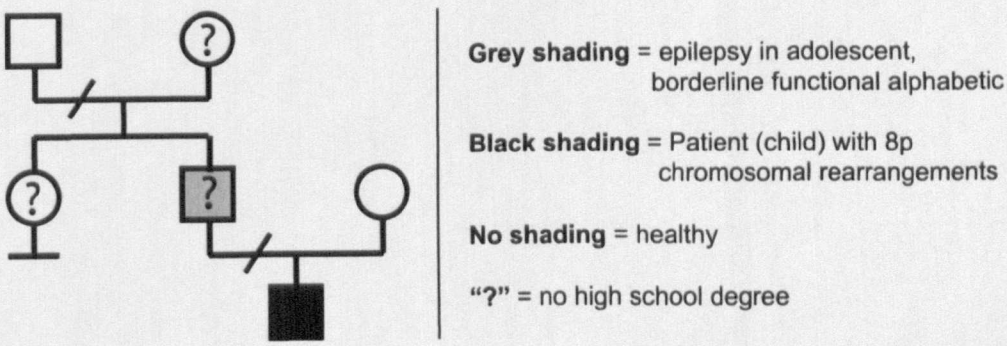


Figure 35: Cologne family pedigree showing the closest relatives of the patient with 8p chromosomal rearrangements.

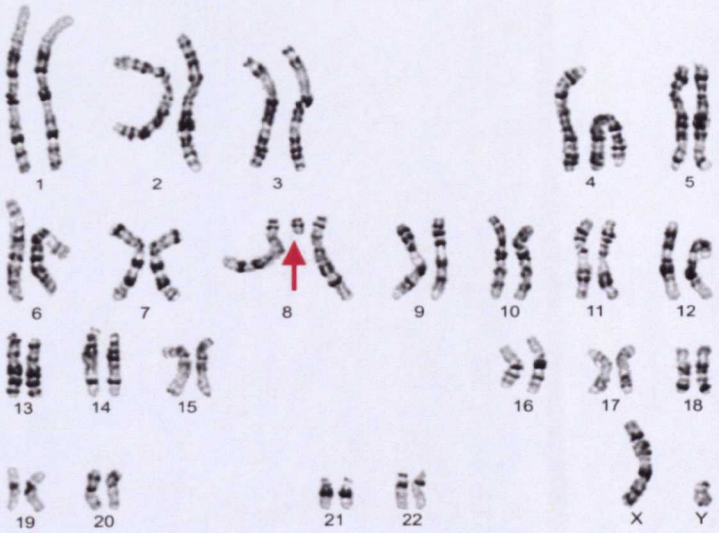


Figure 36: Paternal karyotype, 47, XY, +der(8) (8p23.1pter). The red arrow in the figure indicates the supernumerary marker chromosome at 8p (kindly provided by Dr. Regine Shubert).

4.3.2 Results

Each sample was typed several times with the Triplex assay and copy numbers were calculated from peak heights of each PCR product using a maximum likelihood analysis (section 2.2.7). The Triplex results agreed in unique copy number value of 4 copies for "mother", 6 copies for "father" and 5 copies for "patient" (Figure 37). Regarding the 5DEL analysis on its own, the ratio between the resultant alleles suggested copy number values based on multiples of 4, 6 and 5, respectively for "mother", "father" and "patient" (Figure 37). Furthermore, different copy number haplotypes can be ascertained from the pattern shown by 5DEL analysis (Table 14). However, segregation analysis does not support the exclusion of any paternal or maternal haplotypes in order to follow the segregation of alleles from parents to child.

In addition to the Triplex assay, microsatellites (EPEV3, EPEV1 and EPEV5) and an additional indel assay (9 bp indel) were used. These methodologies, which were mainly used to attempt to follow the segregation of the alleles from parents to child, could also confirm the copy number given by the Triplex assay.

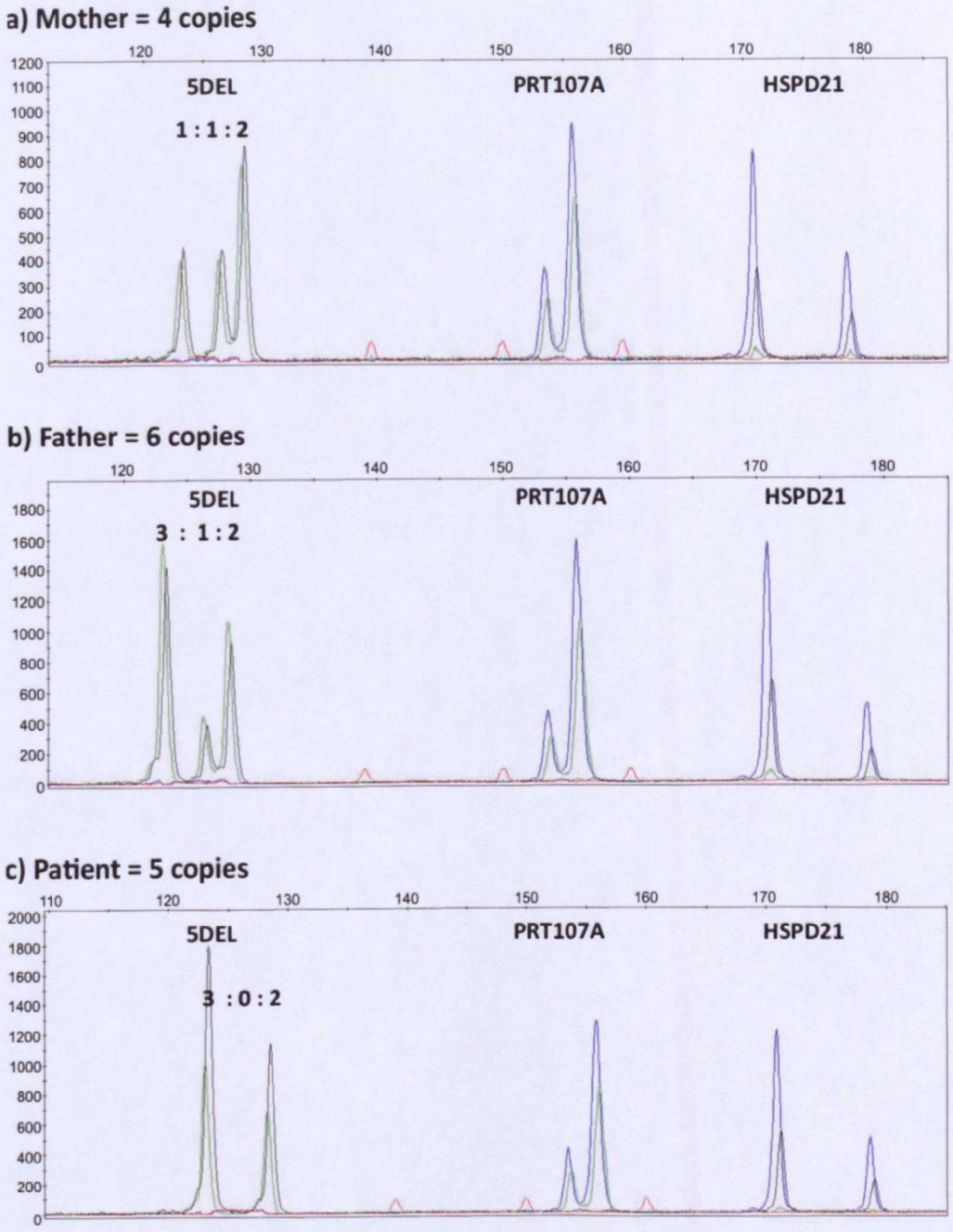


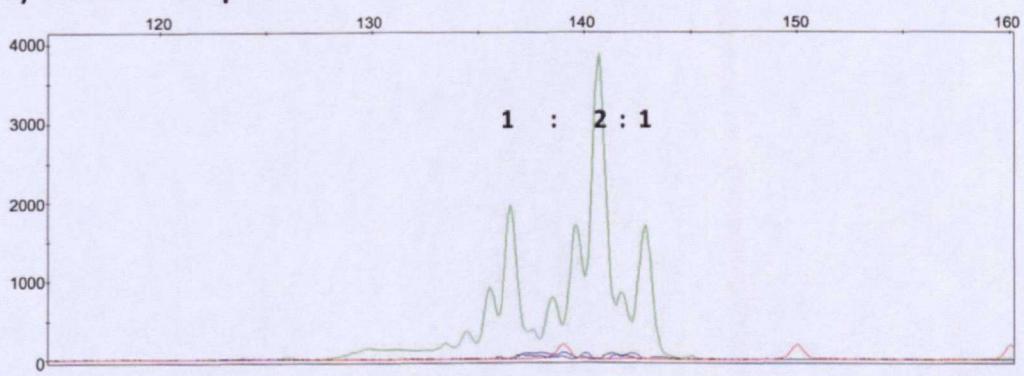
Figure 37: Examples of the GeneMapper electropherogram for Triplex assay after capillary electrophoresis from "mother" (a), "father" (b) and "patient" (c). In each electropherogram is visualised the PCR products of each system (5DEL, PRT107A and HSPD21) in two distinct fluorescent dyes.

Table 14: Allele ratios obtained from 5DEL analysis for "mother", "father" and "patient" (child) were used to define copy number values and predict haplotypes and segregation patterns. The copy number values (CN) represented here were obtained from the Triplex assay. In this analysis the possibility of zero-copy haplotypes was disregarded.

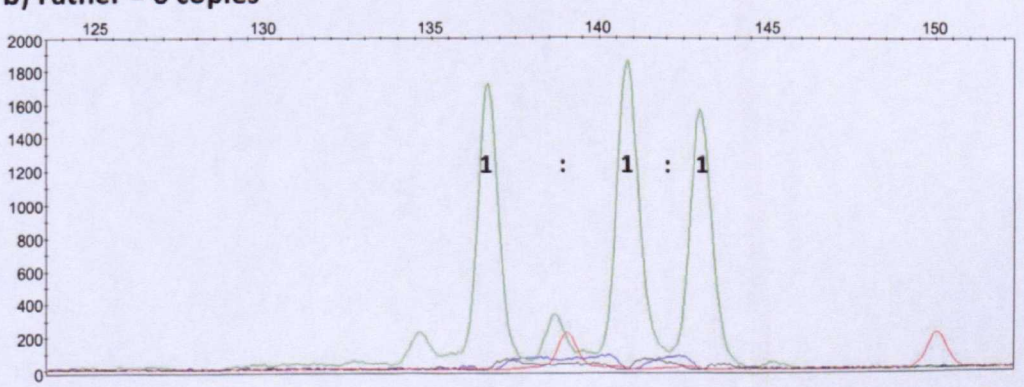
	Alleles			CN	Possible Haplotypes
	123	126	128		
Mother	1	1	2	4	3:1 or 2:2
Father	3	1	2	6	2:4 or 3:3
Patient	3	0	2	5	2:3 or 4:1

The electropherograms in Figure 38 show the alleles obtained from the EPEV3 microsatellite analysis and the ratio between them. The ratios suggest copy number multiples of 4, 3 and 5 for "mother", "father" and "patient", respectively. From the allelic information given by the EPEV3 analysis and the copy number measured earlier by Triplex assay, the possible combination of haplotypes can be determined for each individual (Table 15). The "patient" could have either inherited 2 copies from one parent and 3 copies from the other or 4 copies from "father" and 1 copy from "mother" (Table 15).

a) Mother = 4 copies



b) Father = 6 copies



c) Patient = 5 copies

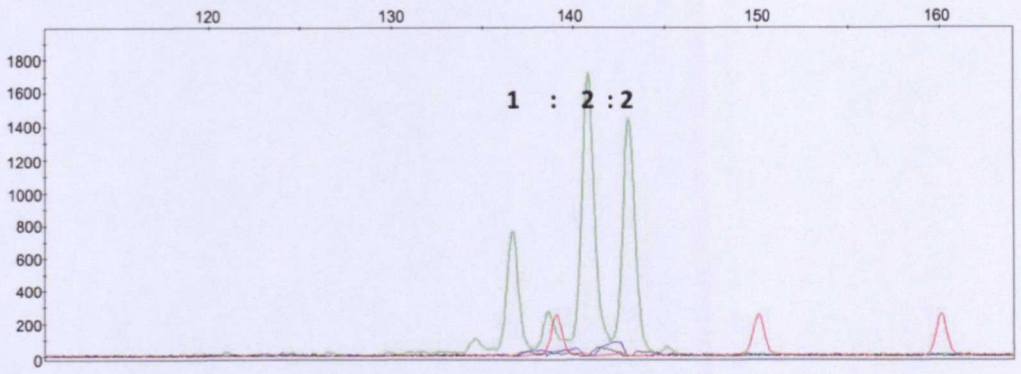


Figure 38: Examples of the GeneMapper electropherogram for EPEV3 assay after capillary electrophoresis from "mother" (a), "father" (b) and "patient" (c). In each electropherogram is visualised the alleles (136 bp, 140 bp and 142 bp) obtained from EPEV3 and the ratios between them.

Table 15: Allele ratios obtained from EPEV3 analysis for "mother", "father" and "patient" (child) were used to define copy number values and predict haplotypes and segregation patterns. The copy number values (CN) represented here are given by the Triplex assay. In this analysis the possibility of zero-copy haplotypes was disregarded.

	Alleles			CN	Possible Haplotypes
	136	140	142		
Mother	1	2	1	4	3:1 or 2:2
Father	2	2	2	6	2:4 or 3:3
Patient	1	2	2	5	2:3 or 4:1

In Figure 39, the three electropherograms, corresponding to the EPEV1 microsatellite allelic pattern of each sample, are not easy to interpret due to the presence of "stutter peaks". However, after "stutter peak" correction (section 2.2.5) the allelic ratios suggested that "mother" has a copy number value multiple of 4, "father" has 6 or 12 copies and "patient" has 5 or 10 copies of the β-defensin repeat at 8p23.1. Table 16 shows the combination of copy number haplotypes possible for each individual taking in consideration the copy number obtained from Triplex analysis and the pattern observed for EPEV1 microsatellite. Following the segregation of the EPEV1 alleles from parents to child, three possible haplotypes, corresponding only to two different copy number states, can be inherited from each parent, as shown in Figure 40. The possible maternal haplotypes inherited include two with one copy number of the locus and one with two copies, while the paternal haplotypes include one with a three copy number state and two with 4 copies.

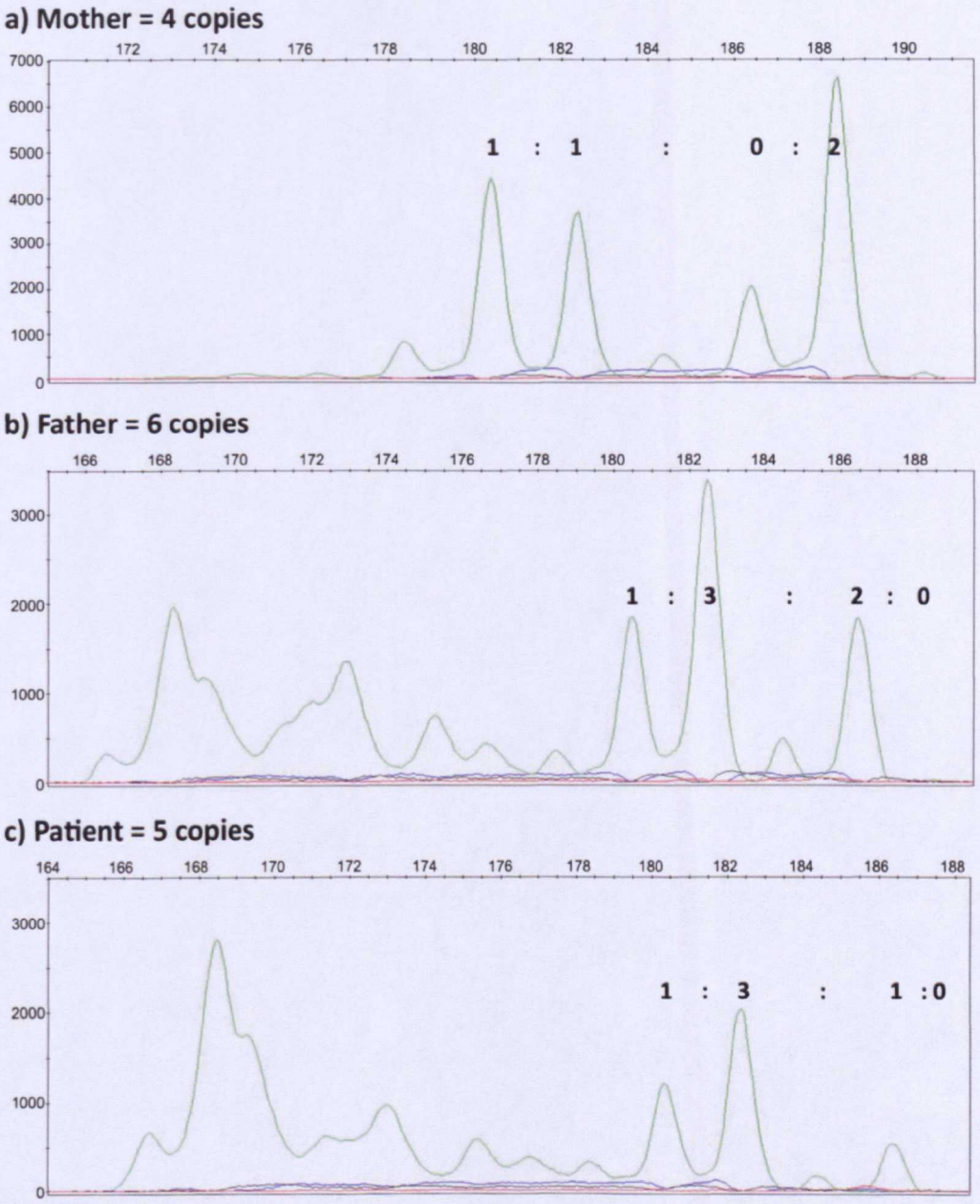


Figure 39: Examples of the GeneMapper electropherogram for EPEV1 assay after capillary electrophoresis from "mother" (a), "father" (b) and "patient" (c). In each electropherogram the alleles obtained from EPEV1 are visualised (180 bp, 182 bp, 186 bp and 188 bp) and the ratios between them displayed, after "slippage peak" correction.

Table 16: Allele ratios obtained from EPEV1 analysis for "mother", "father" and "patient" (child) were used to define copy number values and predict haplotypes and segregation patterns. In this analysis the possibility of zero-copy haplotypes was disregarded.

	Alleles				CN	Possible haplotypes
	180	182	186	188		
Mother	1	1	0	2	4	3:1 or 2:2
Father	1	3	2	0	6	2:4 or 3:3
Patient	1	3	1	0	5	2:3 or 4:1

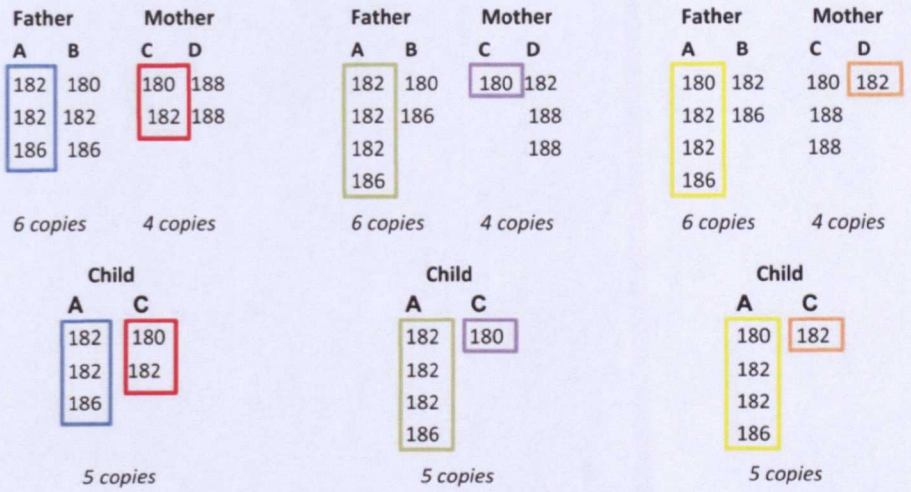
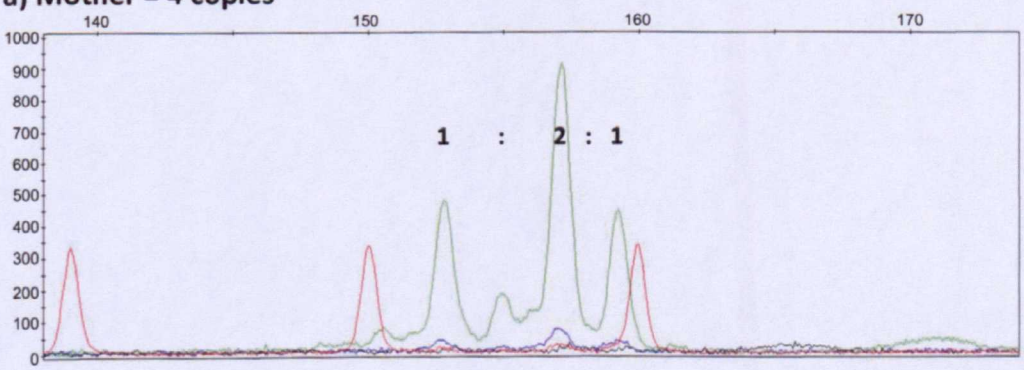


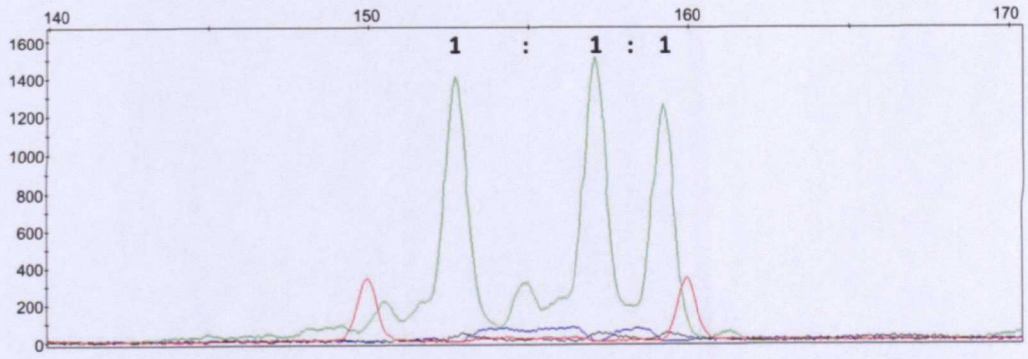
Figure 40: Possible haplotypes suggested from the EPEV1 analysis (microsatellite) for "mother", "father" and "patient" (child).

An additional microsatellite, EPEV5, was developed to further clarify the segregation in this family. The microsatellite patterns obtained from the EPEV5 analysis showed the presence of the same three alleles in all samples (Figure 41). Therefore, it is only possible to predict that the child ("patient") has inherited either 2 copies from one parent and 3 copies from the other or 4 copies from "father" and 1 copy from "mother" (Table 17). However, the ratios between the alleles clearly indicate different copy number values, corresponding to 4, 3 and 5 or multiples of them for "mother", "father" and "patient", respectively.

a) Mother = 4 copies



b) Father = 6 copies



c) Patient = 5 copies

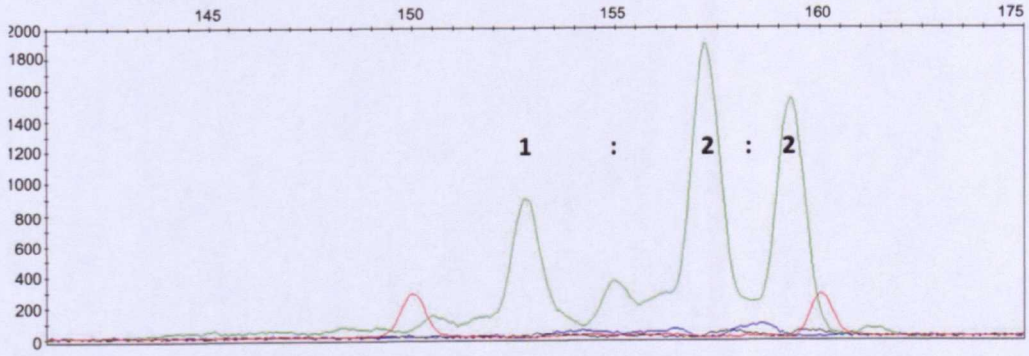


Figure 41: Examples of the GeneMapper electropherogram for EPEV5 assay after capillary electrophoresis from "mother" (a), "father" (b) and "patient" (c). In each electropherogram the alleles (152 bp, 157 bp and 159 bp) obtained from EPEV5 are visualised and the ratios between them displayed, after "slippage peak" correction.

Table 17: Allele ratios obtained from EPEV5 analysis for "mother", "father" and "patient" (child) were used to define copy number values and predict haplotypes and segregation patterns. In this analysis the possibility of zero-copy haplotypes was disregarded.

	Alleles			CN	Possible Haplotypes
	152	157	159		
Mother	1	2	1	4	3:1 or 2:2
Father	2	2	2	6	2:4 or 3:3
Patient	1	2	2	5	2:3 or 4:1

Finally, a last attempt to clarify the segregation in this family was carried out using the 9 bp indel system (Abu Bakar 2010). This assay was designed to detect a 9 bp deletion within the *DEFB4* gene; therefore only two alleles were possible, corresponding to the deleted (297 bp) and inserted 9 bp sequence (306 bp). As shown in the Figure 42, just one allele was observed for "father" and "patient", which make the determination of the copy number impossible.

However, due to the presence of the two alleles in the "mother", the allele ratio suggested a copy number value of 4. The 9 bp indel analysis also suggested the same possible inherited haplotypes, established earlier by 5DEL, EPEV3 and EPEV5, for these individuals (Table 18).

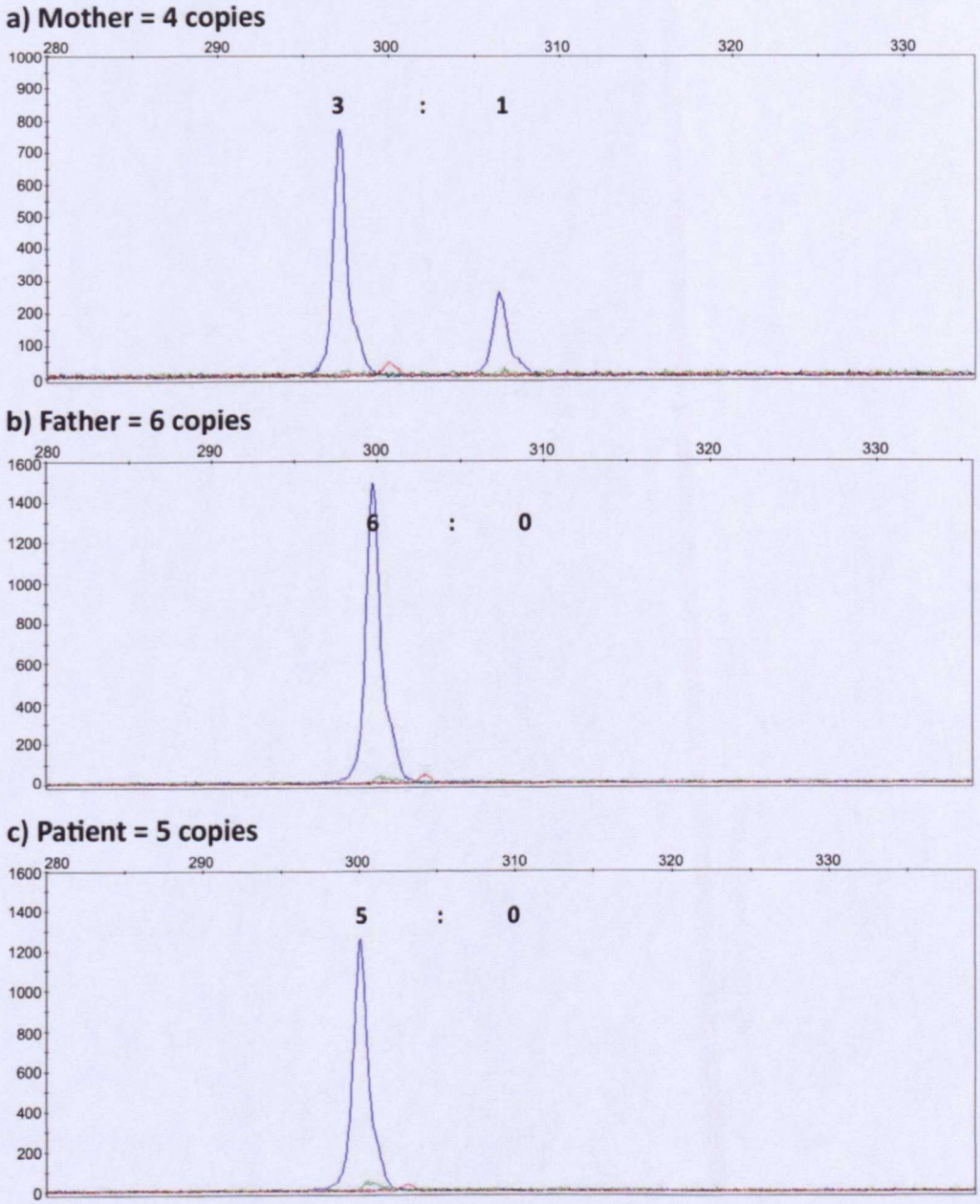


Figure 42: Examples of the GeneMapper electropherogram for 9 bp indel assay after capillary electrophoresis from "mother" (a), "father" (b) and "patient" (c). In each electropherogram the alleles (297 bp and 306 bp) obtained from 9 bp indel are visualised and the ratios between them displayed.

Table 18: Allele ratios obtained from EPEV5 analysis for "mother", "father" and "patient" (child) were used to define copy number values and predict haplotypes and segregation patterns. In this analysis the possibility of zero-copy haplotypes was disregarded.

	Alleles		CN	Possible Haplotypes
	297	306		
Mother	3	1	4	3:1 or 2:2
Father	6	0	6	2:4 or 3:3
Patient	5	0	5	2:3 or 4:1

4.3.3 Discussion

The β-defensin copy number values for the Cologne family trio were well established by the Triplex assay and strongly supported by the microsatellites and indel assays used. While the Triplex assay was able to ascertain precise copy number values for each sample, the microsatellites and indel assays determined ratios between the different alleles suggesting copy numbers based on multiples of the values obtained. Combining the information from all assays, a consensus copy number of 4, 6 and 5 was established for "mother", "father" and "patient" (child), respectively.

Despite the copy number values suggested by 5DEL, EPEV3 and EPEV5 strongly supporting the copy numbers determined by the Triplex assay for each sample, unfortunately these assays were not particularly helpful to completely understand the segregation in this family. The allelic pattern given by these assays was very similar between the three members of this family making it difficult to follow the segregation of the alleles unambiguously from parents to child and so uncover the haplotypes of each sample. Therefore, from these analyses it was only possible to establish the inheritance of either 2 copies from one parent and 3 copies from the other or 4 copies from "father" and 1 copy from "mother". The analysis carried out in this study did not take in consideration the possibility of zero-copy haplotypes, since the frequency of a diploid copy number of one is very low in all populations reported, so far.

While the analysis of EPEV1 microsatellite was informative, the analysis of 9 bp indel assay on its own did not add any further information. However, taking into consideration the combined analysis of EPEV1 and 9 bp indel

assay, it was possible to narrow down the number of potential inherited haplotypes from parents to child. The diagram below shows the possible haplotypes constructed from the analysis of the 9 bp indel (Figure 43). Adding information from the EPEV1 analysis and 9 bp indel assay it is possible to exclude the third hypothesis (indicated with a red cross), so the microsatellite information from EPEV1 only suggests two possible inherited haploid copy number states from each parent; 3 or 4 copies from "father" and 1 or 2 copies from "mother". Therefore, the third hypothesis suggested by the 9 bp indel assay, in which the child would inherit 2 copies from "father" and 3 copies from "mother" was rejected. So, the "patient" could only inherit 3 or 4 paternal copies and 1 or 2 maternal copies.

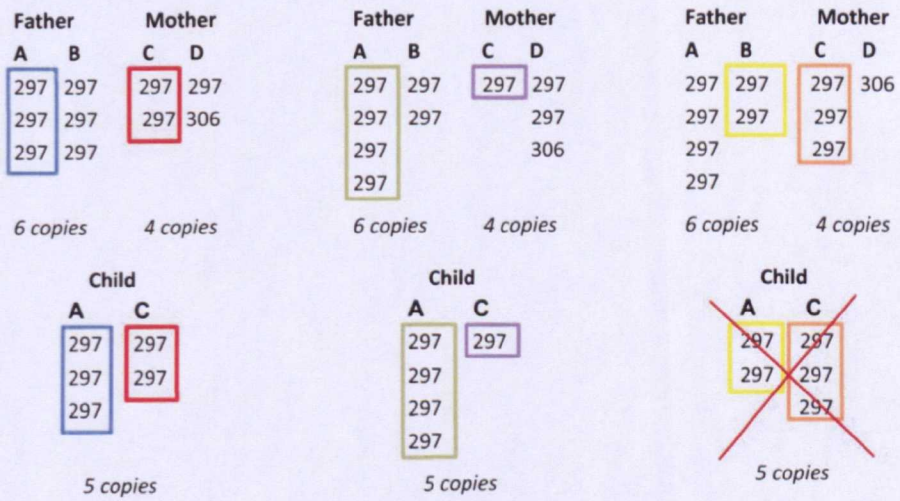


Figure 43: Possible haplotypes suggested from 9 bp indel analysis for "mother", "father" and "patient" (child). Taking into consideration the information given by the EPEV1, it is possible to exclude (crossed in red) one of the combinations.

The analysis carried out here follows the classic segregation pattern, assuming the existence of just two haplotypes; however, father has abnormal karyotype and the segregation patterns could be much more complicated. So, the repeats inherited by child from father may not all be on one chromosome and it is possible that arrangement changes on transmission from father to child. For this reason it would be difficult to go much beyond defining the parental origins of repeats found in the child.

4.3.4 Conclusion

The copy number typing systems used could confidently establish the copy number of 4, 6 and 5 for "mother", "father" and "patient", respectively. Information gathered from the microsatellite and indel analysis did not completely clarify the number of copies that the "patient" inherited from each parent, but contributed to a certain extent to understanding the segregation of alleles in this family, leaving the haplotypes to be identified. Nevertheless, it was shown that the "patient" could only inherit 3 or 4 paternal copies and 1 or 2 maternal copies. These analyses, although efficient in detecting the copy number of these individuals, were not able to completely clarify the origin of the 8p rearrangements observed in the "patient". In order to clearly understand the segregation of alleles in this family it would be necessary to access samples from other family members, such as grandparents or siblings, but unfortunately information was very limited and further samples were not available.

CHAPTER 5: STUDY OF SNPs IN LD WITH THE β-DEFENSIN CNV

5.1 INTRODUCTION

Most common simple deletion and duplication polymorphisms (biallelic CNVs) can be tagged by at least one neighbouring SNP with some even showing a perfect match with proxy SNPs (r^2=1.0) (Hinds *et al.* 2006; McCarroll *et al.* 2006; Redon *et al.* 2006; McCarroll *et al.* 2008). Such characteristics suggest that these polymorphisms share a similar evolutionary history and originated from a unique ancestral mutational event that happened once during human evolutionary history. As a consequence, these variants appear in high linkage disequilibrium (LD) with nearby SNPs and can effectively be tagged by proxy markers (Hinds *et al.* 2006; McCarroll *et al.* 2006; Redon *et al.* 2006). However, the level of linkage disequilibrium around CNVs is not comparable with the LD observed around SNPs (Redon *et al.* 2006; McCarroll *et al.* 2008). This apparent difference could be due to recurrent rearrangements at CNV sites or due to poor coverage of SNPs at repeat-rich regions (Sharp *et al.* 2005; Locke *et al.* 2006; Redon *et al.* 2006). Although SNPs can be use to tag biallelic CNVs, in the case of multiallelic

copy number variants, due to their structurally dynamic nature and complex evolutionary origin, this might not apply. For example at multiallelic CNVs, such as the β-defensin cluster at 8p23.1, Redon *et al.* (2006) described that diploid copy number is poorly predicted by neighbouring SNPs.

Recombination between the distal and proximal sites of 8p23.1 is expected at least in 4% of transmitted chromosomes. Additionally, this CNV shows a germ-line rate of copy number change of 0.7% per gamete, the fastest-changing CN variant currently known (Abu Bakar *et al.* 2009). This evidence suggests that it is very unlikely to find SNPs in strong LD with β-defensin copy number, to allow the prediction of the copy number from flanking SNPs. However, combining information for several SNPs from the proximal and distal sites may provide greater power to predict genomic copy number at these loci. In previous studies SNP analysis just took into account the presence of β-defensin genes at the distal site (REPD) (Redon *et al.* 2006), but a recent study by Abu Bakar *et al.* (2009) showed evidence of a β-defensin copy number variable repeat at the proximal site (REPP), as well. At REPP the copy number haplotypes are more variable than at the distal site, which highlights the importance of including information from both repeat sites, REPP and REPD, when searching for tag SNPs.

To investigate if the β-defensin copy number can be tagged by neighbouring SNPs, both REPP and REPD sites of the β-defensin copy number repeat were investigated. The combination of SNP information from both sites of the β-defensin repeat will contribute to a more realistic and accurate approach to investigate SNPs in LD with copy number. Thus the present study hopes to address the question: can the copy number of β-defensin be predicted by nearby SNPs?

5.2 RESULTS

The Haploview software was used to search for SNPs in LD with both copy number repeats at REPD and REPP site of 8p23.1. SNP analysis was performed using the CEPH/CEU SNP haplotype data from the International HapMap Project collections and the haploid copy number data generated for the same sample cohort by Abu Bakar *et al.* (2009). For CEPH/CEU samples only the diploid copy number was known from using the Triplex assay, so haploid copy numbers were inferred when possible, following the copy number segregation in families. In this analysis, 36 unrelated (CEU parents only) samples (72 haplotypes) were used.

Since Haploview software is designed to just process haplotype analysis of SNPs, in order to analyse the data, copy number data was coded as pseudo-SNP genotypes. Copy numbers were coded in two different ways: in the first instance, 1 and 2 copies were coded as A (adenine) and 3 or more copies as G (guanine), while in the second instance, 1 copy was coded as A and 2 or more copies as G. The adopted coding allowed the introduction of copy number data together with SNP data in the software and thus searches for SNPs in LD with copy number. The SNP data used for the CEPH/CEU population was collected from the HapMap Phase II and III across 1 Mb region at the proximal site (11,700,700-12,700,000) and across approximately 2 Mb around the distal site (6,730,000-8,500,000), each including the respective β-defensin cluster.

At the proximal site of 8p23.1 region three SNPs were found to be in LD with the copy number of β-defensin. rs6989065 and rs12548700 (r^2=0.103, D'=1 and LOD=2.48), mapped around 430 kb downstream (12,653,559 and 12,654,325) of the repeat unit at REPP site, show complete LD with the copy number, as supported by the value of D'=1. Additionally, these SNPs are in strong LD with each other. The third SNP, rs6530976 is located at about 450 kb downstream (12,654,325) of the β-defensin cluster and is in weak LD with the other two SNPs, mentioned above. The measures of LD between rs6530976 and the copy number (r^2=0.116, D'=0.833 and LOD=2.23) do not indicate complete LD but values of D' near 1 (D'=0.833) provides a useful indication of

disruption of LD by recombination. These LD measures indicate a significant association of lower magnitude, weak LD (D'<1 and r^2<1).

At the distal site of 8p23.1 no SNPs were found to be in LD with β-defensin copy number. As the three SNPs found were in LD with each other, just one SNP, rs12548700, was typed in the ECACC Human Random Control (HRC) panels 1 and 2 and in CEPH/CEU cohort from the International HapMap Project collections to test the association further.

5.2.1 Proximal site (REPP) of 8p23.1 - rs12548700

To genotype rs12548700 and investigate the correlation with copy number, an RFLP assay was designed to test the association as described in section 2.2.6.1. This assay was applied in 181 samples of ECACC Human Random Control (HRC) and in 90 CEPH/CEU samples from the HapMap cohorts to confirm the accuracy of genotypes. The histograms below show the genotype frequency distribution according to the copy number in the two cohorts studied (Figure 44). Genotypes for the ECACC cohorts (Figure 44a) are widely distributed across the different copy number haplotypes and all genotypes show a similar mean copy number (between 4.31 and 4.59) (Table 19). From the genotype frequency distribution represented in the histogram it is not possible to visualize any association between each genotype and a particular copy number. In the second histogram (Figure 44b) it is possible to observe that 2 copy number haplotype is only represented by the G/G genotype, but this is based on a very small number of samples (3 samples) with 2 copies. Furthermore, the genotype C/C is apparently shifted towards high copy numbers (4 or more copies), showing a higher mean copy number (mean CN=5) than the other two genotypes (Table 19). The mean copy number observed for each genotype ranged between 4.22 and 5, a slightly broader range than the one observed for ECACC samples.

To test if there was any association between copy number and SNP genotypes a correlation test was applied. No significant association was found between these SNP genotypes and β-defensin copy number for any of the cohorts studied (*p*-value=0.31, Pearson's correlation test, ECACC; *p*-value=0.14 Pearson's correlation test, CEPH/CEU HapMap).

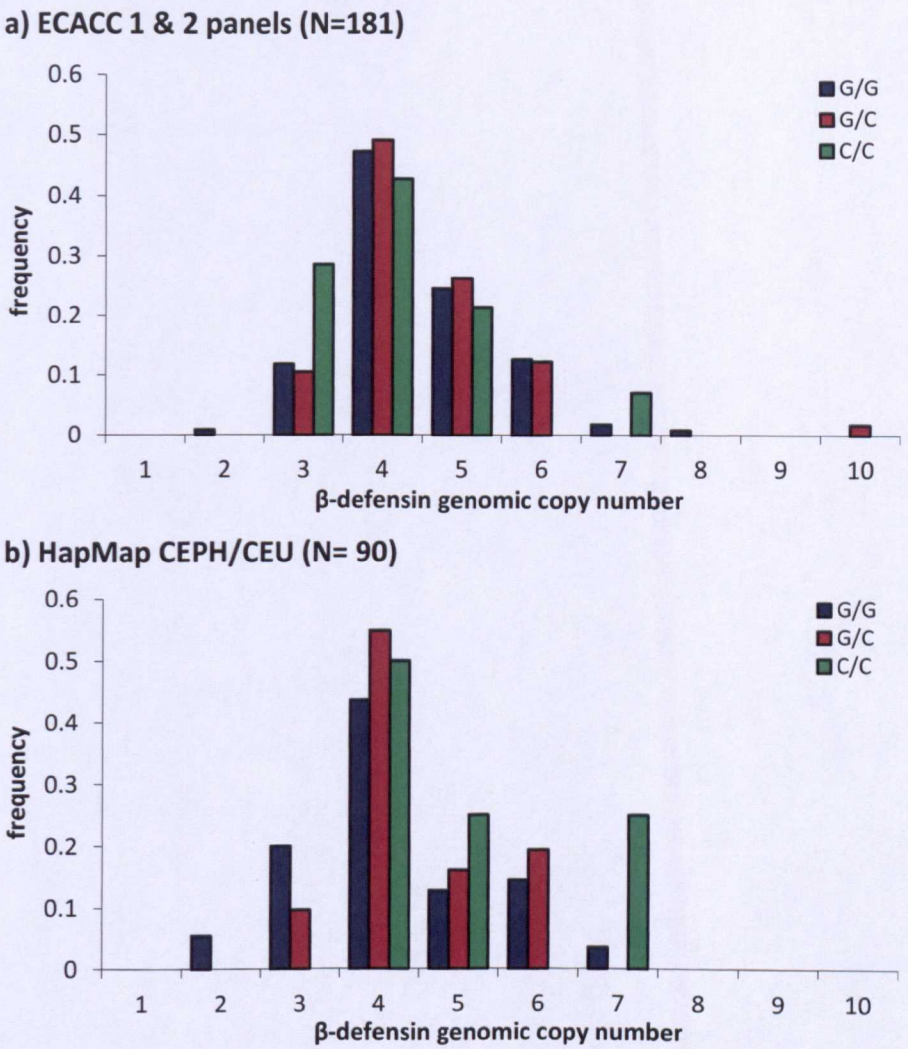


Figure 44: Genotype frequency distribution of rs12548700 for ECACC Human Random Control (HRC) panels 1 and 2 (a) and CEPH/CEU HapMap cohort (b). Genotype frequency distribution is represented according to β-defensin copy number.

Table 19: Genotype results of rs12548700 for the two cohorts studied: ECACC 1 and 2 panels and HapMap CEPH/CEU cohort. The number of samples found with each genotype, as the mean copy number and genotype frequency are indicated for each genotype. Allele frequencies for each cohort are also shown below.

Cohort	Genotypes	Number of samples	Mean CN	Genotype frequency	Allele	Allele Frequency
ECACC 1 & 2 panels	G/G	54	4.59	0.6279	G	0.7924
	G/C	26	4.31	0.3023	C	0.2641
	C/C	6	4.33	0.0698		
	Total	86				
HapMap CEPH/CEU	G/G	55	4.22	0.6111	G	0.7817
	G/C	31	4.45	0.3444	C	0.2108
	C/C	4	5.00	0.0444		
	Total	90				

5.3 DISCUSSION

The results obtained from the genotype of rs12548700 show that this SNP did not tag the copy number of β-defensin at the REPP site, showing no association in ECACC or HapMap CEPH/CEU samples. The apparently significant association initially found in HapMap haplotypes between the rs12548700 and β-defensin copy number for the CEPH/CEU cohort was not confirmed after SNP genotyping, suggesting a type I error.

Since no other SNPs were found in LD with either REPP or REPD site of β-defensin copy number repeat, it suggests that different repeats arose at different times in the human evolutionary history, hence with different genetic "footprints". While most common diallelic CNPs are perfectly captured (r^2=1.0) by at least one SNP tag, multiallelic copy number loci with recurrent CNV formation would show reduced correlation with flanking markers (McCarroll *et al.* 2008). The results obtained are supported by previous evidence showing the high rate of recombination of β-defensin loci that contributes to the continuous formation of new repeats (Abu Bakar *et al.* 2009). Evidence from Redon *et al.* (2006) also agrees with present results; in that study a low Pearson correlation coefficient (r^2) (<0.58) was described for all the 13 multiallelic CNVs studied, showing the poor ability of neighbouring SNPs in predicting the copy number. Among the CNVs studied, the correlation of the best SNP proxy to the β-defensin repeat at REPD site was tested showing a very poor correlation with copy number (r^2=0.20) (Redon *et al.* 2006).

Our study confirms that multiallelic copy number variants are unlikely to be tagged by nearby markers. As such, the study of copy number variation cannot rely on the use of proxy SNPs. Instead, accurate and specific assays, such as the Triplex assay, should be developed, as they are especially important to measure multiallelic copy number variants such as the β-defensin copy number repeat.

CHAPTER 6: β-DEFENSIN CNV IN NON-HUMAN PRIMATES

6.1 INTRODUCTION

CNVs in non-human primates have been investigated, so far, in a very limited number of chimpanzees and rhesus macaques (Perry *et al.* 2006; Lee *et al.* 2008; Perry *et al.* 2008). Furthermore, very little is known about CNV in other great apes, such as orangutan and gorilla. The sequence data available for these primate species is very limited, which makes the study of genomic variation, including CNV, difficult. Genes that are involved in immune response often show high rates of genomic divergence and evidence of adaptive evolution, in response to the rapid evolution of pathogens (Semple *et al.* 2005). The β-defensins are a family of multifunctional genes with key roles in immunity, but also in reproduction and pigmentation. In primates, infection is an important driving force for evolution modulating the survival rate to infectious disease (Hollox and Armour 2008). Therefore, investigation of CNV, such as the CNV at β-defensin locus in great ape could shed light into the origin of CNV in humans and their evolution through the primate lineage.

For rhesus macaque (*Macaca mulatta*), Lee *et al.* (2008) reported multiple copy number gains and losses at the β-defensin locus, including changes in *DEFB4*, using array-CGH. In the same year, using array-CGH and FISH, Perry *et al.*, (2008) showed that the beta-defensin CNV overlaps between humans and chimpanzee (*Pan troglodytes*). The CNV of β-defensins in chimpanzee was recently analysed in more detail by Hardwick *et al.* (2011) using PRT, where copy numbers of 4, 5 and 6 were reported. All genome comparative analysis of chimpanzee, orangutan and rhesus macaque with the human genome shows that β-defensin region is highly conserved in chimpanzee but not as much in orangutan or rhesus macaque (Vista Genome Browser). Here, the newly developed PRT-based assays (PRT107A and HSPD21) were applied to study the recent evolution of β-defensin CNV, investigating the copy number of chimpanzee (*Pan troglodytes)*, gorilla (*Gorilla gorilla*) and orangutan (*Pongo* sp.) species (section 2.1.1.4).

6.2 RESULTS

Considering the genome sequence similarity between the human and other great ape genomes, it is expected that some of the CN measurement systems developed for humans could also amplify orthologous regions of great ape species. To measure the β-defensin copy number in non-human primates, the possibility of using each of the Triplex systems developed for humans was tested in chimpanzee and orangutan by *in silico* PCR and BLAT search using the UCSC genome browser (CGSC 2.1.3/panTro3 and WUGSC 2.0.2/ponAbe2). To date, no sequence data is available in UCSC genome browser for gorilla, so the Ensembl genome browser (gorGor1 (Release 57, Mar2010)) was used instead. However, the sequence data available for gorilla from Ensembl is very incomplete for the β-defensin region and just one β-defensin gene (*SPAG11*) is annotated in the gorilla genome.

In chimpanzee, only HSPD21 PRT primers showed no mismatches with the chimpanzee genome, amplifying uniquely test (chromosome 8) and reference loci (chromosome 21). In orangutan neither of the PRTs (HSPD21 or PRT107A) could be applied and all measurement systems tested showed several mismatches under the primers in the orangutan genome. For HSPD21 primers, no matches were found on chromosome 8 in the orangutan genome, but the primers matched the reference locus on chromosome 21. Further analysis of test and reference loci in the orangutan and chimpanzee genome was carried out by multiple sequence alignment with the human genome using ClustalW2 software (from the European Bioinformatics Institute (EBI) (Figure 45). The results from the alignment confirm the findings for chimpanzee and orangutan. On orangutan chromosome 21, the reference locus showed just one mismatch in each primer, with a transition mismatch in the penultimate 3' position of the forward primer, while on chromosome 8 very poor sequence similarities were found for the HSPD3 pseudogene. This seems to show a sequence gap in the orangutan genome when compared with human and chimpanzee genomes, suggesting that this copy of the processed pseudogene was inserted after the orangutan divergence from the human-chimpanzee common ancestor.

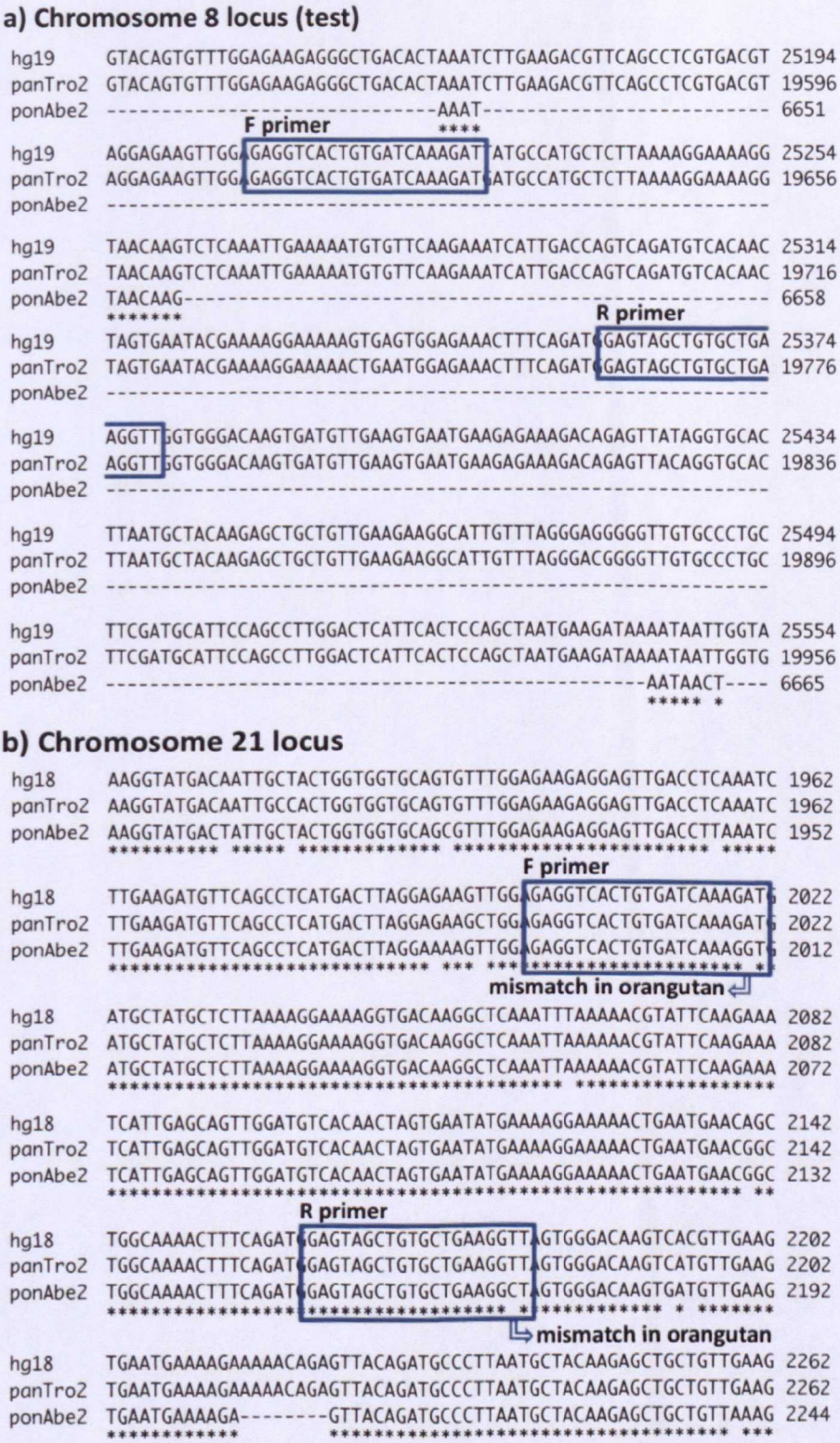


Figure 45: Multiple sequence alignment of test, chromosome 8 (a), and reference locus, chromosome 21 (b), for human (hg18), chimpanzee (panTro2) and orangutan (ponAbe2) genomes using the ClustalW2 software.

To investigate the compatibility of the HSPD21 PRT system with the gorilla genome, the sequence data available from Ensembl was investigated for test (chromosome 8) and reference (chromosome 21) loci. Despite the fact that all β-defensin genes were partially sequenced in Ensembl, only *SPAG11* was annotated for gorilla genome. From sequence alignments of test and reference loci for human and gorilla it was possible to confirm the presence of the test locus on gorilla chromosome 8 (HSPD3 pseudogene), with both primers having only one mismatch with the gorilla genome, with the forward primer showing an internal transition mismatch (Figure 46). However, no matches were found for the reference locus on chromosome 21, suggesting either the absence of the HSPD3 pseudogene in gorilla genome or the poor coverage of the gorilla genome sequence available in Ensembl.

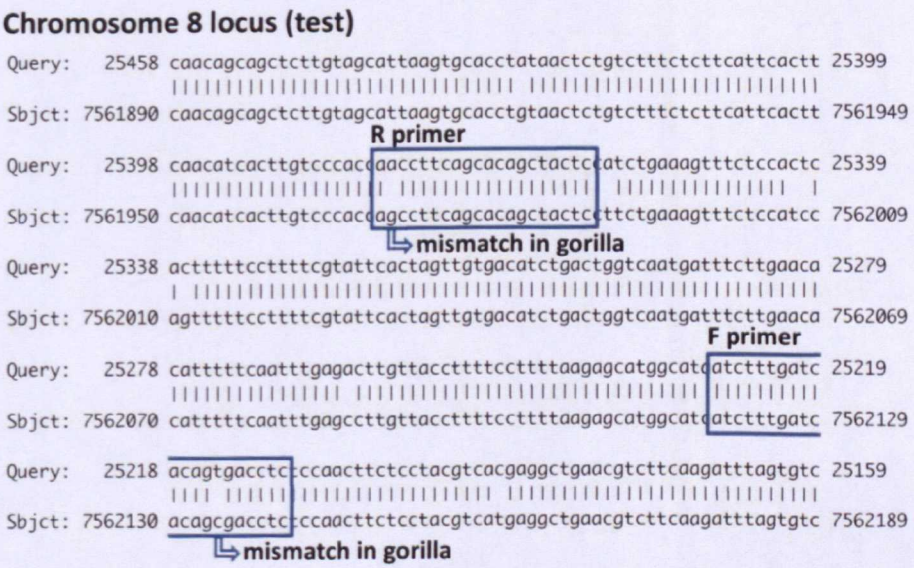


Figure 46: Multiple sequence alignment of test (chromosome 8) for human (Query) and gorilla (Sbjct.) genomes using the ClustalW2 software, showing positions of primers for HSPD21 PRT.

To further investigate the existence of HSPD3 pseudogene on chromosome 21 of gorilla (reference locus of HSPD21 PRT), sequencing of this locus was performed as described in section 2.2.8.1. Sequencing of the chr21 locus was carried out, as well, in all other non-human primates to give additional information about this locus in chimpanzee and orangutan and to confirm the presence of this locus in bonobo since no sequence data is available for this species. The sequencing data obtained for each individual is displayed in Figure 47, showing a multiple sequence alignment comparing with the human genome. As shown, the sequencing results

demonstrated the presence of HSPD3 pseudogene in gorilla with no mismatches observed in the forward and reverse primer of HSPD21 PRT. Moreover, the sequencing data obtained for the other primates, chimpanzee, bonobo and orangutan also confirms the presence of the HSPD3 pseudogene supporting the use of HSPD21 PRT to measure the β-defensin copy number in great ape.

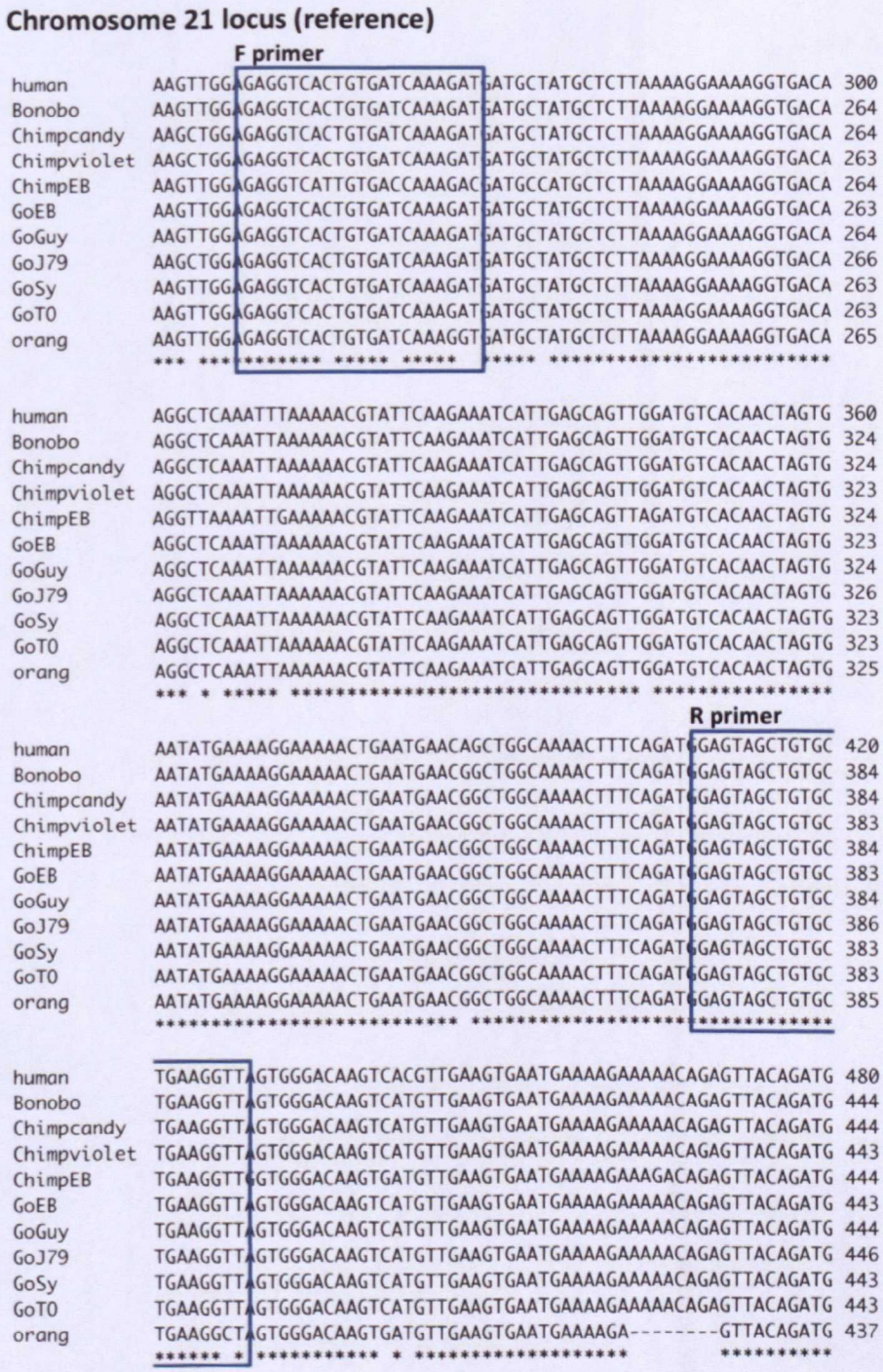
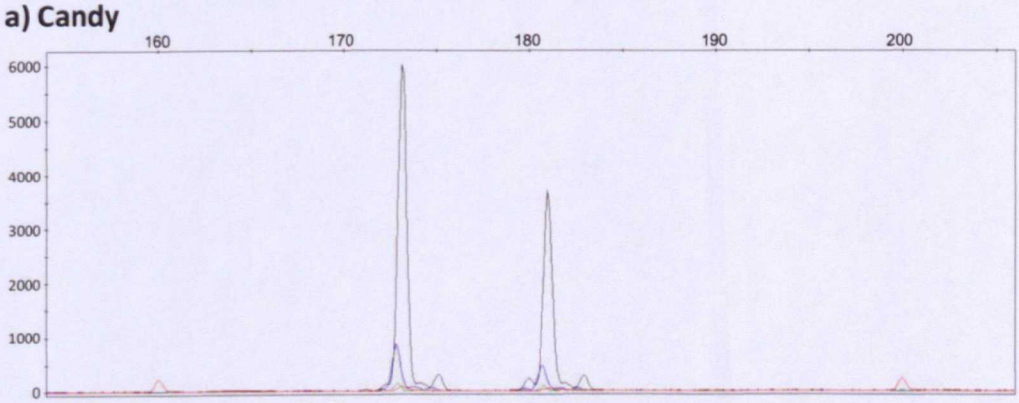


Figure 47: Multiple sequence alignment of reference locus (chromosome 21) for human, bonobo, chimpanzee (Candy, Violet and EB), gorilla and orangutan using the ClustalW2 software, showing the positions of the HSPD21 PRT primers.

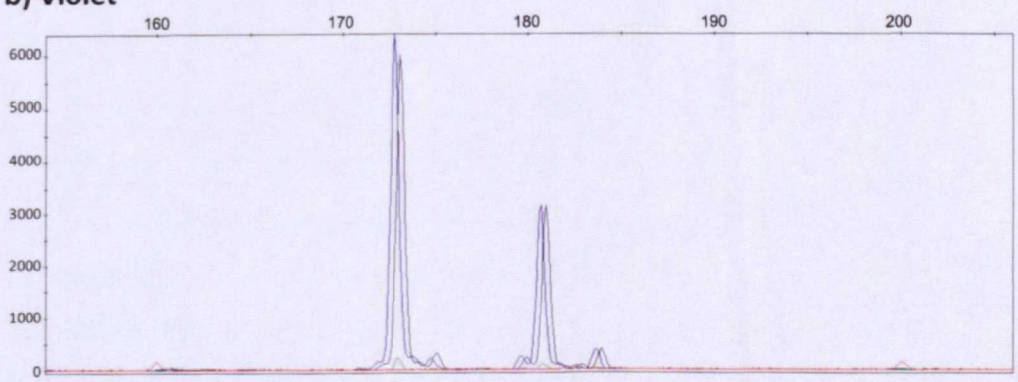
6.2.1 Chimpanzee

Here, the HSPD21 PRT system (section 2.2.3.2) was used to measure the β-defensin copy number in three chimpanzees (*Pan troglodytes*); Candy, Violet and EB and one bonobo (*Pan paniscus*). Each sample was typed with two differently labelled primers (NED and FAM) (section 2.2.3.2), resulting in two independent measurements of the β-defensin copy number (Figure 48). Copy numbers were calculated from the test/reference ratio using peak heights, as described in section 2.2.3.4.

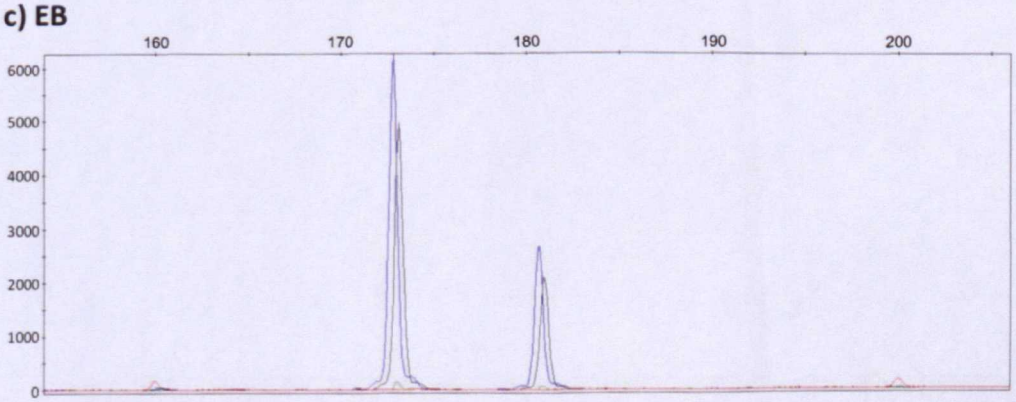
a) Candy



b) Violet



c) EB



d) Bonobo

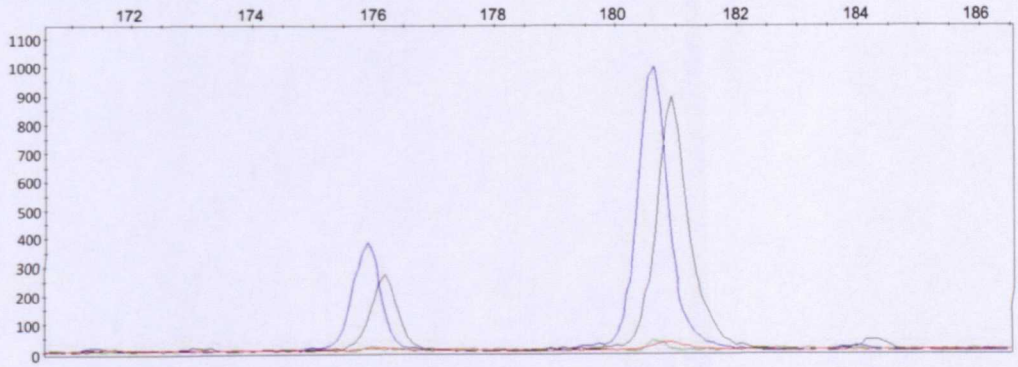


Figure 48: GeneMapper electropherogram for HSPD21 assay after capillary electrophoresis from Candy (a), Violet (b) and EB (c) and Bonobo (b). In each electropherogram test (172 bp) and reference (180 bp) loci are shown in two distinct fluorescent dyes.

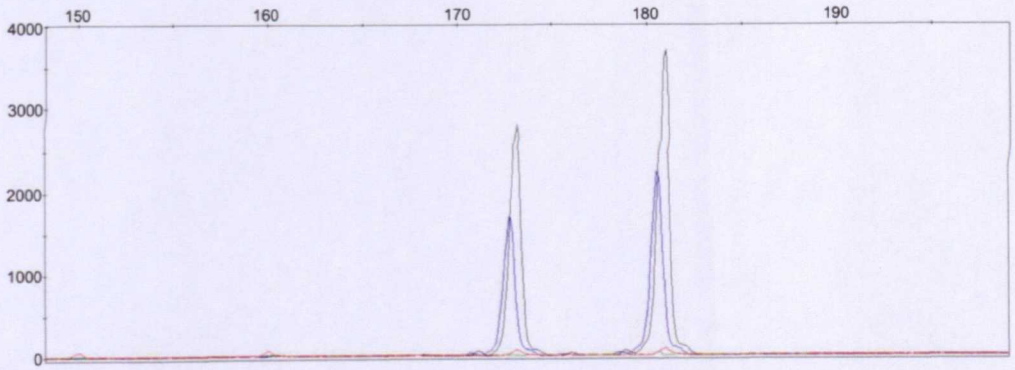
The duplicate HSPD21 PRT measurements agree between each other for each of the samples, indicating a copy number of 1 for bonobo, 3 for Candy and 4 for Violet and Chimpanzee EB when calibrated against human samples of known copy number.

6.2.2 Gorilla

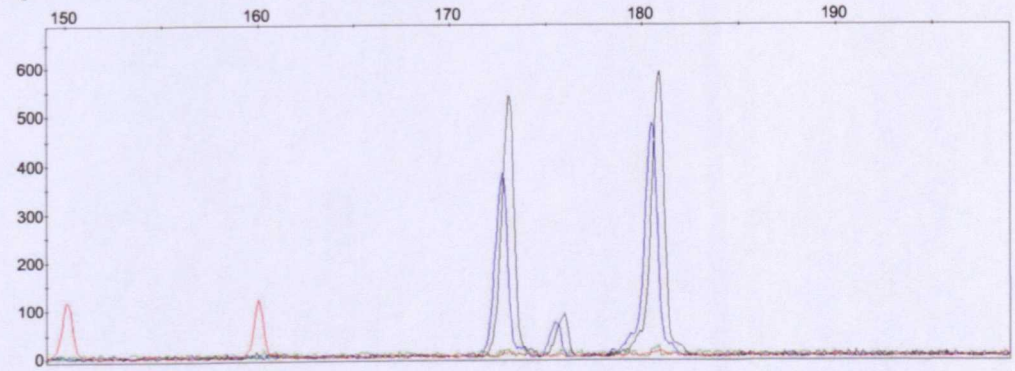
6.2.2.1 HSPD21 PRT

To measure the β-defensin copy number in 5 gorilla individuals (Sylvia, Tomoka, Guy, J79 and EB JC), a modified version of the HSPD21 PRT system was used (section 2.2.3). This included a gorilla-specific forward primer, which takes into consideration the mismatch observed in the gorilla sequence (Figure 46), used alongside the human specific forward primer in the same PCR reaction (section 2.2.3.2). Each sample was typed with two differently labelled primers (NED and FAM) (section 2.2.3.2) and the ratio between test and reference locus, taken from peak heights, was used again to predict the copy number (section 2.2.3.4). The copy number measurements taken from each labelled primer agreed between each other for all samples typed and mean copy numbers were calibrated against human samples of known copy number. For all five gorillas (Sylvia, Tomoka, Guy, J79 and EB JC) an integer copy number of 2 it was suggested by HSPD21 PRT (unrounded copy numbers from 1.79 to 1.95) (Figure 49).

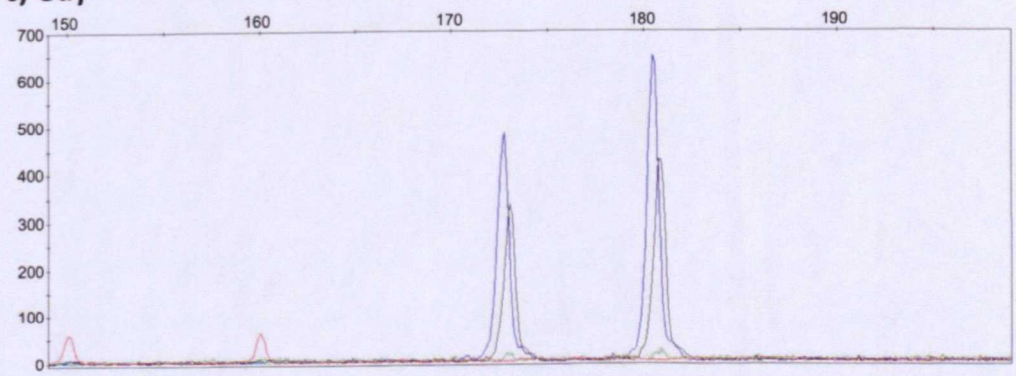
a) Sylvia



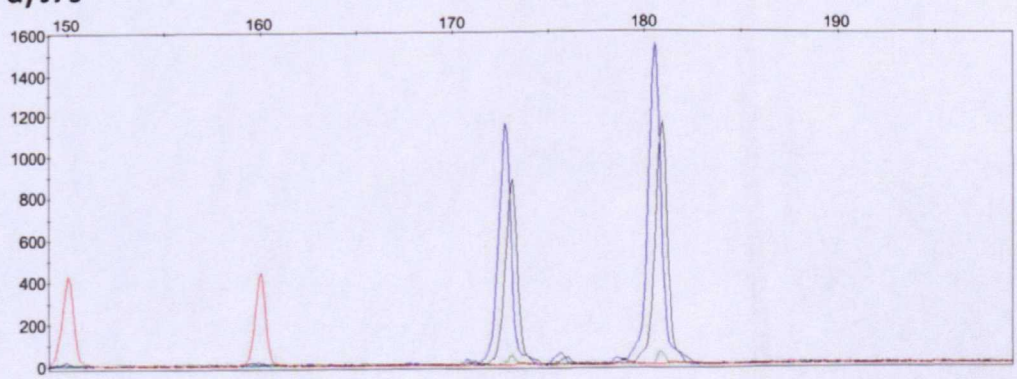
b) Tomoka



c) Guy



d) J79



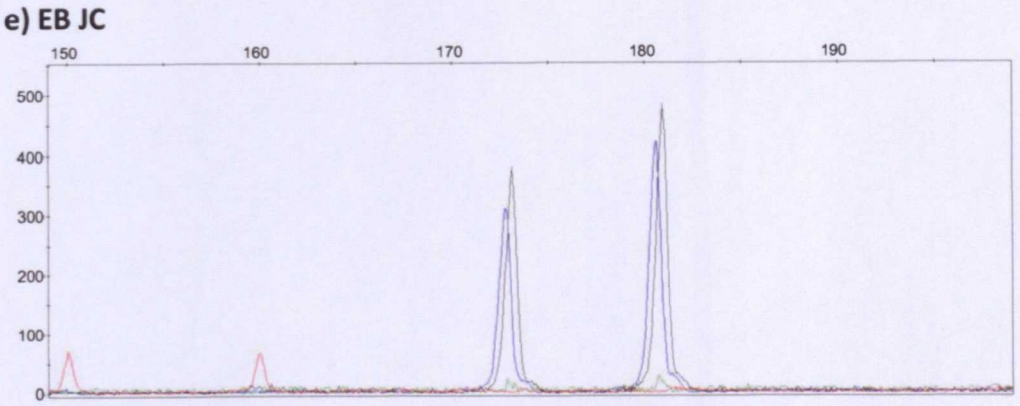


Figure 49: GeneMapper electropherogram for HSPD21 PRT assay after capillary electrophoresis from Sylvia (a), Tomoka (b), Guy (c), J79 (d) and EB JC (e). In each electropherogram, test (172 bp) and reference (180 bp) loci are shown in two distinct fluorescent dyes.

6.2.2.2 *DEFB103* locus sequencing

To further support the evidence of two copies of β-defensins in the gorilla genome, DNA sequence analysis was undertaken at the *DEFB103* locus to investigate sequence polymorphisms within this locus for the 5 different gorilla individuals. By following the segregation of the sequence variants in different haplotypes, it should be possible to distinguish the sequences of each β-defensin copy. In the presence of only two different haplotypes, the copy number of two can be confirmed.

Sequencing data obtained for the *DEFB103* locus was obtained for ~95% of the 1.8 kb sequence (chr8: 7273602-7275433 and chr8: 7775983-7777814, including the entire coding sequence) by using two forward and three reverse primers (section 2.2.8.2, Table 8). Sequencing data does not include poly T repeats within this locus. Only three of the five gorillas, Sylvia, Tomoka and J79 showed sequence variants at this locus. In total, four sequence variants were found to convincingly represent the presence of two different alleles at the same position (Figure 50). Sequence variants 1, 2 and 3 were found in the gorilla J79 and another sequence variant, 4, was found in both Sylvia and Tomoka. For all sequence variants the two alleles are approximately equal in peak height (Figure 51).

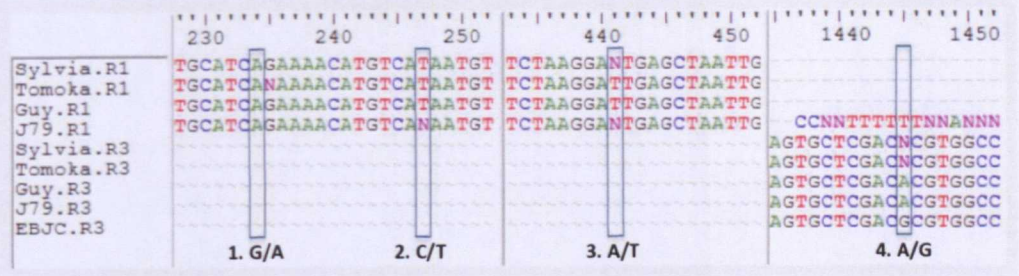


Figure 50: *DEFB103* locus multiple sequence alignment of sequencing data obtained by using DEFB103R and DEFB103R3 primers for gorilla (Sylvia, Guy, Tomoka, J79 and EBJC) and human (J80, C11 and C62). In the figure, the three sequence variants in J79 (1, 2 and 3) and the sequence variant (4) found in both Sylvia and Tomoka are shown.

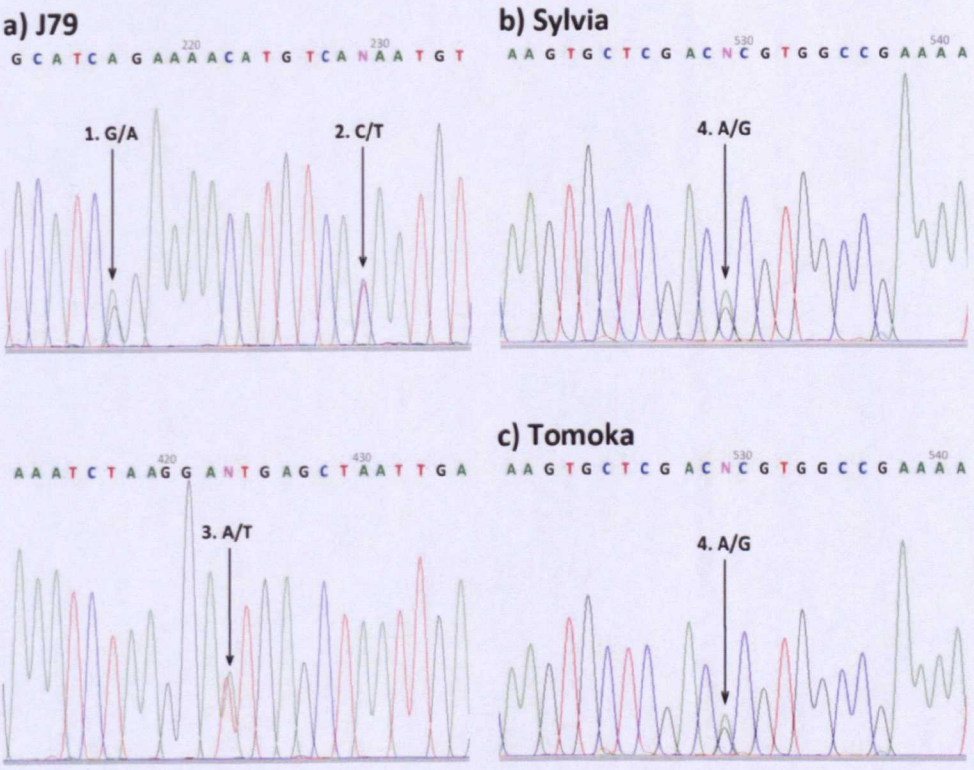


Figure 51: Patch of sequence traces collected from *DEFB103* locus sequencing using DEFB103R and DEFB103R3 primers. Sequence variants 1, 2 and 3 are shown for J79 (a) while sequence variant 4 is shown in Sylvia (b) and Tomoka (c).

6.2.2.3 Allele-specific PCR

In order to follow the segregation of alleles in different haplotypes, allele specific primers were designed for the relevant mixed positions found in J79, Sylvia and Tomoka from the *DEFB103* locus sequencing (section 2.2.8.3). The sequence variant 1 was used to design an allele specific primer that would

allow the analysis of the downstream sequence variants (2 and 3) in gorilla J79. For Sylvia and Tomoka, an allele specific primer was designed from the only sequence variant found (4) in these individuals.

The sequence data obtained for J79 from allelic specific primers was always of very poor quality, even after several attempts. Therefore, no further information about the segregation of the alleles was obtained and no further clarification about the β-defensin copy number of J79 was possible.

The sequence traces shown in Figure 52, show that two alleles at variant 4 in Sylvia and Tomoka were discriminated, confirming the presence of only alleles A and G at this position. However, with only one mixed position found at this locus for each of the gorillas it is not possible to reconstruct their haplotypes.

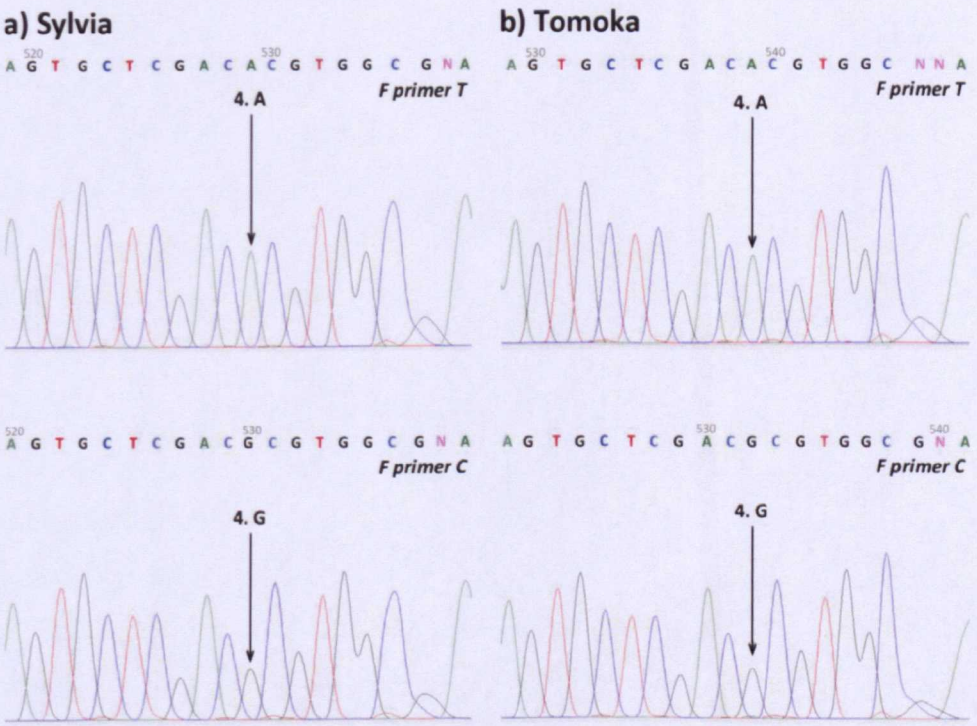


Figure 52: Patch of sequence traces collected from *DEFB103* locus sequencing using the allele specific primers DEFB103F_Sy.To_T and DEFB103F_Sy.To_C. For each gorilla, Sylvia (a) and Tomoka (b) are shown the separated alleles after the allele specific PCR at the sequence variant (4).

6.3 DISCUSSION

To investigate the evolutionary history of beta-defensin CNV in primates, HSPD21 PRT was used to measure the copy number of β-defensins in chimpanzee, bonobo, gorilla and orangutan. Great ape and human genomes have high DNA sequence similarity (at least 97%) (Locke *et al.* 2011). After comparing orthologous sequences, the HSPD21 PRT, initially designed to measure the beta-defensin copy number variation in humans, proved to be a versatile methodology with a vast applicability to different primate species. Moreover, the PRT provides a more detailed analysis of this locus when compared with array-CGH, which relies on the available genome assembly to construct the probes.

HSPD21 showed that the number of copies of β-defensin genes at the chromosome 8 locus in chimpanzees varied between 3 and 4 copies. Our results are also supported by evidence from array-CGH studies showing CNV of the beta-defensin orthologous region in chimpanzee (Perry *et al.* 2008; Marques-Bonet *et al.* 2009). Moreover, copy numbers of 4, 5 and 6 for this locus were also recently reported for chimpanzee by PRT-based methods (Hardwick *et al.* 2011). Despite the differences in the range of copy numbers observed, which may reflect the small size cohorts, both studies report variation in β-defensin copy number. Therefore, this and other studies support the evidence of beta-defensin CNV in chimpanzee in a similar range to that observed in humans. Moreover, these findings contrast with the chimpanzee genome assembly (CGSC 2.1.3/panTro3), where only one copy is annotated at this locus, which is not supported by this work. It would be of great advantage for future investigations of CNV in primates if the chimpanzee genome assembly could be updated to reflect recent findings.

For the bonobo individual, a β-defensin copy number of one was typed with HSPD21 PRT. This species is closely related to chimpanzee, with a common ancestor at around 1 million years ago, therefore high sequence similarity was expected to allow amplification of test and reference locus using HSPD21 PRT. However, the copy number given by PRT, one copy, could not be confirmed using only one individual and without any sequence data

available from the test locus for this species. Even if a copy number of one is described by PRT it is still possible that the bonobo could have a higher copy number, which may not be revealed due to mismatches that prohibit the amplification of all β-defensin copies. Despite these facts, due to the close evolutionary relationship to chimpanzee and the evidence of HSPD3 locus on the bonobo genome shown by sequencing data, it is possible to suggest the presence of only one copy for this individual with the evidence given by PRT.

The investigation of β-defensin CNV in orangutan was inconclusive. Through sequence comparison, no orthologous region of the HSPD21 PRT test locus was found, suggesting the absence of the HSPD3 pseudogene in orangutan. To address the beta-defensin copy number in orangutan, further work would be necessary to design an orangutan-specific PRT assay. However, with the limitations of the existing orangutan genome assembly (WUGSC 2.0.2/ponAbe2), with no annotated β-defensin genes, this would represent a difficult and laborious task that could not be accomplished due to time restrictions.

The β-defensin genes in all gorillas typed with HSPD21 PRT, were found to have two copies. Allele-specific sequencing was carried out for sequence variants found at the *DEFB103* locus to confirm the results given by HSPD21 PRT. However, this approach proved not to be very successful or useful in clarifying the copy number. As this assay relies in the presence of multiple sequence variants in the same individual in order to follow the segregation of alleles into different haplotypes, from our results only two gorillas, Sylvia and J79 (from the 5 gorillas) presented multiple sequence variants. In Sylvia, the two sequence variants, 3 and 4, were located too far apart (around 850 bp) to be targeted by the same sequencing assay. On the other hand, the locations of the sequence variants 1, 2 and 3 (in J79) allowed the design of an allele-specific sequencing assay. Unfortunately, the sequencing assay did not successfully clarify the copy number in J79. Finally, from the allele-specific sequencing assay designed for the only sequence variant found in Sylvia and Tomoka, it was possible to confirm the existence of only two alleles and therefore supports the copy number of 2. Further work, either by designing other allele-specific sequencing assays or by using restriction enzymes, could be developed in order to confirm more confidently the copy number given by HSPD21 PRT.

Despite the fact that no further confirmation of a copy number of two was achieved with the allelic-specific sequencing assays, array-CGH data recently reported for gorilla genome (Marques-Bonet *et al.* 2009) agrees with the copy number reported here by PRT.

In summary, the beta-defensins are copy number variable in chimpanzee and bonobo, but in gorilla, evidence suggests the presence of only two copies. The chimpanzee showed a similar range of variation to that observed in humans, suggesting that beta-defensins have a similar role in immunity in humans and chimpanzees. The study of beta-defensin copy number variation in gorilla suggests for the presence of only two copies in gorilla, implying that the variation in copy number of beta-defensin only originated in the chimpanzee-human lineage, so occurring after the divergence from gorilla in evolutionary history (Figure 53). The possible phenotypic consequences of the lack of variation in copy number of the beta-defensin genes in gorilla are unknown. Nonetheless, it may suggest that gorilla and chimpanzee-human lineages have been subjected to different evolutionary pressures forcing the immune system to evolve in different directions.

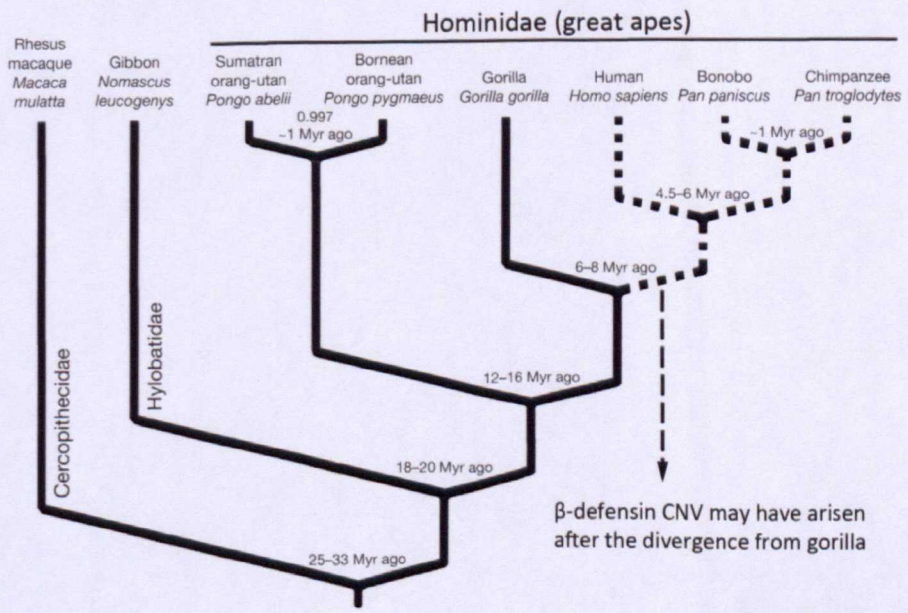


Figure 53: Old World primates (Catarrhine) phylogeny showing the divergence among great apes, a small ape (Hylobatidae) and an Old World monkey (Cercopithecidae) with respect to humans (adapted from Locke et al. (2011)).

The studies of primate CNVs provide a new perspective on the divergence and selective pressures acting on these genomic regions. Moreover, this highlights the significance of CNV in evolution and their contribution for interspecies and phenotypic variation, which consequently can account for disease susceptibility. The knowledge of CNV in other primates is still limited by the quality and coverage of the existing assemblies, which restricts the clarity of the evolution of genomic variation and its importance in the primate lineage. For a good evaluation of CNV in non-human primate genomes, studies need to be scaled up and the use of species-specific techniques needs to be adopted to access the copy number of functional and evolutionary important CNV regions, such as the β-defensin genes.

CHAPTER 7: FINAL DISCUSSION AND CONCLUSIONS

7.1 MEASURING THE β-DEFENSIN CNV

In this study a new copy number measurement method, the Triplex assay, was developed to measure the β-defensin copy number variation. The β-defensins are located on a multiallelic copy number locus at 8p23.1, which shows extensive copy number variation, with individuals showing between 2 and 12 copies per diploid genome.

The measurement of copy number variation has been increasingly reported in several studies by different methods, including real-time PCR and array-CGH based methods. These methods have been extensively used to study genome-wide copy number changes in human genomes (Iafrate *et al.* 2004; Sebat *et al.* 2004; Sharp *et al.* 2005; Redon *et al.* 2006; Wong *et al.* 2007; McCarroll *et al.* 2008). Although array-CGH methods have a good coverage of the genome, they provide very poor coverage of complex regions containing segmental duplications and copy number variation (Carter 2007). Moreover, none of these methodologies combines a straightforward method with accuracy, low cost and general applicability.

PRT-based methods, in particular when combined in triplex, proved to be a simple, inexpensive and high-throughput method to genotype loci such as this, with copy numbers as high as 12. The evidence of the accuracy of this method is shown by its capacity to classify samples into discrete integer copy number classes. Moreover, the triplex gives the advantage of providing three independent copy number measurements for each sample tested, which in the analyses were generally concordant. Therefore, our data allowed us to conclude that the Triplex assay is an accurate method to measure the β-defensin copy number and shows considerably increased power in measuring the multiallelic copy number variant of the β-defensin locus when compared with the methods previously available, in particular with real-time PCR (Aldhous *et al.* 2010; Fode *et al.* 2011). The Triplex assay was directly compared with real-time PCR in two different studies to measure β-defensin copy number variation in different human populations and in a Crohn's disease association study (Aldhous *et al.* 2010; Fode *et al.* 2011). Compared with the clustered distribution of copy number values given by the Triplex, the copy number values obtained from real-time PCR showed a continuous distribution with no clear separation between different copy number classes. Such results highlight the importance of assessing copy number accurately, which otherwise could lead to false positive associations, such as the case of Crohn's disease, when using real-time PCR (Fellermann *et al.* 2006; Bentley *et al.* 2010).

The possibility of using SNPs in linkage disequilibrium to predict the copy number has been explored in several studies (Hinds *et al.* 2006; McCarroll *et al.* 2006; Redon *et al.* 2006; McCarroll *et al.* 2008). Strong linkage disequilibrium between SNPs and CNV regions is most frequently found for biallelic CNVs, where SNPs can effectively tag an evolutionary lineage (Hinds *et al.* 2006; McCarroll *et al.* 2006; Redon *et al.* 2006). Here, we investigated the regions neighbouring the multiallelic β-defensin copy number locus with the aim of finding possible SNPs in LD with the copy number. However, our results indicate that no SNPs tag the β-defensin copy number variation at 8p23.1, agreeing with previous studies describing the poor level of LD around CNVs, especially at multiallelic regions (Redon *et al.* 2006; McCarroll *et al.* 2008). This suggests that the β-defensin copy number is changing faster than

other genomic properties, such as SNPs. The high rate of recurrent recombination previously reported for the β-defensin locus which contributes for their high genomic instability, supports the poor correlation of CN with neighbouring SNPs found here (Abu Bakar *et al.* 2009). Therefore, it is very unlikely that multiallelic CNVs with recurrent rearrangements will show correlation with flanking makers.

The development of new CNV methods is essential to investigate the biological consequences of CNV in susceptibility to disease, human variation and human evolution. Considering the number of copy variable loci in the human genome and the number of these that are candidate loci for susceptibility to several diseases, this work shows that PRT, especially when used in Triplex, can be a powerful technique to determine copy numbers in large case-control association studies.

7.2 β-DEFENSIN CNV IN HUMAN DIVERSITY, DISEASE AND EVOLUTION

In this study PRT-based assays (single and triplex assay) were used to measure the β-defensin copy number variation in a psoriasis case-control association study. From the analysis of β-defensin copy number variation in patients with psoriasis, a new susceptibility locus was identified. A significant association (p-value=9.44×10^{-4}) between psoriasis and β-defensin copy number variation was reported in a total of 516 samples (202 psoriatic cases) by two independent studies, with high copy number showing a higher susceptibility to psoriasis. Considering the significance standards currently defined for candidate disease-loci at genome-wide studies, the p-value obtained here is only of modest significance. The genome-wide level of significance for studies of SNP markers has been determined to be a nominal p-value of around 5×10^{-8} (Dudbridge and Gusnanto 2008). Since the prior probability of any given locus to be associated with a complex disease is very low and the probability of a type I error is much higher than the nominal level of the test, this explains the fact that the nominal level of significance defined at the genome-wide studies is much higher. The p-value for the association of beta-defensin with copy number does not reach this threshold; however it is not clear whether this threshold is too conservative for CNV testing or what threshold would constitute an appropriate significant value for this study, which involves the typing of a single candidate locus. The results show however, a significant but weak association between beta-defensin copy number and psoriasis.

The work presented here shows the importance of β-defensin in susceptibility to psoriasis, as well as the role of structural variants, as opposed to SNPs, in susceptibility to disease. Moreover, this study provides a good example in defining new genetic factors responsible for the observed clinical phenotypes, contributing to a better understanding of the pathogenesis and susceptibility to psoriasis disease. As such, it seems clear that the importance of copy number variation in disease status is still under-investigated and further work should consider CNV as a possible source for disease susceptibility.

Association studies need to rely on accurate and robust techniques to precisely measure copy number variation and report relationships between genetic variants and disease phenotypes. As demonstrated in the past, false associations have been reported using real-time PCR (Aldhous *et al.* 2010), proving that careful analysis of copy number is essential in association studies. Here, the Triplex assay was demonstrated to perform accurately in large case-control association studies and proved to be essential for studies such as this.

The variation in copy number of the β-defensins was further analysed among six human populations to investigate the inter-population copy number diversity. This study did not show any evidence of significant differentiation associated with geographical origin. These findings suggest that the observed mean copy number of 4, in all populations, reflects balancing or intermediate selection, suggesting that high copy number may increase susceptibility to inflammatory or autoimmune diseases, such as psoriasis, and low copy number could contribute to the risk of infectious diseases. This hypothesis fits the multifunctional role of this gene family as "natural antibiotics" and signalling molecules with important roles in several pathways in the innate and adaptive immune system.

Finally, with the aim to track down the origin of beta-defensin copy number variation in the evolutionary history of primates, the copy number in great apes was analysed. The beta-defensins were found to be copy variable in human and chimpanzee but not in gorilla; the data suggest the presence of only two copies at this locus. This indicates that copy number variation of beta-defensins might have originated by duplication in the human and chimpanzee common ancestor after the divergence from gorilla about 6-8 million years ago.

CHAPTER 8: BIBLIOGRAPHY

Aarbiou, J., R. M. Verhoosel, S. van Wetering, W. I. de Boer, J. H. J. M. van Krieken, S. V. Litvinov, K. F. Rabe and P. S. Hiemstra (2004). "Neutrophil Defensins Enhance Lung Epithelial Wound Closure and Mucin Gene Expression In Vitro." American Journal of Respiratory Cell and Molecular Biology **30**(2): 193-201.

Abu Bakar, S. (2010). "Generation of diversity at the human beta-defensin copy number." PhD Thesis. The University of Nottingham, Nottingham, UK.

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

Agúndez, J. A. G., L. Gallardo, M. C. Ledesma, L. Lozano, A. Rodriguez-Lescure, J. C. Pontes, M. C. Iglesias-Moreno, J. Poch, J. M. Ladero and J. Benítez (2001). "Functionally Active Duplications of the *CYP2D6* Gene Are More Prevalent among Larynx and Lung Cancer Patients." Oncology **61**(1): 59-63.

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

Akrami, S. M., J. S. Rowland, G. R. Taylor and J. A. L. Armour (2003). "Diagnosis of gene dosage alterations at the PMP22 gene using MAPH." Journal of Medical Genetics **40**(11): e123.

Akrami, S. M., R. M. Winter, J. D. Brook and J. A. Armour (2001). "Detection of a large TBX5 deletion in a family with Holt-Oram syndrome." Journal of Medical Genetics **38**(12): e44.

Aldhous, M. C., S. Abu Bakar, N. J. Prescott, R. Palla, K. Soo, J. C. Mansfield, C. G. Mathew, J. Satsangi and J. A. Armour (2010). "Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease." Human Molecular Genetics **19**(24): 4930-4938.

Aldred, P., E. Hollox and J. Armour (2005). "Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3." Human Molecular Genetics **14**(14): 2045 - 2052.

Ali, R. S., A. Falconer, M. Ikram, C. E. Bissett, R. Cerio and A. G. Quinn (2001). "Expression of the peptide antibiotics human beta defensin-1 and human beta defensin-2 in normal human skin." Journal of Investigative Dermatology **117**(1): 106-111.

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

Antonacci, F., J. M. Kidd, T. Marques-Bonet, M. Ventura, P. Siswara, Z. Jiang and E. E. Eichler (2009). "Characterization of six human disease-associated inversion polymorphisms." Human Molecular Genetics **18**(14): 2555-2566.

Armengol, L., M. A. Pujana, J. Cheung, S. W. Scherer and X. Estivill (2003). "Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements." Human Molecular Genetics **12**(17): 2201-2208.

Armour, J., C. Sismani, P. Patsalis and G. Cross (2000). "Measurement of locus copy number by hybridisation with amplifiable probes." Nucleic Acids Research **28**(2): 605 - 609.

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

Armour, J. A. L., R. Palla, P. L. J. M. Zeeuwen, M. den Heijer, J. Schalkwijk and E. J. Hollox (2007). "Accurate, high-throughput typing of copy number

variation using paralogue ratios from dispersed repeats." Nucleic Acids Research **35**(3): e19.

Asumalahti, K., T. Laitinen, R. Itkonen-Vatjus, M. L. Lokki, S. Suomela, E. Snellman, U. Saarialho-Kere and J. Kere (2000). "A candidate gene for psoriasis near HLA-C, HCR (Pg8), is highly polymorphic with a disease-associated susceptibility allele." Human Molecular Genetics **9**(10): 1533-1542.

Asumalahti, K., T. Laitinen, P. Lahermo, S. Suomela, R. Itkonen-Vatjus, C. Jansen, J. Karvonen, S.-L. Karvonen, T. Reunala, E. Snellman, T. Uurasmaa, U. Saarialho-Kere and J. Kere (2003). "Psoriasis Susceptibility Locus on 18p Revealed by Genome Scan in Finnish Families Not Associated with PSORS1." Journal of Investigative Dermatology **121**(4): 735-740.

Bacolla, A. and R. D. Wells (2004). "Non-B DNA Conformations, Genomic Rearrangements, and Human Disease." Journal of Biological Chemistry **279**(46): 47411-47414.

Bailey, J., R. Baertsch, W. Kent, D. Haussler and E. Eichler (2004). "Hotspots of mammalian chromosomal evolution." Genome Biology **5**(4): R23.

Bailey, J. A. and E. E. Eichler (2006). "Primate segmental duplications: crucibles of evolution, diversity and disease." Nature Reviews Genetics **7**(7): 552-564.

Bailey, J. A., Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li and E. E. Eichler (2002). "Recent Segmental Duplications in the Human Genome." Science **297**(5583): 1003-1007.

Bailey, J. A., A. M. Yavor, H. F. Massa, B. J. Trask and E. E. Eichler (2001). "Segmental Duplications: Organization and Impact Within the Current Human Genome Project Assembly." Genome Research **11**(6): 1005-1017.

Ballana, E., J. Gonzalez, N. Bosch and X. Estivill (2007). "Inter-population variability of DEFA3 gene absence: correlation with haplotype structure and population variability." BMC Genomics **8**(1): 14.

Barber, J. C., C. A. Joyce, M. N. Collinson, J. C. Nicholson, L. R. Willatt, H. M. Dyson, M. S. Bateman, A. J. Green, J. R. Yates and N. R. Dennis (1998). "Duplication of 8p23.1: a cytogenetic anomaly with no established clinical significance." Journal of Medical Genetics **35**(6): 491-496.

Barber, J. C., C. J. Reed, S. P. Dahoun and C. A. Joyce (1999). "Amplification of a pseudogene cassette underlies euchromatic variation of 16p at the cytogenetic level." Human Genetics **104**(3): 211-218.

Barber, J. C. K. (2005). "Directly transmitted unbalanced chromosome abnormalities and euchromatic variants." Journal of Medical Genetics **42**(8): 609-629.

Barber, J. C. K., V. Maloney, E. J. Hollox, A. Stuke-Sontheimer, G. du Bois, E. Daumiller, U. Klein-Vogler, A. Dufke, J. A. L. Armour and T. Liehr (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.

Barber, J. C. K., V. K. Maloney, S. Huang, D. J. Bunyan, L. Cresswell, E. Kinning, A. Benson, T. Cheetham, J. Wyllie, S. A. Lynch, S. Zwolinski, L. Prescott, Y. Crow, R. Morgan and E. Hobson (2007). "8p23.1 duplication syndrome; a novel genomic condition with unexpected complexity revealed by array CGH." European Journal of Human Genetics **16**(1): 18-27.

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

Bauman, J. G. J., J. Wiegant, P. Borst and P. van Duijn (1980). "A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA." Experimental Cell Research **128**(2): 485-490.

Bennett, S. T., A. M. Lucassen, S. C. L. Gough, E. E. Powell, D. E. Undlien, L. E. Pritchard, M. E. Merriman, Y. Kawaguchi, M. J. Dronsfield, F. Pociot, J. Nerup, N. Bouzekri, A. Cambonthomsen, K. S. Ronningen, A. H. Barnett, S. C. Bain and J. A. Todd (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., M. Raida, H.-J. Mägert, P. Schulz-Knappe and W.-G. Forssmann (1995). "hBD-1: a novel β-defensin from human plasma." Federation of European Biochemical Societies Letters **368**(2): 331-335.

Bentley, R. W., J. Pearson, R. B. Gearry, M. L. Barclay, C. McKinney, T. R. Merriman and R. L. Roberts (2010). "Association of higher DEFB4 genomic copy number with Crohn's disease." American Journal of Gastroenterology **105**(2): 354-359.

Berend, S. A., S. Canun, C. McCaskill, S. L. Page and L. G. Shaffer (1998). "Molecular analysis of mosaicism for two different de novo acrocentric rearrangements demonstrates diversity in robertsonian translocation formation." American Journal of Medical Genetics **80**(3): 252-259.

Beutler, E., M. Yeh and V. F. Fairbanks (1962). "The Normal Human Female as mosaic of X-chromosome activity: studies uisng the gene for G-6-PD-deficiency as a marker " Proceedings of the National Academy of Sciences, USA **48**(1): 9-16.

Bhalerao, J. and A. M. Bowcock (1998). "The genetics of psoriasis: a complex disorder of the skin and immune system." Human Molecular Genetics **7**(10): 1537-1545.

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

Bowcock, A. M. (2005). "The genetics of psoriasis and autoimmunity." Annual Reviews of Genomics and Human Genetics **6**: 93-122.

Bowcock, A. M. and J. G. Krueger (2005). "Getting under the skin: the immunogenetics of psoriasis." Nature Reviews Immunology **5**(9): 699-711.

Brandrup, F., M. Hauge, K. Henningsen and B. Eriksen (1978). "Psoriasis in an unselected series of twins." Archives of Dermatology **114**(6): 874-878.

Buiting, K., S. Saitoh, S. Gross, B. Bittrich, S. Schwartz, R. D. Nicholls and B. Horsthemke (1995). "Inherited microdeletions in the Angleman and Prader-Willi syndromes define an imprinting centre on human chromosome 15." Nature Genetics **9**(4): 395-400.

Campbell, C D., N. Sampas, A. Tsalenko, P H. Sudmant, J M. Kidd, M. Malig, T H. Vu, L. Vives, P. Tsang, L. Bruhn and E E. Eichler (2011). "Population-Genetic Properties of Differentiated Human Copy-Number Polymorphisms." American Journal of Human Genetics **88**(3): 317-332.

Candille, S. I., C. B. Kaelin, B. M. Cattanach, B. Yu, D. A. Thompson, M. A. Nix, J. A. Kerns, S. M. Schmutz, G. L. Millhauser and G. S. Barsh (2007). "A β-Defensin Mutation Causes Black Coat Color in Domestic Dogs." Science **318**(5855): 1418-1423.

Capon, F., M. J. Bijlmakers, N. Wolf, M. Quaranta, U. Huffmeier, M. Allen, K. Timms, V. Abkevich, A. Gutin, R. Smith, R. B. Warren, H. S. Young, J. Worthington, A. D. Burden, C. E. M. Griffiths, A. Hayday, F. O. Nestle, A. Reis, J. Lanchbury, J. N. Barker and R. C. Trembath (2008). "Identification of ZNF313/RNF114 as a novel psoriasis susceptibility gene." Human Molecular Genetics **17**(13): 1938-1945.

Capon, F., P. Di Meglio, J. Szaub, N. J. Prescott, C. Dunster, L. Baumber, K. Timms, A. Gutin, V. Abkevic, A. D. Burden, J. Lanchbury, J. N. Barker, R. C. Trembath and F. O. Nestle (2007). "Sequence variants in the genes for the interleukin-23 receptor (IL23R) and its ligand (IL12B) confer protection against psoriasis." Human Genetics **122**(2): 201-206.

Capon, F., C. Helms, C. D. Veal, D. Tillman, A. D. Burden, J. N. Barker, A. M. Bowcock and R. C. Trembath (2004). "Genetic analysis of PSORS2 markers in a UK dataset supports the association between RAPTOR SNPs and familial psoriasis." Journal of Medical Genetics **41**(6): 459-460.

Capon, F., G. Novelli, S. Semprini, M. Clementi, M. Nudo, P. Vultaggio, C. Mazzanti, T. Gobello, A. Botta, G. Fabrizi and B. Dallapiccola (1999). "Searching for psoriasis susceptibility genes in Italy: genome scan and evidence for a new locus on chromosome 1." Journal of Investigative Dermatology **112**(1): 32-35.

Cargill, M., S. J. Schrodi, M. Chang, V. E. Garcia, R. Brandon, K. P. Callis, N. Matsunami, K. G. Ardlie, D. Civello, J. J. Catanese, D. U. Leong, J. M. Panko, L. B. McAllister, C. B. Hansen, J. Papenfuss, S. M. Prescott, T. J. White, M. F. Leppert, G. G. Krueger and A. B. Begovich (2007). "A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes." American Journal of Human Genetics **80**(2): 273-290.

Carsten Münk, G. W., Otto O. Yang, Alan J. Waring, Wei Wang, Teresa Hong, Robert I. Lehrer, Nathaniel R. Landau, Alexander M. Cole. (2004). "The θ-Defensin, Retrocyclin, Inhibits HIV-1 Entry." AIDS Research and Human Retroviruses **19**(10).

Carter, N. P. (2007). "Methods and strategies for analyzing copy number variation using DNA microarrays." Nature Genetics.

Caspersson, T., L. Zech and C. Johansson (1999). "Analysis of Human Metaphase Chromosome Set by Aid of DNA-Binding Fluorescent Agents." Experimental Cell Research **253**(2): 302-304.

Chaly, Y. V., E. M. Paleolog, T. S. Kolesnikova, Tikhonov, II, E. V. Petratchenko and N. N. Voitenok (2000). "Neutrophil alpha-defensin human neutrophil peptide modulates cytokine production in human monocytes and adhesion molecule expression in endothelial cells." European Cytokine Network **11**(2): 257-266.

Chang, T. L., J. Vargas, A. DelPortillo and M. E. Klotman (2005). "Dual role of α-defensin-1 in anti–HIV-1 innate immunity." The Journal of Clinical Investigation **115**(3): 765-773.

Chartier-Harlin, M.-C., J. Kachergus, C. Roumier, V. Mouroux, X. Douay, S. Lincoln, C. Levecque, L. Larvor, J. Andrieux, M. Hulihan, N. Waucquier, L. Defebvre, P. Amouyel, M. Farrer and A. Destée (2004). "Alpha-synuclein locus duplication as a cause of familial Parkinson's disease." The Lancet **364**(9440): 1167-1169.

Chavakis, T., D. B. Cines, J.-S. Rhee, O. D. Liang, U. Schubert, H.-P. Hammes, A. A.-R. Higazi, P. P. Nawroth, K. T. Preissner and K. Bdeir (2004). "Regulation of neovascularization by human neutrophil peptides (α-defensins): a link between inflammation and angiogenesis." The Journal of the Federation of American Societies for Experimental Biology **18**(11): 1306-1308.

Chen, G. K., E. Slaten, R. A. Ophoff and K. Lange (2006a). "Accommodating chromosome inversions in linkage analysis." American Journal of Human Genetics **79**(2): 238-251.

Chen, Q. X., M. Book, X. M. Fang, A. Hoeft and F. Stuber (2006b). "Screening of copy number polymorphisms in human beta-defensin genes using modified real-time quantitative PCR." Journal of Immunological Methods **308**(1-2): 231-240.

Chertov, O., D. F. Michiel, L. Xu, J. M. Wang, K. Tani, W. J. Murphy, D. L. Longo, D. D. Taub and J. J. Oppenheim (1996). "Identification of Defensin-1, Defensin-2, and CAP37/Azurocidin as T-cell Chemoattractant Proteins Released from Interleukin-8-stimulated Neutrophils." Journal of Biological Chemistry **271**(6): 2935-2940.

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

Clayton, D. G., N. M. Walker, D. J. Smyth, R. Pask, J. D. Cooper, L. M. Maier, L. J. Smink, A. C. Lam, N. R. Ovington, H. E. Stevens, S. Nutland, J. M. M. Howson, M. Faham, M. Moorhead, H. B. Jones, M. Falkowski, P. Hardenbol, T. D. Willis and J. A. Todd (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., T. Ganz, A. M. Liese, M. D. Burdick, L. Liu and R. M. Strieter (2001). "Cutting Edge: IFN-Inducible ELR- CXC Chemokines Display Defensin-Like Antimicrobial Activity." The Journal of Immunology **167**(2): 623-627.

Cole, A. M., T. Hong, L. M. Boo, T. Nguyen, C. Zhao, G. Bristol, J. A. Zack, A. J. Waring, O. O. Yang and R. I. Lehrer (2002). "Retrocyclin: a primate peptide that protects cells from infection by T- and M-tropic strains of HIV-1." Proceedings of the National Academy of Sciences of the USA **99**(4): 1813-1818.

Conejo García, J.-R., A. Krause, S. Schulz, F.-J. Rodríguez-Jiménez, E. Klüver, K. Adermann, U. Forssmann, A. Frimpong-Boateng, R. Bals and W.-G. Forssmann (2001). "Human β-defensin 4: a novel inducible peptide with a specific salt-sensitive spectrum of antimicrobial activity." The Journal of the Federation of American Societies for Experimetal Biology.

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

Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer and M. E. Hurles (2010). "Origins and functional

impact of copy number variation in the human genome." Nature **464**(7289): 704-712.

Coussens, L. M. and Z. Werb (2002). "Inflammation and cancer." Nature **420**(6917): 860-867.

Crovella, S., N. Antcheva, I. Zelezetsky, M. Boniotto, S. Pacor, M. V. V. Falzacappa and A. Tossi (2005). "Primate beta-defensins - Structure, Function and Evolution." Current Protein and Peptide Science **6**: 7-21.

Cummings, J. R., T. Ahmad, A. Geremia, J. Beckly, R. Cooney, L. Hancock, S. Pathan, C. Guo, L. R. Cardon and D. P. Jewell (2007). "Contribution of the novel inflammatory bowel disease gene IL23R to disease susceptibility and phenotype." Inflammatory Bowel Diseases **13**(9): 1063-1068.

Cutler, G., L. A. Marshall, N. Chin, H. Baribault and P. D. Kassner (2007). "Significant gene content variation characterizes the genomes of inbred mouse strains." Genome Research **17**(12): 1743-1754.

Dall'Ozzo, S., C. Andres, P. Bardos, H. Watier and G. Thibault (2003). "Rapid single-step FCGR3A genotyping based on SYBR Green I fluorescence in real-time multiplex allele-specific PCR." Journal of Immunological Methods **277**(1-2): 185-192.

Davidson, A. and B. Diamond (2001). "Autoimmune diseases." The New England Journal of Medicine **345**(5): 340-350.

de Cid, R., E. Riveira-Munoz, P. L. Zeeuwen, J. Robarge, W. Liao, E. N. Dannhauser, E. Giardina, P. E. Stuart, R. Nair, C. Helms, G. Escaramis, E. Ballana, G. Martin-Ezquerra, M. den Heijer, M. Kamsteeg, I. Joosten, E. E. Eichler, C. Lazaro, R. M. Pujol, L. Armengol, G. Abecasis, J. T. Elder, G. Novelli, J. A. Armour, P. Y. Kwok, A. Bowcock, J. Schalkwijk and X. Estivill (2009). "Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis." Nature Genetics **41**(2): 211-215.

De Marzo, A. M., Elizabeth A. Platz , Siobhan Sutcliffe , Jianfeng Xu , Henrik Grönberg , Charles G. Drake , Yasutomo Nakai , William B. Isaacs & William G. Nelson (2007). "Inflammation in prostate carcinogenesis." Nature Reviews Cancer **7**(4): 256-269.

de Smith, A. J., R. G. Walters, P. Froguel and A. I. Blakemore (2008). "Human genes involved in copy number variation: mechanisms of origin, functional effects and implications for disease." Cytogenetic and Genome Research **123**(1-4): 17-26.

Deutsch, S., U. Choudhury, G. Merla, C. Howald, A. Sylvan and S. E. Antonarakis (2004). "Detection of aneuploidies by paralogous sequence quantification." Journal of Medical Genetics **41**(12): 908-915.

Devriendt, K., K. De Mars, P. De Cock, M. Gewillig and J. P. Fryns (1995). "Terminal deletion in chromosome region 8p23.1-8pter in a child with features of velo-cardio-facial syndrome." Annals of Human Genetics **38**(4): 228-230.

Dhami, P., A. J. Coffey, S. Abbs, J. R. Vermeesch, J. P. Dumanski, K. J. Woodward, R. M. Andrews, C. Langford and D. Vetrie (2005). "Exon array CGH: Detection of copy-number changes at the resolution of individual exons in the human genome." American Journal of Human Genetics **76**(5): 750-762.

Diamond, G., M. Zasloff, H. Eck, M. Brasseur, W. L. Maloy and C. L. Bevins (1991). "Tracheal antimicrobial peptide, a cysteine-rich peptide from mammalian tracheal mucosa: peptide isolation and cloning of a cDNA." Proceedings of the National Academy of Sciences, USA **88**(9): 3952-3956.

Donald, C. D., C. Q. Sun, S. D. Lim, J. Macoska, C. Cohen, M. B. Amin, A. N. Young, T. A. Ganz, F. F. Marshall and J. A. Petros (2003). "Cancer-Specific Loss of β-Defensin 1 in Renal and Prostatic Carcinomas." Laboratory Investigation **83**(4): 501-505.

Dopman, E. B. and D. L. Hartl (2007). "A portrait of copy-number polymorphism in Drosophila melanogaster." Proceedings of the National Academy of Sciences, USA **104**(50): 19920-19925.

Down, J. (1866). "Observations on an Ethnic Classification of Idiots " Clinical Lecture Reports **3**: 259–262.

Dudbridge, F. and A. Gusnanto (2008). "Estimation of significance thresholds for genomewide association scans." Genetic Epidemiology **32**(3): 227-234.

Duerr, R. H., K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, A. H. Steinhart, C. Abraham, M. Regueiro, A. Griffiths, T. Dassopoulos, A. Bitton, H. Yang, S. Targan, L. W. Datta, E. O. Kistner, L. P. Schumm, A. T. Lee, P. K. Gregersen, M. M. Barmada, J. I. Rotter, D. L. Nicolae and J. H. Cho (2006). "A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene." Science **314**(5804): 1461-1463.

Duffin, K. C., J. Woodcock and G. G. Krueger (2010). "Genetic variations associated with psoriasis and psoriatic arthritis found by genome-wide association." Dermatologic Therapy **23**(2): 101-113.

Duffy, D. L., L. S. Spelman and N. G. Martin (1993). "Psoriasis in Australian twins." Journal of the American Academy of Dermatology **29**(3): 428-434.

Duvic, M. (1990). "Immunology of AIDS Related to Psoriasis." Journal of Investigative Dermatololgy **95**(s5): 38S-40S.

Edwards, J. H., D. G. Harnden, A. H. Cameron, V. M. Crosse and O. H. Wolff (1960). "A new trisomic syndrome." Lancet **1**(7128): 787-790.

Egan, C. M., S. Sridhar, M. Wigler and I. M. Hall (2007). "Recurrent DNA copy number variation in the laboratory mouse." Nature Genetics **39**(11): 1384-1389.

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

Ellinghaus, E., D. Ellinghaus, P. E. Stuart, R. P. Nair, S. Debrus, J. V. Raelson, M. Belouchi, H. Fournier, C. Reinhard, J. Ding, Y. Li, T. Tejasvi, J. Gudjonsson, S. W. Stoll, J. J. Voorhees, S. Lambert, S. Weidinger, B. Eberlein, M. Kunz, P. Rahman, D. D. Gladman, C. Gieger, H. E. Wichmann, T. H. Karlsen, G. Mayr, M. Albrecht, D. Kabelitz, U. Mrowietz, G. R. Abecasis, J. T. Elder, S. Schreiber, M. Weichenthal and A. Franke (2010). "Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2." Nature Genetics **42**(11): 991-995.

Emerson, J. J., M. Cardoso-Moreira, J. O. Borevitz and M. Long (2008). "Natural Selection Shapes Genome-Wide Patterns of Copy-Number Polymorphism in Drosophila melanogaster." Science **320**(5883): 1629-1631.

Enlund, F., L. Samuelsson, C. Enerback, A. Inerot, J. Wahlstrom, M. Yhr, A. Torinsson, J. Riley, G. Swanbeck and T. Martinsson (1999). "Psoriasis susceptibility locus in chromosome region 3q21 identified in patients from southwest Sweden." European Journal of Human Genetics **7**(7): 783-790.

Evans, J. A., N. Canning, A. G. W. Hunter, J. T. Martsolf, M. Ray, D. R. Thompson and J. L. Hamerton (1978). "A cytogenetic survey of 14,069 newborn infants. III. An analysis of the significance and cytologic behavior of the Robertsonian and reciprocal translocations." Cytogenetics and Cell Genetics **20**(1-6): 96-123.

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

Farber, E. M., M. L. Nall and W. Watson (1974). "Natural History of Psoriasis in 61 Twin Pairs." Archives of Dermatology **109**(2): 207-211.

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

Fernando, M. M. A., L. Boteva, D. L. Morris, B. Zhou, Y. L. Wu, M.-L. Lokki, C. Y. Yu, J. D. Rioux, E. J. Hollox and T. J. Vyse (2010). "Assessment of

complement C4 gene copy number using the paralog ratio test." Human Mutation **31**(7): 866-874.

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

Feuk, L., J. R. MacDonald, T. Tang, A. R. Carson, M. Li, G. Rao, R. Khaja and S. W. Scherer (2005). "Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies." PLoS Genetics **1**(4): e56.

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

Florijn, R. J., A. J. Bonden, H. Vrolijk, J. Wiegant, J.-W. Vaandrager, F. Bass, J. T. den Dunnen, H. J. Tanke, G.-J. B. van Ommen and A. K. Raap (1995). "High-resolution DNA Fiber-FISH for genomic DNA mapping and colour bar-coding of large genes." Human Molecular Genetics **4**(5): 831-836.

Fode, P., C. Jespersgaard, R. J. Hardwick, H. Bogle, M. Theisen, D. Dodoo, M. Lenicek, L. Vitek, A. Vieira, J. Freitas, P. S. Andersen and E. J. Hollox (2011). "Determination of Beta-Defensin Genomic Copy Number in Different Populations: A Comparison of Three Methods." PLoS ONE **6**(2): e16768.

Foerster, J., I. Nolte, S. Schweiger, C. Ehlert, M. Bruinenberg, K. Spaar, G. van der Steege, M. Mulder, V. Kalscheuer, B. Moser, Z. Kijas, P. Seeman, M. Stander, W. Sterry and G. te Meerman (2004). "Evaluation of the IRF-2 Gene as a Candidate for PSORS3." Journal of Investigative Dermatology **122**(1): 61-64.

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

Fulton, C., G. M. Anderson, M. Zasloff, R. Bull and A. G. Quinn (1997). "Expression of natural peptide antibiotics in human skin." The Lancet **350**(9093): 1750-1751.

Gabay, J. (1994). "Ubiquitous natural antibiotics." Science **264**(5157): 373-374.

Ganz, T. (1999a). "Defensins and host defense." Science **286**(5439): 420-421.

Ganz, T. (1999b). "Immunology - Defensins and host defense." Science **286**(5439): 420-421.

Ganz, T. (2003). "Defensins: antimicrobial peptides of innate immunity." Nature Reviews Immunology **3**(9): 710-720.

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

Ganz, T. and R. I. Lehrer (1998). "Antimicrobial peptides of vertebrates." Current Opinion in Immunology **10**(1): 41-44.

Ganz, T., M. E. Selsted, D. Szklarek, S. S. Harwig, K. Daher, D. F. Bainton and R. I. Lehrer (1985). "Defensins. Natural peptide antibiotics of human neutrophils." The Journal of Clinical Investigation **76**(4): 1427-1435.

García-Closas, M., N. Malats, D. Silverman, M. Dosemeci, M. Kogevinas, D. W. Hein, A. Tardón, C. Serra, A. Carrato, R. García-Closas, J. Lloreta, G. Castaño-Vinyals, M. Yeager, R. Welch, S. Chanock, N. Chatterjee, S. Wacholder, C. Samanic, M. Torà, F. Fernández, F. X. Real and N. Rothman (2005). "NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses." The Lancet **366**(9486): 649-659.

Garcia, J. R., F. Jaumann, S. Schulz, A. Krause, J. Rodriguez-Jimenez, U. Forssmann, K. Adermann, E. Kluver, C. Vogelmeier, D. Becker, R. Hedrich, W. G. Forssmann and R. Bals (2001). "Identification of a novel, multifunctional beta-defensin (human beta-defensin 3) with specific antimicrobial activity. Its interaction with plasma membranes of Xenopus oocytes and the induction of macrophage chemoattraction." Cell and Tissue Research **306**(2): 257-264.

Ghosh, D., E. Porter, B. Shen, S. K. Lee, D. Wilk, J. Drazba, S. P. Yadav, J. W. Crabb, T. Ganz and C. L. Bevins (2002). "Paneth cell trypsin is the processing enzyme for human defensin-5." Nature Immunology **3**(6): 583-590.

Gibbs, R. A., George M. Weinstock, Michael L. Metzker1, Donna M. Muzny1, Erica J. Sodergren1, Steven Scherer1, Graham Scott1, David Steffen1, Kim C. Worley1 and Paula E. Burch1 (2004). "Genome sequence of the Brown Norway rat yields insights into mammalian evolution." Nature **428**(6982): 493-521.

Giglio, S., K. W. Broman, N. Matsumoto, V. Calvari, G. Gimelli, T. Neumann, H. Ohashi, L. Voullaire, D. Larizza, R. Giorda, J. L. Weber, D. H. Ledbetter and O. Zuffardi (2001). "Olfactory Receptor Gene Clusters, Genomic-Inversion Polymorphisms, and Common Chromosome Rearrangements." American Journal of Human Genetics **68**(4): 874-883.

Giglio, S., V. Calvari, G. Gregato, G. Gimelli, S. Camanini, R. Giorda, A. Ragusa, S. Guerneri, A. Selicorni, M. Stumm, H. Tonnies, M. Ventura, M. Zollino, G. Neri, J. Barber, D. Wieczorek, M. Rocchi and O. Zuffardi (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): 276-285.

Gimelli, G., M. A. Pujana, M. G. Patricelli, S. Russo, D. Giardino, L. Larizza, J. Cheung, L. Armengol, A. Schinzel, X. Estivill and O. Zuffardi (2003). "Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions." Human Molecular Genetics **12**(8): 849-858.

Gokcumen, O., P. Babb, R. Iskow, Q. Zhu, X. Shi, R. Mills, I. Ionita-Laza, E. Vallender, A. Clark, W. Johnson and C. Lee (2011). "Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection." Genome Biology **12**(5): R52.

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

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

Gottlieb, A. B. and J. G. Krueger (1990). "HLA region genes and immune activation in the pathogenesis of psoriasis." Archives of Dermatology **126**(8): 1083-1086.

Graubert, T. A., P. Cahan, D. Edwin, R. R. Selzer, T. A. Richmond, P. S. Eis, W. D. Shannon, X. Li, H. L. McLeod, J. M. Cheverud and T. J. Ley (2007). "A High-Resolution Map of Segmental DNA Copy Number Variation in the Mouse Genome." PLoS Genetics **3**(1): e3.

Groot, P. C., M. J. Bleeker, J. C. Pronk, F. Arwert, W. H. Mager, R. J. Planta, A. W. Eriksson and R. R. Frants (1989). "The Human Alpha-Amylase Multigene Family Consists of Haplotypes with Variable Numbers of Genes." Genomics **5**(1): 29-42.

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

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

Gudjonsson, J. E., A. Karason, A. Antonsdottir, E. H. Runarsdottir, V. B. Hauksson, R. Upmanyu, J. Gulcher, K. Stefansson and H. Valdimarsson (2003). "Psoriasis patients who are homozygous for the HLA-Cw*0602 allele

have a 2 5-fold increased risk of developing psoriasis compared with Cw6 heterozygotes." British Journal of Dermatology **148**(2): 233-235.

Guo, W.-J., F. Callif-Daley, M. C. Zapata and M. E. Miller (1995). "Clinical and cytogenetic findings in seven cases of inverted duplication of 8p with evidence of a telomeric deletion using fluorescence in situ hybridization." American Journal of Medical Genetics **58**(3): 230-236.

Guryev, V., K. Saar, T. Adamovic, M. Verheul, S. A. A. C. van Heesch, S. Cook, M. Pravenec, T. Aitman, H. Jacob, J. D. Shull, N. Hubner and E. Cuppen (2008). "Distribution and functional impact of DNA copy number variation in the rat." Nature Genetics **40**(5): 538-545.

Hamerton, J. L., N. Canning, M. Ray and S. Smith (1975). "A cytogenetic survey of 14,069 newborn infants. I. Incidence of chromosome abnormalities." Clinical Genetics **8**(4): 223-243.

Handsaker, R. E., J. M. Korn, J. Nemesh and S. A. McCarroll (2011). "Discovery and genotyping of genome structural polymorphism by sequencing on a population scale." Nature Genetics **43**(3): 269-276.

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

Harder, J., J. Bartels, E. Christophers and J. M. Schroder (2001). "Isolation and characterization of human beta-defensin-3, a novel human inducible peptide antibiotic." Journal of Biological Chemistry **276**(8): 5707-5713.

Harder, J., R. Siebert, Y. Zhang, P. Matthiesen, E. Christophers, B. Schlegelberger and J. M. Schroder (1997b). "Mapping of the gene encoding human beta-defensin-2 (DEFB2) to chromosome region 8p22-p23.1." Genomics **46**(3): 472-475.

Hardwick, R. J., L. R. Machado, L. W. Zuccherato, S. Antolinos, Y. Xue, N. Shawa, R. H. Gilman, L. Cabrera, D. E. Berg, C. Tyler-Smith, P. Kelly, E. Tarazona-Santos and E. J. Hollox (2011). "A worldwide analysis of beta-defensin copy number variation suggests recent selection of a high-expressing DEFB103 gene copy in East Asia." Human Mutation **32**(7): 743-750.

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

Harwig, S. S. L., K. M. Swiderek, V. N. Kokryakov, L. Tan, T. D. Lee, E. A. Panyutich, G. M. Aleshina, O. V. Shamova and R. I. Lehrer (1994). "Gallinacins: cysteine-rich antimicrobial peptides of chicken leukocytes." Federation of European Biochemical Societies Letters **342**(3): 281-285.

Hastings, P. J., G. Ira and J. R. Lupski (2009a). "A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation." PLoS Genetics **5**(1): e1000327.

Hastings, P. J., J. R. Lupski, S. M. Rosenberg and G. Ira (2009b). "Mechanisms of change in gene copy number." Nature Reviews Genetics **10**(8): 551-564.

Heid, C. A., J. Stevens, K. J. Livak and P. M. Williams (1996). "Real time quantitative PCR." Genome Research **6**(10): 986-994.

Helms, C., L. Cao, J. G. Krueger, E. M. Wijsman, F. Chamian, D. Gordon, M. Heffernan, J. A. W. Daw, J. Robarge, J. Ott, P.-Y. Kwok, A. Menter and A. M. Bowcock (2003). "A putative RUNX1 binding site variant between SLC9A3R1 and NAT9 is associated with susceptibility to psoriasis." Nature Genetics **35**(4): 349-356.

Helms, C., N. L. Saccone, L. Cao, J. A. Daw, K. Cao, T. M. Hsu, P. Taillon-Miller, S. Duan, D. Gordon, B. Pierce, J. Ott, J. Rice, M. A. Fernandez-Vina, P. Y. Kwok, A. Menter and A. M. Bowcock (2005). "Localization of PSORS1 to a haplotype block harboring HLA-C and distinct from corneodesmosin and HCR." Human Genetics **118**(3-4): 466-476.

Henrichsen, C. N., N. Vinckenbosch, S. Zollner, E. Chaignat, S. Pradervand, F. Schutz, M. Ruedi, H. Kaessmann and A. Reymond (2009). "Segmental copy number variation shapes tissue transcriptomes." Nature Genetics **41**(4): 424-429.

Herodez, S. S., B. Zagradisnik and N. K. Vokac (2005). "MLPA method for PMP22 gene analysis." Acta Chimica Slovenica **52**(2): 105-110.

Hewett, D., L. Samuelsson, J. Polding, F. Enlund, D. Smart, K. Cantone, C. G. See, S. Chadha, A. Inerot, C. Enerback, D. Montgomery, C. Christodolou, P. Robinson, P. Matthews, M. Plumpton, J. Wahlstrom, G. Swanbeck, T. Martinsson, A. Roses, J. Riley and I. Purvis (2002). "Identification of a psoriasis susceptibility candidate gene by linkage disequilibrium mapping with a localized single nucleotide polymorphism map." Genomics **79**(3): 305-314.

Higgs, D., M. Vickers, A. Wilkie, I. Pretorius, A. Jarman and D. Weatherall (1989). "A review of the molecular genetics of the human alpha-globin gene cluster." Blood **73**(5): 1081-1104.

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

Holla, O. L., C. Teie, K. E. Berge and T. P. Leren (2005). "Identification of deletions and duplications in the low density lipoprotein receptor gene by MLPA." Clinica Chimica Acta **356**(1-2): 164-171.

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

Hollox, E., T. Atia, G. Cross, T. Parkin and J. Armour (2002a). "High-throughput screening of human subtelomeric DNA for copy number changes

using Multiplex Amplifiable Probe Hybridisation (MAPH)." Journal of Medical Genetics **39**: 790 - 795.

Hollox, E. J. (2008). "Copy number variation of beta-defensins and relevance to disease." Cytogenetic and Genome Research **123**(1-4): 148-155.

Hollox, E. J., S. M. Akrami and J. A. Armour (2002b). "DNA copy number analysis by MAPH: molecular diagnostic applications." Expert Review of Molecular Diagnostics **2**(4): 370-378.

Hollox, E. J., J. A. L. Armour and J. C. K. Barber (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., J. C. K. Barber, A. J. Brookes and J. A. L. Armour (2008a). "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., J. Davies, U. Griesenbach, J. Burgess, E. W. F. W. Alton and J. A. L. Armour (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., J.-C. Detering and T. Dehnugara (2009). "An integrated approach for measuring copy number variation at the FCGR3 (CD16) locus." Human Mutation **30**(3): 477-484.

Hollox, E. J., U. Huffmeier, P. L. Zeeuwen, R. Palla, J. Lascorz, D. Rodijk-Olthuis, P. C. van de Kerkhof, H. Traupe, G. de Jongh, M. den Heijer, A. Reis, J. A. Armour and J. Schalkwijk (2008b). "Psoriasis is associated with increased beta-defensin genomic copy number." Nature Genetics **40**(1): 23-25.

Hoover, D. M., K. R. Rajashankar, R. Blumenthal, A. Puri, J. J. Oppenheim, O. Chertov and J. Lubkowski (2000). "The structure of human beta-defensin-2 shows evidence of higher order oligomerization." Journal of Biological Chemistry **275**(42): 32911-32918.

Hornstra, L. K., D. L. Nelson, S. T. Warren and T. P. Yang (1993). "High resolution methylation analysis of the FMR1 gene trinucleotide repeat region in fragile X syndrome." Human Molecular Genetics **2**(10): 1659-1665.

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

Hüffmeier, U., J. Lascorz, T. Becker, F. Schürmeier-Horst, A. Magener, A. B. Ekici, S. Endele, C. T. Thiel, S. Thoma-Uszynski, R. Mössner, K. Reich, W. Kurrat, T. F. Wienker, H. Traupe and A. Reis (2009). "Characterisation of psoriasis susceptibility locus 6 (PSORS6) in patients with early onset psoriasis

and evidence for interaction with PSORS1." Journal of Medical Genetics **46**(11): 736-744.

Hüffmeier, U., J. Lascorz, H. Traupe, B. Böhm, F. Schürmeier-Horst, M. Ständer, R. Kelsch, C. Baumann, W. Küster, H. Burkhardt and A. Reis (2005). "Systematic linkage disequilibrium analysis of SLC12A8 at PSORS5 confirms a role in susceptibility to psoriasis vulgaris." Journal of Investigative Dermatology **125**(5): 906-912.

Huffmeier, U., S. Uebe, A. B. Ekici, J. Bowes, E. Giardina, E. Korendowych, K. Juneblad, M. Apel, R. McManus, P. Ho, I. N. Bruce, A. W. Ryan, F. Behrens, J. Lascorz, B. Bohm, H. Traupe, J. Lohmann, C. Gieger, H.-E. Wichmann, C. Herold, M. Steffens, L. Klareskog, T. F. Wienker, O. FitzGerald, G.-M. Alenius, N. J. McHugh, G. Novelli, H. Burkhardt, A. Barton and A. Reis (2010). "Common variants at TRAF3IP2 are associated with susceptibility to psoriatic arthritis and psoriasis." Nature Genetics **42**(11): 996-999.

Huse, K., S. Taudien, M. Groth, P. Rosenstiel, K. Szafranski, M. Hiller, J. Hampe, K. Junker, J. Schubert, S. Schreiber, G. Birkenmeier, M. Krawczak and M. Platzer (2008). "Genetic variants of the copy number polymorphic beta-defensin locus are associated with sporadic prostate cancer." Tumour Biology **29**(2): 83-92.

Iafrate, A. J., L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer and C. Lee (2004). "Detection of large-scale variation in the human genome." Nature Genetics **36**(9): 949-951.

Ibrahim, G., R. Waxman and P. S. Helliwell (2009). "The prevalence of psoriatic arthritis in people with psoriasis." Arthritis & Rheumatism **61**(10): 1373-1378.

Inoue, K. and J. R. Lupski (2002). "Molecular mechanisms for genomic disorders." Annual Review of Genomics and Human Genetics **3**: 199-242.

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

Jacobs, P. A. and J. A. Strong (1959). "A Case of Human Intersexuality Having a Possible XXY Sex-Determining Mechanism." Nature **183**(4657): 302-303.

Janssen, B., C. Hartmann, V. Scholz, A. Jauch and J. Zschocke (2005). "MLPA analysis for the detection of deletions, duplications and complex rearrangements in the dystrophin gene: potential and pitfalls." Neurogenetics **6**(1): 29-35.

Janssens, W., H. Nuytten, L. J. Dupont, J. Van Eldere, S. Vermeire, D. Lambrechts, K. Nackaerts, M. Decramer, J.-J. Cassiman and H. Cuppens (2010). "Genomic Copy Number Determines Functional Expression of β-Defensin 2 in Airway Epithelial Cells and Associates with Chronic Obstructive

Pulmonary Disease." American Journal of Respiratory and Critical Care Medicine **182**(2): 163-169.

Jeffreys, A. J., V. Wilson and S. L. Thein (1985a). "Hypervariable "minisatellite" regions in human DNA." Nature **314**(6006): 67-73.

Jeffreys, A. J., V. Wilson and S. L. Thein (1985b). "Individual-specific "fingerprints" of human DNA." Nature **316**(6023): 76-79.

Ji, Y., E. E. Eichler, S. Schwartz and R. D. Nicholls (2000). "Structure of chromosomal duplicons and their role in mediating human genomic disorders." Genome Research **10**(5): 597-610.

Kallioniemi, A., T. Visakorpi, R. Karhu, D. Pinkel and O.-P. Kallioniemi (1996). "Gene Copy Number Analysis by Fluorescencein SituHybridization and Comparative Genomic Hybridization." Methods **9**(1): 113-121.

Kamysz, W., M. Okroj and J. Lukasiak (2003). "Novel properties of antimicrobial peptides." Acta Biochimica Polonica **50**(2): 461-469.

Karason, A., J. E. Gudjonsson, R. Upmanyu, A. A. Antonsdottir, V. B. Hauksson, E. H. Runasdottir, H. H. Jonsson, D. F. Gudbjartsson, M. L. Frigge, A. Kong, K. Stefansson, H. Valdimarsson and J. R. Gulcher (2003). "A susceptibility gene for psoriatic arthritis maps to chromosome 16q: evidence for imprinting." American Journal of Human Genetics **72**(1): 125-131.

Karin, M., T. Lawrence and V. Nizet (2006). "Innate immunity gone awry: linking microbial infections to chronic inflammation and cancer." Cell **124**(4): 823-835.

Kehrer-Sawatzki, H. and D. N. Cooper (2008). "Comparative analysis of copy number variation in primate genomes." Cytogenetic and Genome Research **123**(1-4): 288-296.

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

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

Knoll, J. H. M., R. D. Nicholls, R. E. Magenis, J. M. Graham, M. Lalande, S. A. Latt, J. M. Opitz and J. F. Reynolds (1989). "Angelman and Prader-Willi

syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion." American Journal of Medical Genetics **32**(2): 285-290.

Korbel, J. O., P. M. Kim, X. Chen, A. E. Urban, S. Weissman, M. Snyder and M. B. Gerstein (2008). "The current excitement about copy-number variation: how it relates to gene duplications and protein families." Current Opinion in Structural Biology **18**(3): 366-374.

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

Korver, W., M. W. Schilham, P. Moerer, M. J. van den Hoff, K. Dam, W. H. Lamers, R. H. Medema and H. Clevers (1998). "Uncoupling of S phase and mitosis in cardiomyocytes and hepatocytes lacking the winged-helix transcription factor Trident." Current Biology **8**(24): 1327-1330, S1321.

Lebofsky, R. and A. Bensimon (2003). "Single DNA molecule analysis: Applications of molecular combing." Briefings in Functional Genomics & Proteomics **1**(4): 385-396.

Lebwohl, M. (2003). "Psoriasis." Lancet **361**(9364): 1197-1204.

Lebwohl, M., S. K. Tyring, T. K. Hamilton, D. Toth, S. Glazer, N. H. Tawfik, P. Walicke, W. Dummer, X. Wang, M. R. Garovoy and D. Pariser (2003). "A novel targeted T-cell modulator, efalizumab, for plaque psoriasis." New England Journal of Medicine **349**(21): 2004-2013.

Ledbetter, D. H., V. M. Riccardi, S. D. Airhart, R. J. Strobel, B. S. Keenan and J. D. Crawford (1981). "Deletions of Chromosome 15 as a Cause of the Prader-Willi Syndrome." New England Journal of Medicine **304**(6): 325-329.

Lee, A. S., M. Gutiérrez-Arcelus, G. H. Perry, E. J. Vallender, W. E. Johnson, G. M. Miller, J. O. Korbel and C. Lee (2008). "Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies." Human Molecular Genetics **17**(8): 1127-1136.

Lee, C. and S. W. Scherer (2010). "The clinical context of copy number variation in the human genome." Expert Reviews in Molecular Medicine **12**: e8.

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

Lee, Y.-A., F. Rüschendorf, C. Windemuth, M. Schmitt-Egenolf, A. Stadelmann, G. Nürnberg, M. Ständer, T. F. Wienker, A. Reis and H. Traupe (2000). "Genomewide Scan in German Families Reveals Evidence for a Novel Psoriasis-Susceptibility Locus on Chromosome 19p13." American Journal of Human Genetics **67**(4): 1020-1024.

Lehrer, R. I. (2004). "Primate defensins." Nature Reviews Microbiology **2**(9): 727-738.

Lehrer, R. I. and T. Ganz (1999). "Antimicrobial peptides in mammalian and insect host defence." Current Opinion in Immunology **11**(1): 23-27.

Lejeune, J., Gautier, M. and Turpin, R. (1959). "Study of somatic chromosomes from 9 mongoloid children." Comptes rendus hebdomadaires des séances de l'Académie des sciences **248**: 1721-1722.

Leonova, L., V. N. Kokryakov, G. Aleshina, T. Hong, T. Nguyen, C. Zhao, A. J. Waring and R. I. Lehrer (2001). "Circular minidefensins and posttranslational generation of molecular diversity." Journal of Leukocyte Biology **70**(3): 461-464.

Leung, D. Y. M., J. B. Travers, R. Giorno, D. A. Norris, R. Skinner, J. Aelion, L. V. Kazemi, M. H. Kim, A. E. Trumble, M. Koth and P. M. Schlievert (1995). "Evidence for a streptococcal superantigen-driven process in acute guttate psoriasis." Journal of Clinical Investigation **96**(5): 2106-2112.

Levsky, J. M. and R. H. Singer (2003). "Fluorescence in situ hybridization: past, present and future." Journal of Cell Science **116**(14): 2833-2838.

Li, J., T. Jiang, J.-H. Mao, A. Balmain, L. Peterson, C. Harris, P. H. Rao, P. Havlak, R. Gibbs and W.-W. Cai (2004). "Genomic segmental polymorphisms in inbred mouse strains." Nature Genetics **36**(9): 952-954.

Lichtenstein, A., T. Ganz, M. Selsted and R. Lehrer (1986). "In vitro tumor cell cytolysis mediated by peptide defensins of human and rabbit granulocytes." Blood **68**(6): 1407-1410.

Lieber, M. R., Y. Ma, U. Pannicke and K. Schwarz (2003). "Mechanism and regulation of human non-homologous DNA end-joining." Nature Reviews Molecular Cell Biology **4**(9): 712-720.

Lim, S.-Y., T. Chan, R. S. Gelman, J. B. Whitney, K. L. O'Brien, D. H. Barouch, D. B. Goldstein, B. F. Haynes and N. L. Letvin (2010). "Contributions of *Mamu-A*01* Status and *TRIM5* Allele Expression, But Not *CCL3L* Copy Number Variation, to the Control of SIVmac251 Replication in Indian-Origin Rhesus Monkeys." PLoS Genetics **6**(6): e1000997.

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

Linzmeier, R. and T. Ganz (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 - 430.

Linzmeier, R., C. H. Ho, B. V. Hoang and T. Ganz (1999). "A 450-kb contig of defensin genes on human chromosome 8p23." Gene **233**(1-2): 205-211.

Linzmeier, R., D. Michaelson, L. Liu and T. Ganz (1993). "The structure of neutrophil defensin genes." Federation of European Biochemical Societies Letters **321**(2-3): 267-273.

Liu, L., C. Zhao, H. H. Heng and T. Ganz (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): 316-320.

Liu, Y., C. Helms, W. Liao, L. C. Zaba, S. Duan, J. Gardner, C. Wise, A. Miner, M. J. Malloy, C. R. Pullinger, J. P. Kane, S. Saccone, J. Worthington, I. Bruce, P. Y. Kwok, A. Menter, J. Krueger, A. Barton, N. L. Saccone and A. M. Bowcock (2008). "A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci." PLoS Genetics **4**(3): e1000041.

Locke, D. P., L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, D. M. Muzny, S.-P. Yang, Z. Wang, A. T. Chinwalla, P. Minx, M. Mitreva, L. Cook, K. D. Delehaunty, C. Fronick, H. Schmidt, L. A. Fulton, R. S. Fulton, J. O. Nelson, V. Magrini, C. Pohl, T. A. Graves, C. Markovic, A. Cree, H. H. Dinh, J. Hume, C. L. Kovar, G. R. Fowler, G. Lunter, S. Meader, A. Heger, C. P. Ponting, T. Marques-Bonet, C. Alkan, L. Chen, Z. Cheng, J. M. Kidd, E. E. Eichler, S. White, S. Searle, A. J. Vilella, Y. Chen, P. Flicek, J. Ma, B. Raney, B. Suh, R. Burhans, J. Herrero, D. Haussler, R. Faria, O. Fernando, F. Darre, D. Farre, E. Gazave, M. Oliva, A. Navarro, R. Roberto, O. Capozzi, N. Archidiacono, G. D. Valle, S. Purgato, M. Rocchi, M. K. Konkel, J. A. Walker, B. Ullmer, M. A. Batzer, A. F. A. Smit, R. Hubley, C. Casola, D. R. Schrider, M. W. Hahn, V. Quesada, X. S. Puente, G. R. Ordonez, C. Lopez-Otin, T. Vinar, B. Brejova, A. Ratan, R. S. Harris, W. Miller, C. Kosiol, H. A. Lawson, V. Taliwal, A. L. Martins, A. Siepel, A. RoyChoudhury, X. Ma, J. Degenhardt, C. D. Bustamante, R. N. Gutenkunst, T. Mailund, J. Y. Dutheil, A. Hobolth, M. H. Schierup, O. A. Ryder, Y. Yoshinaga, P. J. de Jong, G. M. Weinstock, J. Rogers, E. R. Mardis, R. A. Gibbs and R. K. Wilson (2011). "Comparative and demographic analysis of orang-utan genomes." Nature **469**(7331): 529-533.

Locke, D. P., A. J. Sharp, S. A. McCarroll, S. D. McGrath, T. L. Newman, Z. Cheng, S. Schwartz, D. G. Albertson, D. Pinkel, D. M. Altshuler and E. E. Eichler (2006). "Linkage Disequilibrium and Heritability of Copy-Number Polymorphisms within Duplicated Regions of the Human Genome." American Journal of Human Genetics **79**(2): 275-290.

Lowes, M. A., A. M. Bowcock and J. G. Krueger (2007). "Pathogenesis and therapy of psoriasis." Nature **445**(7130): 866-873.

Lu, H., W. Ouyang and C. Huang (2006). "Inflammation, a Key Event in Cancer Development." Molecular Cancer Research **4**(4): 221-233.

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., C. M. Ambrose, M. P. Duyao, R. H. Myers, C. Lin, L. Srinidhi, G. Barnes, S. A. Taylor, M. James, N. Groot, H. MacFarlane, B. Jenkins, M. A. Anderson, N. S. Wexler, J. F. Gusella, G. P. Bates, S. Baxendale, H. Hummerich, S. Kirby, M. North, S. Youngman, R. Mott, G. Zehetner, Z. Sedlacek, A. Poustka, A.-M. Frischauf, H. Lehrach, A. J. Buckler, D. Church, L. Doucette-Stamm, M. C. O'Donovan, L. Riba-Ramirez, M. Shah, V. P. Stanton, S. A. Strobel, K. M. Draths, J. L. Wales, P. Dervan, D. E. Housman, M. Altherr, R. Shiang, L. Thompson, T. Fielder, J. J. Wasmuth, D. Tagle, J. Valdes, L. Elmer, M. Allard, L. Castilla, M. Swaroop, K. Blanchard, F. S. Collins, R. Snell, T. Holloway, K. Gillespie, N. Datson, D. Shaw and P. S. Harper (1993). "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes." Cell **72**(6): 971-983.

Mackay, C. R. (2005). "CCL3L1 dose and HIV-1 susceptibility." Trends in Molecular Medicine **11**(5): 203-206.

Mackewicz, C. E., J. Yuan, P. Tran, L. Diaz, E. Mack, M. E. Selsted and J. A. Levy (2003). "alpha-Defensins can have anti-HIV activity but are not CD8 cell anti-HIV factors." Journal of Acquired Immune Deficiency Syndromes **17**(14): F23-F32.

Mamtani, M., B. Rovin, R. Brey, J. F. Camargo, H. Kulkarni, M. Herrera, P. Correa, S. Holliday, J.-M. Anaya and S. K. Ahuja (2008). "CCL3L1 gene-containing segmental duplications and polymorphisms in CCR5 affect risk of systemic lupus erythaematosus." Annals of the Rheumatic Diseases **67**(8): 1076-1083.

Marques-Bonet, T., J. M. Kidd, M. Ventura, T. A. Graves, Z. Cheng, L. W. Hillier, Z. Jiang, C. Baker, R. Malfavon-Borja, L. A. Fulton, C. Alkan, G. Aksay, S. Girirajan, P. Siswara, L. Chen, M. F. Cardone, A. Navarro, E. R. Mardis, R. K. Wilson and E. E. Eichler (2009). "A burst of segmental duplications in the genome of the African great ape ancestor." Nature **457**(7231): 877-881.

Mars, W. M., P. Patmasiriwat, T. Maity, V. Huff, M. M. Weil and G. F. Saunders (1995). "Inheritance of Unequal Numbers of the Genes Encoding the Human Neutrophil Defensins HP-1 and HP-3." Journal of Biological Chemistry **270**(51): 30371-30376.

Matthews, D., L. Fry, A. Powles, J. Weber, M. McCarthy, E. Fisher, K. Davies and R. Williamson (1996). "Evidence that a locus for familial psoriasis maps to chromosome 4q." Nature Genetics **14**(2): 231-233.

McCarroll, S. A. and D. M. Altshuler (2007). "Copy-number variation and association studies of human disease." Nature Genetics **39**(7 Suppl): S37-42.

McCarroll, S. A., T. N. Hadnott, G. H. Perry, P. C. Sabeti, M. C. Zody, J. C. Barrett, S. Dallaire, S. B. Gabriel, C. Lee, M. J. Daly and D. M. Altshuler (2006). "Common deletion polymorphisms in the human genome." Nature Genetics **38**(1): 86-92.

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

McClintock, B. (1939). "The behavior in successive nuclear divisions of a chromosome broken at meiosis." Proceedings of the National Academy of Sciences, USA **25**(8): 405-416.

McKinney, C., M. E. Merriman, P. T. Chapman, P. J. Gow, A. A. Harrison, J. Highton, P. B. B. Jones, L. McLean, J. L. O'Donnell, V. Pokorny, M. Spellerberg, L. K. Stamp, J. Willis, S. Steer and T. R. Merriman (2008). "Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis." Annals of the Rheumatic Diseases **67**(3): 409-413.

McNaughton, D., W. Knight, R. Guerreiro, N. Ryan, J. Lowe, M. Poulter, D. J. Nicholl, J. Hardy, T. Revesz, J. Lowe, M. Rossor, J. Collinge and S. Mead (2012). "Duplication of amyloid precursor protein (APP), but not prion protein (PRNP) gene is a significant cause of early onset dementia in a large UK series." Neurobiology of Aging **33**(2): 426.e413-426.e421.

Michalik, A. and C. Van Broeckhoven (2003). "Pathogenesis of polyglutamine disorders: aggregation revisited." Human Molecular Genetics **12**(suppl 2): R173-R186.

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

Mills, R. E., C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, W. S. Pittard and S. E. Devine (2006). "An initial map of insertion and deletion (INDEL) variation in the human genome." Genome Research **16**(9): 1182-1190.

Mills, R. E., W. S. Pittard, J. M. Mullaney, U. Farooq, T. H. Creasy, A. A. Mahurkar, D. M. Kemeza, D. S. Strassler, C. P. Ponting, C. Webber and S. E. Devine (2011). "Natural genetic variation caused by small insertions and deletions in the human genome." Genome Research **21**(6): 830-839.

Murphy, C. J., B. A. Foster, M. J. Mannis, M. E. Selsted and T. W. Reid (1993). "Defensins are mitogenic for epithelial cells and fibroblasts." Journal of Cellular Physiology **155**(2): 408-413.

Nair, R. P., K. C. Duffin, C. Helms, J. Ding, P. E. Stuart, D. Goldgar, J. E. Gudjonsson, Y. Li, T. Tejasvi, B.-J. Feng, A. Ruether, S. Schreiber, M. Weichenthal, D. Gladman, P. Rahman, S. J. Schrodi, S. Prahalad, S. L. Guthery, J. Fischer, W. Liao, P.-Y. Kwok, A. Menter, G. M. Lathrop, C. A. Wise, A. B. Begovich, J. J. Voorhees, J. T. Elder, G. G. Krueger, A. M. Bowcock and G. R. Abecasis (2009). "Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways." Nature Genetics **41**(2): 199-204.

Nair, R. P., T. Henseler, S. Jenisch, P. Stuart, C. K. Bichakjian, W. Lenk, E. Westphal, S.-W. Guo, E. Christophers, J. J. Voorhees and J. T. Elder (1997). "Evidence for Two Psoriasis Susceptibility Loci (HLA and 17q) and Two Novel Candidate Regions (16q and 20p) by Genome-Wide Scan." Human Molecular Genetics **6**(8): 1349-1356.

Nair, R. P., P. Stuart, T. Henseler, S. Jenisch, N. V. C. Chia, E. Westphal, N. J. Schork, J. Kim, H. W. Lim, E. Christophers, J. J. Voorhees and J. T. Elder (2000). "Localization of Psoriasis-Susceptibility Locus PSORS1 to a 60-kb Interval Telomeric to HLA-C." American Journal of Human Genetics **66**(6): 1833-1844.

Nair, R. P., P. E. Stuart, I. Nistor, R. Hiremagalore, N. V. Chia, S. Jenisch, M. Weichenthal, G. R. Abecasis, H. W. Lim, E. Christophers, J. J. Voorhees and J. T. Elder (2006). "Sequence and haplotype analysis supports HLA-C as the psoriasis susceptibility 1 gene." American Journal of Human Genetics **78**(5): 827-851.

Nakashima, H., N. Yamamoto, M. Masuda and N. Fujii (1993). "Defensins inhibit HIV replication in vitro." Journal of Acquired Immune Deficiency Syndromes **7**(8): 1129.

Nathans, J., T. Piantanida, R. Eddy, T. Shows and D. Hogness (1986). "Molecular genetics of inherited variation in human color vision." Science **232**(4747): 203-210.

Nguyen, D.-Q., C. Webber, J. Hehir-Kwa, R. Pfundt, J. Veltman and C. P. Ponting (2008). "Reduced purifying selection prevails over positive selection in human copy number variant evolution." Genome Research **18**(11): 1711-1723.

Nguyen, D.-Q., C. Webber and C. P. Ponting (2006). "Bias of Selection on Human Copy-Number Variants." PLoS Genetics **2**(2): e20.

Nguyen, K., P. Walrafen, R. Bernard, S. Attarian, C. Chaix, C. Vovan, E. Renard, N. Dufrane, J. Pouget, A. Vannier, A. Bensimon and N. Lévy (2011). "Molecular combing reveals allelic combinations in facioscapulohumeral dystrophy." Annals of Neurology **70**(4): 627-633.

Nguyen, T. X., A. M. Cole and R. I. Lehrer (2003). "Evolution of primate theta-defensins: a serpentine path to a sweet tooth." Peptides **24**(11): 1647-1654.

Nicastro, G., L. Franzoni, C. de Chiara, A. C. Mancin, J. R. Giglio and A. Spisni (2003). "Solution structure of crotamine, a Na+ channel affecting toxin from *Crotalus durissus terrificus* venom." European Journal of Biochemistry **270**(9): 1969-1979.

Nickoloff, B., R. Mitra, J. Green, X. Zheng, Y. Shimizu, C. Thompson and L. Turka (1993). "Accessory cell function of keratinocytes for superantigens. Dependence on lymphocyte function-associated antigen-1/intercellular adhesion molecule-1 interaction." The Journal of Immunology **150**(6): 2148-2159.

Nickoloff, B. J., L. A. Turka, R. S. Mitra and F. O. Nestle (1995). "Direct and indirect control of T-cell activation by keratinocytes." Journal of Investigative Dermatology **105**(1 Suppl): 25S-29S.

Niyonsaba, F., H. Ogawa and I. Nagaoka (2004). "Human beta-defensin-2 functions as a chemotactic agent for tumour necrosis factor-alpha-treated human neutrophils." Immunology **111**(3): 273-281.

Niyonsaba, F., H. Ushio, N. Nakano, W. Ng, K. Sayama, K. Hashimoto, I. Nagaoka, K. Okumura and H. Ogawa (2007). "Antimicrobial peptides human beta-defensins stimulate epidermal keratinocyte migration, proliferation and production of proinflammatory cytokines and chemokines." Journal of Investigative Dermatology **127**(3): 594-604.

Nozawa, M., Y. Kawahara and M. Nei (2007). "Genomic drift and copy number variation of sensory receptor genes in humans." Proceedings of the National Academy of Sciences, USA **104**(51): 20421-20426.

Ohno, S., W. D. Kaplan and R. Kinosita (1959). "Formation of the sex chromatin by a single X-chromosome in liver cells of Rattus norvegicus." Experimental Cell Research **18**(2): 415-418.

Ollmann, M. M., M. L. Lamoreux, B. D. Wilson and G. S. Barsh (1998). "Interaction of Agouti protein with the melanocortin 1 receptor in vitro and in vivo." Genes & Development **12**(3): 316-330.

Orrù, S., E. Giuressi, C. Carcassi, M. Casula and L. Contu (2005). "Mapping of the Major Psoriasis-Susceptibility Locus (PSORS1) in a 70-Kb Interval around the Corneodesmosin Gene (CDSN)." American Journal of Human Genetics **76**(1): 164-171.

Osborne, L. R., M. Li, B. Pober, D. Chitayat, J. Bodurtha, A. Mandel, T. Costa, T. Grebe, S. Cox, L. C. Tsui and S. W. Scherer (2001). "A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome." Nature Genetics **29**(3): 321-325.

Ottolenghi, S., W. Lanyon, J. Paul, R. Williamson, D. Weatherall, J. Clegg, J. Pritchard, S. Pootrakul and W. Boon (1974). "The severe form of alpha thalassaemia is caused by a haemoglobin gene deletion." Nature **251**((5474)): 389-392.

Patau, K., D. W. Smith, E. Therman, S. L. Inhorn and H. P. Wagner (1960). "Multiple congenital anomaly caused by an extra autosome." Lancet **1**(7128): 790-793.

Patil, A., A. L. Hughes and G. Zhang (2004). "Rapid evolution and diversification of mammalian α-defensins as revealed by comparative analysis of rodent and primate genes." Physiological Genomics **20**(1): 1-11.

Patil, A. A., Y. B. Cai, Y. M. Sang, F. Blecha and G. L. Zhang (2005). "Cross-species analysis of the mammalian beta-defensin gene family: presence of syntenic gene clusters and preferential expression in the male reproductive tract." Physiological Genomics **23**(1): 5-17.

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

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

Perry, G. H., F. Yang, T. Marques-Bonet, C. Murphy, T. Fitzgerald, A. S. Lee, C. Hyland, A. C. Stone, M. E. Hurles, C. Tyler-Smith, E. E. Eichler, N. P. Carter, C. Lee and R. Redon (2008). "Copy number variation and evolution in humans and chimpanzees." Genome Research **18**(11): 1698-1710.

Philip, M., D. A. Rowley and H. Schreiber (2004). "Inflammation as a tumor promoter in cancer induction." Seminars in Cancer Biology **14**(6): 433-439.

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

Pinkel, D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray and D. G. Albertson (1998). "High resolution analysis of DNA copy number variation

using comparative genomic hybridization to microarrays." Nature Genetics **20**(2): 207-211.

Plagnol, V., J. D. Cooper, J. A. Todd and D. G. Clayton (2007). "A Method to Address Differential Bias in Genotyping in Large-Scale Association Studies." PLoS Genetics **3**(5): e74.

Quiñones-Mateu, M. E., M. M. Lederman, Z. Feng, B. Chakraborty, J. Weber, H. R. Rangel, M. L. Marotta, M. Mirza, B. Jiang, P. Kiser, K. Medvik, S. F. Sieg and A. Weinberg (2003). "Human epithelial beta-defensins 2 and 3 inhibit HIV-1 replication." Journal of Acquired Immune Deficiency Syndromes **17**(16): F39-F48.

Rabe, K. F., S. Hurd, A. Anzueto, P. J. Barnes, S. A. Buist, P. Calverley, Y. Fukuchi, C. Jenkins, R. Rodriguez-Roisin, C. van Weel and J. Zielinski (2007). "Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: GOLD Executive Summary." American Journal of Respiratory and Critical Care Medicine **176**(6): 532-555.

Rad, I. A., A. L. Sharif, J. A. Raeburn, G. Evans, F. Lalloo, P. Morrison, J. A. L. Armour and G. S. Cross (2001). "Detection of BRCA1 whole exon deletions and duplications in breast/ovarian cancer families from UK by Multiplex Amplifiable Probe Hybridisation (MAPH)." Journal of Medical Genetics **38**: S22.

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

Ritchie, R. J., M.-G. Mattei and M. Lalande (1998). "A Large Polymorphic Repeat in the Pericentromeric Region of Human Chromosome 15q Contains Three Partial Gene Duplications." Human Molecular Genetics **7**(8): 1253-1260.

Robertson, W. R. B. (1916). "Chromosome studies. I. Taxonomic relationships shown in the chromosomes of tettigidae and acrididae: V-shaped chromosomes and their significance in acrididae, locustidae, and gryllidae: Chromosomes and variation." Journal of Morphology **27**(2): 179-331.

Rodriguez-Jimenez, F. J., A. Krause, S. Schulz, W. G. Forssmann, J. R. Conejo-Garcia, R. Schreeb and D. Motzkus (2003). "Distribution of new human beta-defensin genes clustered on chromosome 20 in functionally different segments of epididymis." Genomics **81**(2): 175-183.

Rowley, J. D. (1973). "A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining." Nature **243**(5405): 290-293.

Russell, T. J., L. M. Schultes and D. J. Kuban (1972). "Histocompatibility (HL-A) Antigens Associated with Psoriasis." New England Journal of Medicine **287**(15): 738-740.

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

Schaschl, H., T. J. Aitman and T. J. Vyse (2009). "Copy number variation in the human genome and its implication in autoimmunity." Clinical & Experimental Immunology **156**(1): 12-16.

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

Schroder, J. M. (1999). "Epithelial antimicrobial peptides: innate local host response elements." Cellular and Molecular Life Sciences **56**(1-2): 32-46.

Schroeder, B. O., Z. Wu, S. Nuding, S. Groscurth, M. Marcinowski, J. Beisner, J. Buchner, M. Schaller, E. F. Stange and J. Wehkamp (2011). "Reduction of disulphide bonds unmasks potent antimicrobial activity of human [bgr]-defensin 1." Nature **469**(7330): 419-423.

Schullerus, D., R. von Knobloch, J. Chudek, J. Herbers and G. Kovacs (1999). "Microsatellite analysis reveals deletion of a large region at chromosome 8p in conventional renal cell carcinoma." International Journal of Cancer **80**(1): 22-24.

Schutte, B. C. and P. B. McCray, Jr. (2002). "β-defensins in lung host defense." Annual Review of Physiology **64**: 709-748.

Schutte, B. C., J. P. Mitros, J. A. Bartlett, J. D. Walters, H. P. Jia, M. J. Welsh, T. L. Casavant and P. B. McCray, Jr. (2002). "Discovery of five conserved beta -defensin gene clusters using a computational search strategy." Proceedings of the National Academy of Sciences, USA **99**(4): 2129-2133.

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

Selsted, M. E. and A. J. Ouellette (2005). "Mammalian defensins in the antimicrobial immune response." Nature Immunology **6**(6): 551-557.

Semple, C., P. Gautier, K. Taylor and J. Dorin (2006). "The changing of the guard: Molecular diversity and rapid evolution of beta-defensins." Molecular Diversity **10**(4): 575 - 584.

Semple, C., A. Maxwell, P. Gautier, F. Kilanowski, H. Eastwood, P. Barran and J. Dorin (2005). "The complexity of selection at the major primate beta-defensin locus." BMC Evolotionary Biology **5**(1): 32.

Semple, F., S. Webb, H.-N. Li, H. B. Patel, M. Perretti, I. J. Jackson, M. Gray, D. J. Davidson and J. R. Dorin (2010). "Human β-defensin 3 has immunosuppressive activity in vitro and in vivo." European Journal of Immunology **40**(4): 1073-1078.

Shanahan, F. (2002). "Crohn's disease." The Lancet **359**(9300): 62-69.

Shao, W., J. Tang, W. Song, C. Wang, Y. Li, C. M. Wilson and R. A. Kaslow (2007). "CCL3L1 and CCL4L1: variable gene copy number in adolescents with and without human immunodeficiency virus type 1 (HIV-1) infection." Genes and Immunity **8**(3): 224-231.

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

Sharp, A. J., D. P. Locke, S. D. McGrath, Z. Cheng, J. A. Bailey, R. U. Vallente, L. M. Pertz, R. A. Clark, S. Schwartz, R. Segaves, V. V. Oseroff, D. G. Albertson, D. Pinkel and E. E. Eichler (2005). "Segmental Duplications and Copy-Number Variation in the Human Genome." American Journal of Human Genetics **77**(1): 78-88.

She, X., J. E. Horvath, Z. Jiang, G. Liu, T. S. Furey, L. Christ, R. Clark, T. Graves, C. L. Gulden, C. Alkan, J. A. Bailey, C. Sahinalp, M. Rocchi, D. Haussler, R. K. Wilson, W. Miller, S. Schwartz and E. E. Eichler (2004). "The structure and evolution of centromeric transition regions within the human genome." Nature **430**(7002): 857-864.

Sismani, C., J. A. Armour, J. Flint, C. Girgalli, R. Regan and P. C. Patsalis (2001). "Screening for subtelomeric chromosome abnormalities in children with idiopathic mental retardation using multiprobe telomeric FISH and the new MAPH telomeric assay." European Journal of Human Genetics **9**(7): 527-532.

Slater, H. R., D. L. Bruno, H. Ren, M. Pertile, J. P. Schouten and K. H. A. Choo (2003). "Rapid, high throughput prenatal detection of aneuploidy using a novel quantitative method (MLPA)." Journal of Medical Genetics **40**(12): 907-912.

Sleegers, K., N. Brouwers, I. Gijselinck, J. Theuns, D. Goossens, J. Wauters, J. Del-Favero, M. Cruts, C. M. v. Duijn and C. V. Broeckhoven (2006). "APP duplication is sufficient to cause early onset Alzheimer's dementia with cerebral amyloid angiopathy." Brain **129**(11): 2977-2983.

Solinas-Toldo, S., S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Döhner, T. Cremer and P. Lichter (1997). "Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances." Genes, Chromosomes and Cancer **20**(4): 399-407.

Soto, E., J. Espinoza, J. K. Nien, J. P. Kusanovic, O. Erez, K. Richani, J. Santolaya-Forgas and R. Romero (2007). "Human β-defensin-2: A natural antimicrobial peptide present in amniotic fluid participates in the host response to microbial invasion of the amniotic cavity." Journal of Maternal-Fetal and Neonatal Medicine **20**(1): 15-22.

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

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

Stankiewicz, P. and J. R. Lupski (2010). "Structural variation in the human genome and its role in disease." Annual Review of Medicine **61**: 437-455.

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

Stimpson, K. M., I. Y. Song, A. Jauch, H. Holtgreve-Grez, K. E. Hayden, J. M. Bridger and B. A. Sullivan (2010). "Telomere disruption results in non-random formation of de novo dicentric chromosomes involving acrocentric human chromosomes." PLoS Genetics **6**(8): e1001061.

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

Stuart, P. E., R. P. Nair, E. Ellinghaus, J. Ding, T. Tejasvi, J. E. Gudjonsson, Y. Li, S. Weidinger, B. Eberlein, C. Gieger, H. E. Wichmann, M. Kunz, R. Ike, G. G. Krueger, A. M. Bowcock, U. Mrowietz, H. W. Lim, J. J. Voorhees, G. R. Abecasis, M. Weichenthal, A. Franke, P. Rahman, D. D. Gladman and J. T. Elder (2010). "Genome-wide association analysis identifies three psoriasis susceptibility loci." Nature Genetics **42**(11): 1000-1004.

Sudmant, P. H., J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, A. Tsalenko, N. Sampas, L. Bruhn, J. Shendure, E. E. Eichler and G. Project (2010).

"Diversity of Human Copy Number Variation and Multicopy Genes." Science **330**(6004): 641-646.

Sugawara, H., N. Harada, T. Ida, T. Ishida, D. H. Ledbetter, K.-i. Yoshiura, T. Ohta, T. Kishino, N. Niikawa and N. Matsumoto (2003). "Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23." Genomics **82**(2): 238-244.

Sun, C. Q., R. Arnold, C. Fernandez-Golarz, A. B. Parrish, T. Almekinder, J. He, S.-m. Ho, P. Svoboda, J. Pohl, F. F. Marshall and J. A. Petros (2006). "Human β-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.

Sun, L.-D., H. Cheng, Z.-X. Wang, A.-P. Zhang, P.-G. Wang, J.-H. Xu, Q.-X. Zhu, H.-S. Zhou, E. Ellinghaus, F.-R. Zhang, X.-M. Pu, X.-Q. Yang, J.-Z. Zhang, A.-E. Xu, R.-N. Wu, L.-M. Xu, L. Peng, C. A. Helms, Y.-Q. Ren, C. Zhang, S.-M. Zhang, R. P. Nair, H.-Y. Wang, G.-S. Lin, P. E. Stuart, X. Fan, G. Chen, T. Tejasvi, P. Li, J. Zhu, Z.-M. Li, H.-M. Ge, M. Weichenthal, W.-Z. Ye, C. Zhang, S.-K. Shen, B.-Q. Yang, Y.-Y. Sun, S.-S. Li, Y. Lin, J.-H. Jiang, C.-T. Li, R.-X. Chen, J. Cheng, X. Jiang, P. Zhang, W.-M. Song, J. Tang, H.-Q. Zhang, L. Sun, J. Cui, L.-J. Zhang, B. Tang, F. Huang, Q. Qin, X.-P. Pei, A.-M. Zhou, L.-M. Shao, J.-L. Liu, F.-Y. Zhang, W.-D. Du, A. Franke, A. M. Bowcock, J. T. Elder, J.-J. Liu, S. Yang and X.-J. Zhang (2010). "Association analyses identify six new psoriasis susceptibility loci in the Chinese population." Nature Genetics **42**(11): 1005-1009.

Sun, L., C. M. Finnegan, T. Kish-Catalone, R. Blumenthal, P. Garzino-Demo, G. M. La Terra Maggiore, S. Berrone, C. Kleinman, Z. Wu, S. Abdelwahab, W. Lu and A. Garzino-Demo (2005). "Human β-Defensins Suppress Human Immunodeficiency Virus Infection: Potential Role in Mucosal Protection." Journal of Virology **79**(22): 14318-14329.

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

Taudien, S., P. Galgoczy, K. Huse, K. Reichwald, M. Schilhabel, K. Szafranski, A. Shimizu, S. Asakawa, A. Frankish, I. F. Loncarevic, N. Shimizu, R. Siddiqui and M. Platzer (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.

Territo, M. C., T. Ganz, M. E. Selsted and R. Lehrer (1989). "Monocyte-chemotactic activity of defensins from human neutrophils." The Journal of Clinical Investigation **84**(6): 2017-2020.

The International HapMap 3 Consortium (2010). "Integrating common and rare genetic variation in diverse human populations." Nature **467**(7311): 52-58.

The Wellcome Trust Case Control Consortium (2010). "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls." Nature **464**(7289): 713-720.

Thouzeau, C., Y. Le Maho, G. Froget, L. Sabatier, C. Le Bohec, J. A. Hoffmann and P. Bulet (2003). "Spheniscins, Avian β-Defensins in Preserved Stomach Contents of the King Penguin, Aptenodytes patagonicus." Journal of Biological Chemistry **278**(51): 51053-51058.

Tiilikainen, A., A. Lassus, J. Karvonen, P. Vartiainen and M. Julin (1980). "Psoriasis and HLA-Cw6." British Journal of Dermatology **102**(2): 179-184.

Tjio, J. H. and A. Levan (1956). "The chromosome number of man." Hereditas **42**: 1-6.

Tomfohrde, J., A. Silverman, R. Barnes, M. A. Fernandez-Vina, M. Young, D. Lory, L. Morris, K. D. Wuepper, P. Stastny, A. Menter and et al. (1994). "Gene for familial psoriasis susceptibility mapped to the distal end of human chromosome 17q." Science **264**(5162): 1141-1145.

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

Tran, D., P. A. Tran, Y.-Q. Tang, J. Yuan, T. Cole and M. E. Selsted (2002). "Homodimeric θ-Defensins from Rhesus macaqueLeukocytes." Journal of Biological Chemistry **277**(5): 3079-3084.

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

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

Tyson, J., T. Majerus, S. Walker and J. Armour (2009). "Quadruplex MAPH: improvement of throughput in high-resolution copy number screening." BMC Genomics **10**(1): 453.

Tyson, J., T. M. O. Majerus, S. Walker and J. A. L. Armour (2010). "Screening for common copy-number variants in cancer genes." Cancer Genetics and Cytogenetics **203**(2): 316-323.

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

Valdimarsson, H., B. S. Baker, I. Jónsdóttir, A. Powles and L. Fry (1995). "Psoriasis: a T-cell-mediated autoimmune disease induced by streptococcal superantigens?" Immunology Today **16**(3): 145-149.

van der Maarel, S. M. and R. R. Frants (2005). "The D4Z4 Repeat-Mediated Pathogenesis of Facioscapulohumeral Muscular Dystrophy." The American Journal of Human Genetics **76**(3): 375-386.

Van Wetering, S., S. Mannesse-Lazeroms, J. Dijkman and P. Hiemstra (1997). "Effect of neutrophil serine proteinases and defensins on lung epithelial cells: modulation of cytotoxicity and IL-8 production." Journal of Leukocyte Biology **62**(2): 217-226.

Vaurs-Barriere, C., M. N. Bonnet-Dupeyron, P. Combes, F. Gauthier-Barichard, X. T. Reveles, R. Schiffmann, E. Bertini, D. Rodriguez, P. Vago, J. A. L. Armour, P. Saugier-Veber, T. Frebourg, R. J. Leach and O. Boespflug-Tanguy (2006). "Golli-MBP Copy Number Analysis by FISH, QMPSF and MAPH in 195 Patients with Hypomyelinating Leukodystrophies." Annals of Human Genetics **70**(1): 66-77.

Veal, C. D., R. L. Clough, R. C. Barber, S. Mason, D. Tillman, B. Ferry, A. B. Jones, M. Ameen, N. Balendran, S. H. Powis, A. D. Burden, J. N. W. N. Barker and R. C. Trembath (2001). "Identification of a novel psoriasis susceptibility locus at 1p and evidence of epistasis between PSORS1 and candidate loci." Journal of Medical Genetics **38**(1): 7-13.

Visser, R., O. Shimokawa, N. Harada, A. Kinoshita, T. Ohta, N. Niikawa and N. Matsumoto (2005). "Identification of a 3.0-kb major recombination hotspot in patients with Sotos syndrome who carry a common 1.9-mb microdeletion." American Journal of Human Genetics **76**(1): 52-67.

Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard (2006). "A Map of Recent Positive Selection in the Human Genome." PLoS Biology **4**(3): e72.

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

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

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

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

Wat, M. J., O. A. Shchelochkov, A. M. Holder, A. M. Breman, A. Dagli, C. Bacino, F. Scaglia, R. T. Zori, S. W. Cheung, D. A. Scott and S.-H. L. Kang (2009). "Chromosome 8p23.1 deletions as a cause of complex congenital heart defects and diaphragmatic hernia." American Journal of Medical Genetics Part A **149A**(8): 1661-1677.

Weisenseel, P., B. Laumbacher, P. Besgen, D. Ludolph-Hauser, T. Herzinger, M. Roecken, R. Wank and J. C. Prinz (2002). "Streptococcal infection distinguishes different types of psoriasis." Journal of Medical Genetics **39**(10): 767-768.

White, S., M. Kalf, Q. Liu, M. Villerius, D. Engelsma, M. Kriek, E. Vollebregt, B. Bakker, G. J. Van Ommen, M. H. Breuning and J. T. Den Dunnen (2002). "Comprehensive Detection of Genomic Duplications and Deletions in the DMD Gene, by Use of Multiplex Amplifiable Probe Hybridization." American Journal of Human Genetics **71**(2): 365-374.

White, S. H., W. C. Wimley and M. E. Selsted (1995). "Structure, function, and membrane integration of defensins." Current Opinion in Structural Biology **5**(4): 521-527.

Wolf, N., M. Quaranta, N. J. Prescott, M. Allen, R. Smith, A. D. Burden, J. Worthington, C. E. M. Griffiths, C. G. Mathew, J. N. Barker, F. Capon and R. C. Trembath (2008). "Psoriasis is associated with pleiotropic susceptibility loci identified in type II diabetes and Crohn's disease." Journal of Medical Genetics **45**(2): 114-116.

Wong, K. K., R. J. deLeeuw, N. S. Dosanjh, L. R. Kimm, Z. Cheng, D. E. Horsman, C. MacAulay, R. T. Ng, C. J. Brown, E. E. Eichler and W. L. Lam (2007). "A Comprehensive Analysis of Common Copy-Number Variations in the Human Genome." American Journal of Human Genetics **80**(1): 91-104.

Xiao, Y., A. Hughes, J. Ando, Y. Matsuda, J.-F. Cheng, D. Skinner-Noble and G. Zhang (2004). "A genome-wide screen identifies a single beta-defensin gene cluster in the chicken: implications for the origin and evolution of mammalian defensins." BMC Genomics **5**(1): 56.

Xue, Y., D. Sun, A. Daly, F. Yang, X. Zhou, M. Zhao, N. Huang, T. Zerjal, C. Lee, N. P. Carter, M. E. Hurles and C. Tyler-Smith (2008). "Adaptive Evolution of UGT2B17 Copy-Number Variation." American Journal of Human Genetics **83**(3): 337-346.

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

Yang, D., A. Biragyn, L. W. Kwak and J. J. Oppenheim (2002). "Mammalian defensins in immunity: more than just microbicidal." Trends in Immunology **23**(6): 291-296.

Yang, D., Q. Chen, O. Chertov and J. J. Oppenheim (2000). "Human neutrophil defensins selectively chemoattract naive T and immature dendritic cells." Journal of Leukocyte Biology **68**(1): 9-14.

Yang, D., O. Chertov, S. N. Bykovskaia, Q. Chen, M. J. Buffo, J. Shogan, M. Anderson, J. M. Schroder, J. M. Wang, O. M. Howard and J. J. Oppenheim (1999). "Beta-defensins: linking innate and adaptive immunity through dendritic and T cell CCR6." Science **286**(5439): 525-528.

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

Yang, Y., E. K. Chung, Y. L. Wu, S. L. Savelli, H. N. Nagaraja, B. Zhou, M. Hebert, K. N. Jones, Y. Shu, K. Kitzmiller, C. A. Blanchong, K. L. McBride, G. C. Higgins, R. M. Rennebohm, R. R. Rice, K. V. Hackshaw, R. A. S. Roubey, J. M. Grossman, B. P. Tsao, D. J. Birmingham, B. H. Rovin, L. A. Hebert and C. Y. Yu (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.

Young, J. M., R. M. Endicott, S. S. Parghi, M. Walker, J. M. Kidd and B. J. Trask (2008). "Extensive Copy-Number Variation of the Human Olfactory Receptor Gene Family." American Journal of Human Genetics **83**(2): 228-242.

Zhang, F., C. M. B. Carvalho and J. R. Lupski (2009a). "Complex human chromosomal and genomic rearrangements." Trends in Genetics **25**(7): 298-307.

Zhang, F., M. Khajavi, A. M. Connolly, C. F. Towne, S. D. Batish and J. R. Lupski (2009b). "The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans." Nature Genetics **41**(7): 849-853.

Zhang, X.-J., P.-P. He, Z.-X. Wang, J. Zhang, Y.-B. Li, H.-Y. Wang, S.-C. Wei, S.-Y. Chen, S.-J. Xu, L. Jin, S. Yang and W. Huang (2002). "Evidence for a Major Psoriasis Susceptibility Locus at 6p21(PSORS1) and a Novel Candidate Region at 4q31 by Genome-wide Scan in Chinese Hans." Journal of Investigative Dermatology **119**(6): 1361-1366.

Zhang, X.-J., W. Huang, S. Yang, L.-D. Sun, F.-Y. Zhang, Q.-X. Zhu, F.-R. Zhang, C. Zhang, W.-H. Du, X.-M. Pu, H. Li, F.-L. Xiao, Z.-X. Wang, Y. Cui, F. Hao, J. Zheng, X.-Q. Yang, H. Cheng, C.-D. He, X.-M. Liu, L.-M. Xu, H.-F. Zheng, S.-M. Zhang, J.-Z. Zhang, H.-Y. Wang, Y.-L. Cheng, B.-H. Ji, Q.-Y. Fang, Y.-Z. Li, F.-S. Zhou, J.-W. Han, C. Quan, B. Chen, J.-L. Liu, D. Lin, L. Fan, A.-P. Zhang, S.-X. Liu, C.-J. Yang, P.-G. Wang, W.-M. Zhou, G.-S. Lin, W.-D. Wu, X. Fan, M. Gao, B.-Q. Yang, W.-S. Lu, Z. Zhang, K.-J. Zhu, S.-K. Shen, M. Li, X.-Y. Zhang, T.-T. Cao, W. Ren, X. Zhang, J. He, X.-F. Tang, S. Lu, J.-Q. Yang, L. Zhang, D.-N. Wang, F. Yuan, X.-Y. Yin, H.-J. Huang, H.-F. Wang, X.-Y. Lin and J.-J. Liu (2009c). "Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21." Nature Genetics **41**(2): 205-210.

Zhang, X.-J., K.-L. Yan, Z.-M. Wang, S. Yang, G.-L. Zhang, X. Fan, F.-L. Xiao, M. Gao, Y. Cui, P.-G. Wang, L.-d. Sun, K.-Y. Zhang, B. Wang, D.-Z. Wang, S.-J. Xu, W. Huang and J.-J. Liu (2007). "Polymorphisms in Interleukin-15 Gene on Chromosome 4q31.2 Are Associated with Psoriasis Vulgaris in Chinese Population." Journal of Investigative Dermatology **127**(11): 2544-2551.

Zhao, C., T. Nguyen, L. Liu, R. E. Sacco, K. A. Brogden and R. I. Lehrer (2001). "Gallinacin-3, an inducible epithelial beta-defensin in the chicken." Infection and Immunity **69**(4): 2684-2691.

Zheng, S. L., K. Augustsson-Bälter, B. Chang, M. Hedelin, L. Li, H.-O. Adami, J. Bensen, G. Li, J.-E. Johnasson, A. R. Turner, T. S. Adams, D. A. Meyers, W. B. Isaacs, J. Xu and H. Grönberg (2004). "Sequence Variants of Toll-Like Receptor 4 Are Associated with Prostate Cancer Risk." Cancer Research **64**(8): 2918-2922.