

University of Nottingham
UK | CHINA | MALAYSIA

ANALYSIS OF ALPHA DEFENSIN VARIATION IN SAMPLES FROM PERIODONTITIS PATIENTS

Eric Constantin Moise, BSc
Student ID Number: 4289572

THESIS SUBMITTED TO THE
UNIVERSITY OF NOTTINGHAM
FOR THE DEGREE OF MASTER BY RESEARCH
JANUARY 2018

Table of Contents

ACKNOWLEDGMENTS ---------- 4

ABSTRACT ---------- 5

1. INTRODUCTION ---------- 6

1.1 Innate immunity ---------- 6

1.2 Periodontal health ---------- 7

1.3 Host defence peptides ---------- 9

1.4 Defensins ---------- 9

1.5 Defensin mode of action ---------- 11

1.6 Human genetic variation ---------- 12

1.6.1 Types of genetic variation ---------- 12

1.7 Copy number variation in human defensin regions ----- 14

1.8 Copy number variant formation ---------- 18

1.9 Copy number measurement ---------- 22

1.10 Phenotypic consequences of CNV ---------- 25

1.11 Aims ---------- 27

2. MATERIALS AND METHODS----------28

2.1 DNA Samples----------28

2.2 PRT methods ----------28

2.2.1 Del4 and CEN1 Multiplex PCR ----------31

2.3 Allelic ratios ----------33

2.3.1 DefHae3 Assay----------33

2.3.2 Indel5 and 7bpdup Multiplex PCR ----------34

2.4 Electrophoresis----------36

2.5 Data analysis ----------37

2.5.1 Gene Mapper Analysis----------37

2.5.2 Defining Ratios ----------39

2.5.3 Maximum likelihood analysis. ----------41

3. Results ----------44

3.1 DEFA1A3 COPY NUMBERS ----------44

3.2 Quality control ----------47

3.3 PRT agreement----------49

3.4 Copy Number Assignment Quality. ----------52

3.5 Concordance between PRT measurements ----------60

3.6 CNV Association ---------- 62

4. DISCUSSION ---------- 69

4.1 Measurement of CNVs ---------- 69

4.2 CNV Association ---------- 71

5. BIBLIOGRAPHY ---------- 73

ACKNOWLEDGMENTS

Firstly, I would like to thank my supervisor, Prof John Armour, for giving me the opportunity to do my MRes in his research laboratory and for his supervising and endless support provided throughout my time as his student. I would also like to thank to all the members of JALA lab who helped me during my research. In particular I would like to thank Nzar Shwan for his advices throughout my work.

I would like to thank Reader Dan Cimponeriu, Assistant Prof. Alina Maria Holban and Scientific Researcher Consuel-Severus Ionică for their guidance and support during my bachelor's degree and the encouragement to pursue this MRes at the University of Nottingham.

Last, but not least, I would like to thank my mother, Violeta, for her unfailing support and continuous encouragement throughout the year of my study. This accomplishment would not have been possible without her.

ABSTRACT

Background

In recent years, many studies have found positive associations between copy number variants and predisposition to human complex disease. On chromosome 8p23.1, a copy number variation locus, named *DEFA1A3* is associated with susceptibility to IgA nephropathy, and is more generally thought to be involved in the host immune response in people with periodontitis.

Objective

The main objective of this study was to undertake *DEFA1A3* copy number measurement in 744 European samples with periodontitis and also to determine if there is an association between the copy numbers and periodontitis.

Results and Conclusions

In this study we have used the Paralogue Ratio Tests to type 741 DNA samples from individuals with periodontitis and assessed the accuracy in typing copy number. As a second part of our work, we have found that there is an association between copy number distribution and periodontitis, but further work needs to be done in order to fully confirm and explore this conclusion.

1. INTRODUCTION

1.1 Innate immunity

Innate immunity or non-specific immunity represents the first line of defence against different types of pathogens. The innate response system is prompt and consists of many non-specific defence mechanisms that are ready to be mobilized within minutes or hours after the infection takes place (Levy et al., 2003). It is these defence mechanisms such as physical barriers, secretory molecules and cellular components that can establish a state of immunity by preventing the physical movement of the pathogens from exterior or limit their ability to move and spread throughout the body. The innate immune system is active against a broader spectrum of pathogens like gram-positive and gram-negative bacteria and also against broader pathogens-associated molecular patterns like double stranded RNA. In comparison, the adaptive immune system response takes longer but is more accurate, as it cannot be stimulated by other than certain types of epitopes and antibodies.

In vertebrates, among other epithelial surfaces, the skin represents an anti-infective, protective barrier between the outside world and the inside of the body. The outermost layer of the skin, the epidermis, makes the skin surface resistant to degradation and creates a dry and acidic environment, where the growth of most harmful bacteria and fungi is inhibited. Furthermore, the mucous membrane represents another physical barrier against potential

pathogens. The inner surface of the body is covered and protected by a mucus layer, secreted by epithelial cells, which traps the microbes. In addition, there are a few other specific defence barriers such as salivary glands and the gastrointestinal tract that contribute to the homeostasis of the organism.

Occasionally, pathogens breach physical barriers and enter the body, and then it is up to the second line of defence of innate immunity to isolate and destroy or inactivate the pathogen. To do so, cells of the innate immune system rely on their pattern recognition receptors (PRR) to discriminate between self and non-self. As soon as an invasion occurs, the complement system initiates an inflammatory response that attracts neutrophils or macrophages to the infection site where they begin phagocytosis. If the innate immune system is not capable of destroying the pathogens, after a certain period, the specific adaptive response sets in (Elliott et al., 2014).

1.2 Periodontal health

Epithelial tissues are physical barriers that form boundaries between different environments and among other functions provide protection against microbial infections (Dolińska et al., 2017). The periodontal epithelium is a stratified squamous epithelium tight around the tooth that plays an important role in oral cavity hygiene and homeostasis. It is the only place in the body in which a hard tissue passes the epithelium and connects the inside of the body with the outside environment (Dale and Fredericks, 2005). This particular

characteristic leads to a certain vulnerability towards infections, especially towards periodontal disease.

The periodontal tissue is constantly exposed to pathogenic biofilms that form at the junction between the tooth and soft tissue. Long exposure to subgingival plaque bacteria leads to inflammatory response (Page, 1991). When inflammation is limited to gums the condition is called gingivitis, and if the inflammation progresses and includes the bone then periodontitis arises. Periodontal disease can be described as a chronic inflammatory condition of teeth-supporting tissues, such as gums and bones. The severe form of the disease, called periodontitis, can lead to tooth loss (Page, 1986; Page, 1991).

Recent studies have shown that a healthy homeostasis of the oral cavity is maintained not only through physical barriers but also through chemical barriers such as antimicrobial peptides (AMPs), proinflammatory cytokines and chemokines (McCormick and Weinberg, 2010). Proinflammatory cytokine and chemokine secretions allow the infiltration of macrophages and neutrophils to the infection site as an inflammatory response (Preshaw and Taylor, 2011). There are several types of antimicrobial peptides present in the oral cavity including defensins and histatins with only three major sites of AMP secretion: gingival epithelium, neutrophils and salivary glands (Chung et al., 2007).

1.3 Host defence peptides

Antimicrobial peptides (AMP) are oligopeptides that help the innate immune system to maintain the balance between healthy and infected states. Found in both eukaryotes and prokaryotes, AMP's have a similar role, to fight against pathogens. Being part of the first line of defence, AMPs have a broad spectrum of antimicrobial activity ranging from viruses to cancer cells (Bals, 2000; Zhang and Gallo, 2016). In addition, in the case of periodontal disease, they have been found to be involved in the host's inflammatory response. Uncontrolled inflammation caused by excessive production of AMP is considered the leading mechanism for the pathogenesis of periodontitis (Cekici et al., 2014).

Base on their secondary structure, AMPs can be categorised in 4 groups: α-helix, ß-sheet, loop and extended. Among these groups, defensins and cathelicidins are the most abundant and well-studied human AMPs (Lehrer and Ganz, 2002).

1.4 Defensins

Defensins are small antimicrobial peptides, composed of 29-34 amino acids, with a triple stranded β-sheet structure (Ganz, 1999). Found in high concentration and widely distributed around the body, they have multiple antimicrobial functions in both innate and adaptive immunity, fighting against Gram-positive and negative bacteria as well as fungi, yeast and viruses (de Smet and Contreras, 2005). These are divided into three classes: α-defensins, ß-defensins and theta-

defensins that differ not only in length but also in the arrangement of their cysteine intramolecular disulphide bridges. Alpha defensins follow a 1-6, 2-4, 3-5 cysteine link pattern (Selsted and Harwig, 1989), beta defensins are 1-5, 2-4, 3-6 (Tang and Selsted, 1993), while theta defensins' disulphide pattern is formed from peptide bonds between two hemi-α-defensins (Tang et al., 1999).

Beta defensins are mainly expressed in airways, epididymis epithelia and leucocytes and they present broad antimicrobial properties that are induced upon pathogen invasion. Beta defensins genes form a major cluster on 8p23.1 (Linzmeier et al., 1999); clusters of presumed ß-defensin genes have also been identified on 6p,20p and 20q (Schutte et al., 2002).

Alpha defensins are divided in two subclasses. The first subclass, known as HNP 1 to 4 (Human neutrophil peptides 1 to 4) is synthesized in promyelocytes and stored in neutrophils whereas the other subclass known as HD 5 and 6 (Human defensin 5 and 6) is synthesized and stored in Paneth cells. In the beta defensin classes, there are four well-characterised human ß-defensins (hBD 1-4) expressed within epithelial cells (de Smet and Contreras, 2005).

In the oral cavity, defensins are present in salivary glands, salivary secretions and epithelial cells. Both alpha defensins and beta defensins are strategically located, ready to trigger the immune response. More specifically alpha defensins HNP1 to 3 are found in the junctional epithelium, while beta defensins hBD1, hBD2 and hBD3 are

expressed in the oral and sulcular stratified epithelium (Greer et al., 2013).

Cathelicidins are a family of antimicrobial peptides, mainly found in neutrophils and macrophages. They differ in size, structure and in amino acid sequences, but a common pro-region is present at their N-terminus known as the cathelin domain. Substantial heterogeneity is shown at their C-terminus where the mature antimicrobial peptide, LL-37, is cleaved and released when required (Yang et al., 2004). In oral cavity, the LL-37 peptide is expressed in tongue, and buccal mucosa and is found in saliva. LL-37 concentration levels are upregulated in the case of chronic periodontitis and it has been found that low levels of LL-37 are associated with periodontal disease. Furthermore, in gingival crevicular fluids (GCF), elevated levels of LL-37 have been found in patients with periodontitis in comparison with healthy individuals (Khurshid et al., 2017). Therefore, LL-37 can be used as diagnostic tool in the identification of periodontal diseases.

1.5 Defensin mode of action

Firstly, defensins main ability is to create voltage dependent ion channels that permeabilise the bacterial cellular membrane, thus allowing other peptides to invade, ultimately resulting in cell death (Rice et al., 1987; Wimley et al., 1994). Because of their electrostatic charge they are highly effective against negative charged membranes, allowing them to degrade both Gram positive and Gram-negative

bacteria as well as fungi and enveloped viruses, though with the exception of non-enveloped viruses (Fujii et al., 1993; Ericksen et al., 2005, Ganz et al., 1985).

Secondly, defensins are known to regulate the complement system thus activating the host immune response. They have been shown to induce migration of macrophages, mast cells, naïve dendritic cells and naïve T-cells which activates the immune response through various processes, such as inflammation and wound healing (Yang et al., 2000, Hazlett and Wu, 2011).

1.6 Human genetic variation

Genetic diversity is defined as variation within the genome. Thus, variants such as single nucleotide polymorphism (SNP), insertion-deletion polymorphism (Indels) or copy number variation (CNV) have been associated with genetic and phenotypic variability including diversity among populations, individuals, and susceptibility to common and rare disease. Many studies including Genome-Wide Association Studies (GWASs) and The 1000 Genomes Project have turned their attention to the identification and characterisation of variants within human genomes, and have provided a dataset of informative variants.

1.6.1 Types of genetic variation

Single nucleotide polymorphisms are the most common and well-characterised form of variation. Widely spread across the human genome, SNPs can be described as a substitution of one single

nucleotide with another alternative nucleotide. They are referred as SNPs only if the allele frequency of the SNP is greater than 1% otherwise they are referred as single nucleotide variants (SNV) (The International HapMap Project). The location of a SNP in the genome usually dictates the effect it has on phenotype and contributes to their classification. On the assumption that a SNP is found within non-coding regions, the effect on phenotype is usually minimal or there is no effect. When present within a coding region, the mutation caused can be nonsynonymous if the amino acid has been altered and it can result in a missense mutation or a nonsense mutation (1000 Genome Project). Alternatively, the mutation can be synonymous if the amino acid structure has not been changed, and it does not affect the protein structure (1000 Genome Project).

Insertions (IN) and deletions (DEL) or indels represent the second most common form of genetic variations in human genomes. Dispersed throughout the genome, this is a type of genetic variation characterised by the presence (insertion) or absence (deletion) of different nucleotide sequences (Mullaney et al., 2010). With a variable length (between 1 and up to 10,000 base pair (bp)), indels can have a far more significant effect on gene expression if found within a coding region than SNPs (Mullaney et al., 2010). Assuming that the number of added or deleted nucleotides is not a multiple of three, indels often lead to severe genetic mutation known as a frameshift mutation (Feuk et al., 2006). If the length exceeds 1 kb, indels are classified as a copy number variation (CNV) and are

commonly referred to as duplication, deletion or amplification (Stankiewicz et al., 2010).

1.7 Copy number variation in human defensin regions

DNA copy number variation (CNV) represents a well-established source of genetic diversity between individuals. The DNA fragment involved in a CNV polymorphism is usually longer than or equal to 1 kilobase (kb) and is present in a variable copy number. Due to this increased dimension, CNV region can encompass entire genes, their regulatory regions, disease loci and segmental duplications (Feuk et al., 2006). There are several ways in which CNV can be classified. The first and most common method of classification refers to the changes made in the DNA structure. Therefore CNV, which are also known as copy number polymorphisms (CNPs) can be present as insertion, deletion, duplication and multi-site variants (Redon et al., 2006) (Figure 1). Together, SNPs, CNVs and other structural genomic variants contribute to phenotypic variation such as rhesus blood group variation and can take part in genomic disorders such as Charcot-Marie Tooth disease type 1A (Lupski, 2009).

Copy number variants can also be classified according to the number of alleles underlying variation in the number of copies. Simple copy number variants have just two alleles. If a deletion occurs only within one allele, then the region in question can be present from 0 to 2 copies per diploid genome, and if an allele carries a duplication,

then the region in question can be present in 2-4 copies. However, multiallelic copy number variants encompass regions with highly variable number of copies (Black et al., 2014).

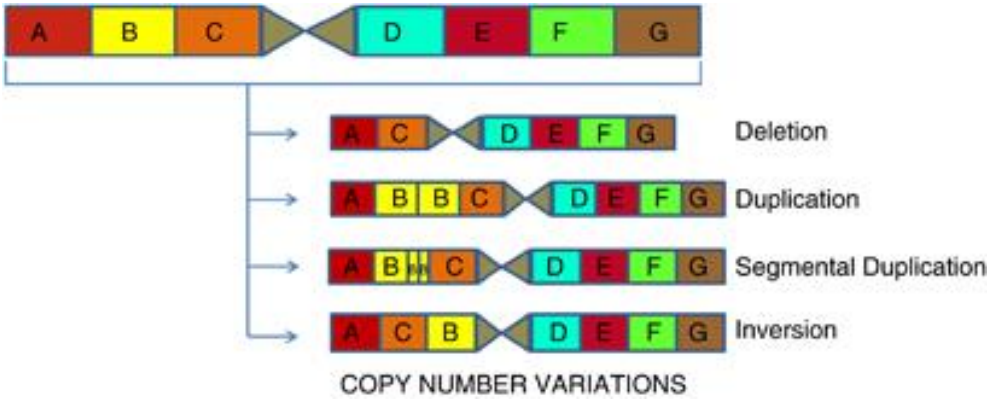


Figure 1 - Different types of genetic variation. There are 4 primarily classes of copy number variants: deletion, duplication, segmental duplication and inversion (adaptation after Almal and Padh. 2012).

The alpha defensin gene cluster is found within chromosome 8p23.1. There are five different α-defensin genes present in humans: *DEFA1*, *DEFA3*, *DEFA4*, *DEFA5* and *DEFA6* whereas *DEFA3* is a human specific gene (Aldred, 2004; Khan et al., 2013; Linzmeier and Ganz, 2005). *DEFA1* and *DEFA3* genes are copy number variable, with a 19-kb sequence unit in the alpha defensin locus known as *DEFA1A3* and they only differ by a single base substitution. One single copy of the gene is always found within the centromeric partial repeat region, whereas a variable number of copies are found in 19-kb full repeat regions (Figure 2). *DEFA1A3* copy number ranges from 3 to 16 copies per diploid genome, with a median of 6 found in European populations (Khan et al., 2013). More importantly, *DEFA3* gene is absent in between 10-37% of people, depending on the population. The configuration of the *DEFA1A3* locus is shown in Figure 2 (Aldred,

2004). In addition, *DEFA1*, *DEFA3 and DEFA4* encode four human neutrophil peptides, HNP-1, HNP-2, HNP-3 and HNP-4. DEFA4 produces the minor neutrophil peptide HNP-4. HNP-1 and HNP-3 amino acid structure differs only in their N-terminal amino acid while HNP-2 lacks the N-terminal amino acid that distinguishes HNP-1 from HNP-3. Since there are no separate genes coding for HNP-2 it is assumed that one or both of *DEFA1* and *DEFA3* must encode the peptide, with subsequent removal of the N-terminal amino acid during proteolytic processing (Linzmeier and Ganz, 2005). HNP1-3 are synthesised in promyelocytes and then stored in the primary azurophilic granules of neutrophils. DEFA5 and DEFA6 encode the prodefensins HD5 and HD6 and are stored in the Paneth cells of the ileum (Cunliffe et al, 2001).

It is still unclear if there is a positive relationship between *DEFA1A3* copy number and HNP1-3 expression. Whilst large scale studies have been, so far, unable to prove that the level of HNP1-3 expression is directly proportional to the *DEFA1A3* copy number (Ai et al., 2016), one small-scale study showed a relationship with neutrophil peptide content (Linzmeier and Ganz., 2005).

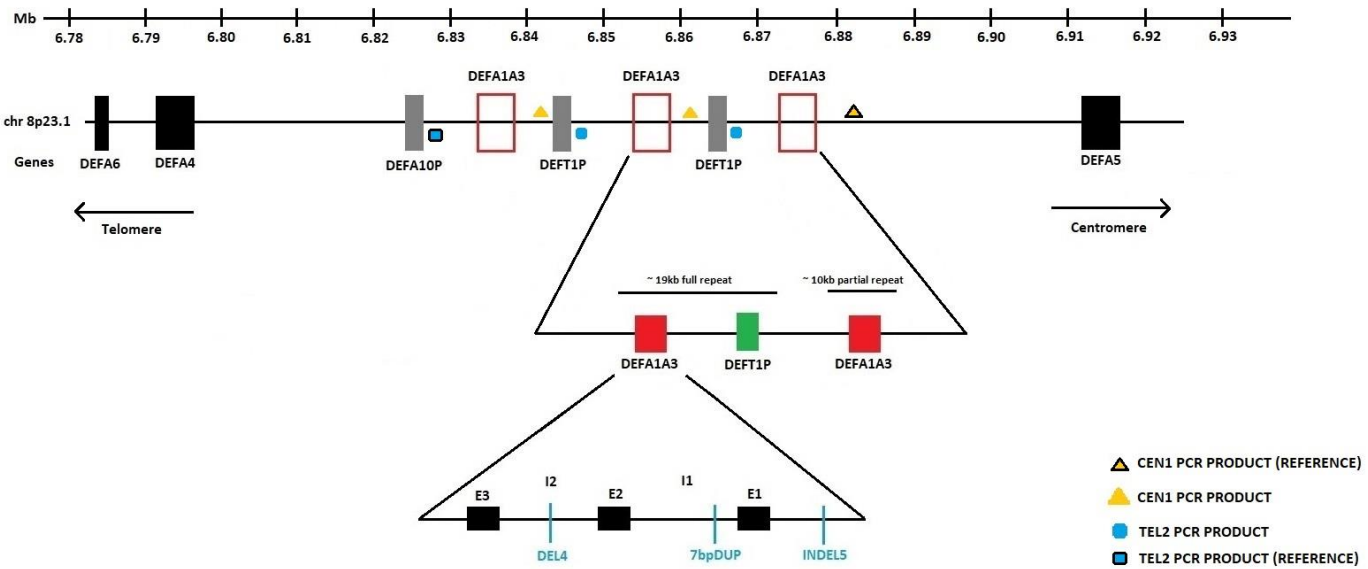


Figure 2 - Schematic section of DEFA1A3 CNV on chromosome 8. One single copy partial repeat of 10kb is shown, flanking two variable regions of 19kb full repeats. The locus is referred to as DEFA1A3 because both DEFA1 and DEFA3 can occupy the full repeat and the centromeric partial repeat. For each full repeat unit there is a copy of the DEFT1P pseudogene. The DEFA1A3 locus comprises 3 exons (E1, E2 and E3) and 2 introns (I1 and I2). The precise location of the measured variants used in this study (Del4, 7bpDUP, INDEL5) are shown in the figure. Both CEN1 and TEL2 PCR products are repetitive sequences found in each copy of either DEFA1 or DEFA3. By comparing the PCR products with the reference products derived from the end copies, we can estimate the total copy number. Adapted figure after the Feb. 2009 genome assembly (GRCh37/hg19).

1.8 Copy number variant formation

Three major mechanism are known to be involved in the formation of copy number variants in the human genome: NAHR – non-allelic homologous recombination, NHEJ – non-homologous end-joining and FoSTeS – Fork Stalling and Template Switching models. Usually, a non-recurrent mechanism is involved in the formation of simple copy (diallelic) number variants, whereas through recurrent mechanisms, multiallelic copy number variants arise.

NAHR is the main cause of recurrent rearrangements between LCRs (low copy repeats) and is thought to be involved in approximately 10% to 22% of all human genomic rearrangements (Matthew et al., 1999). NAHR is also involved in chromosome structural repair, however unlike NHEJ, the repair process is mediated by homologous templates. The allelic homologous recombination (AHR) is classified as another pathway used in the repair process of double stranded breaks (DSBs). In this pathway, the AHR mechanisms will use as a template an allele found in the sister chromatid at the same region where DSB occurs (Gu et al., 2008). Usually, if the DSB is present within a unique region then finding the sister allele will not be problematic, whereas if found within a copy variable region, the recombination process might use a paralog of the repeat. Therefore, by selection a non-allelic region, NAHR can occur. Orientation and location of the repeats can influence the rearrangement type. If the repeats are tandemly repeated, then deletion or duplications can occur and if they are found in an inverted orientation this can lead to

inversions in the repaired sequence. Consequently, NAHR can lead to recurrent changes in the copy number sequence (Gu et al., 2008) (Figure 3).

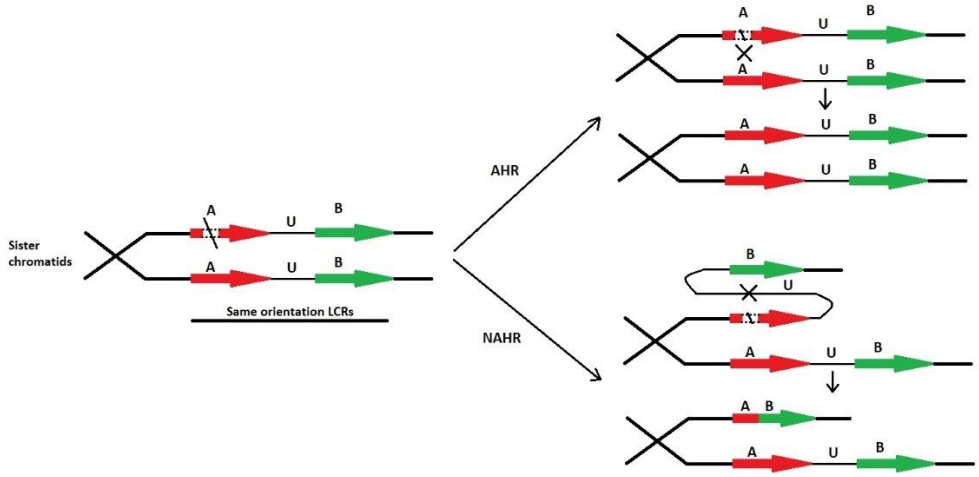


Figure 3 – Example of homologous recombination. The two thin lines represent the sister chromatids joined at the centromere. Red (A) and green (B) arrows are homologous low copy repeats (LCRs) and the unique segment between the two homologous regions is labelled U. In this schematic representation, both LCRs have the same orientation. The location of DSB is marked with an oblique line through the red LCR. The allelic homologous recombination (AHR) process is using the allelic region on the sister chromatid to perfectly repair the red LCR region where DSB is present. The NAHR occurs when a non-allelic rather than allelic position is used as a templated to repair the DSB (green LCR). This leads to the creation of a hybrid LCR (red and green). If the LCRs have the same orientation (arrows point in the same direction, the subsequent recombination leads to deletions or/and duplications. When the LCRs have opposite orientation, the repair process results in an inversion.

NHEJ is an important mechanism used in both physiological and pathological double strand break (DSB) repair process. There are 4 stages involved in the NHEJ repair process as shown in Figure 4: detection of DSB, molecular interaction between the two DNA ends, nucleotide rearrangement at the site of both ends and the ligation process (Gu et al., 2008). Because the repair process is not based on the presence of a homologous template, variation usually occurs within the re-joining stage, where nucleotides are added to/or deleted from or both ends. Therefore, NHEJ can lead to genomic

rearrangements such as small deletions and duplications. Alternatively, microhomology-mediated end joining (MMEJ) has been known to be involved in DSB repair but this process always leads to deletions between double stranded regions (Gu et al., 2008; Hastings et al., 2009) (Figure 4).

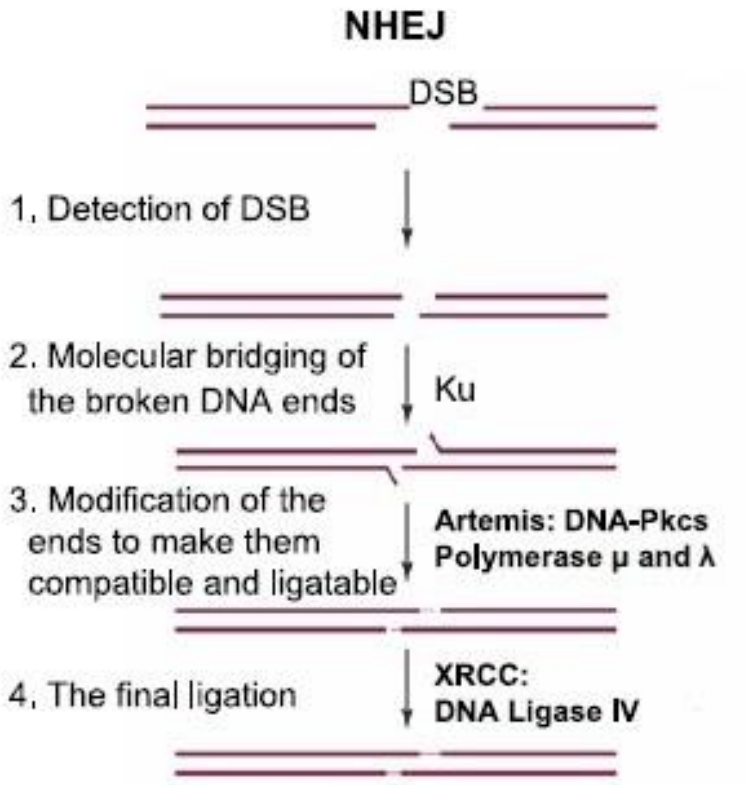


Figure 4 - Genomic rearrangement mechanisms adapted from Gu et al., 2008. Non-homologous end-joining formation process in vertebrates.

Fork Stalling and Template Switching (FoStEs) – represents another way of copy number variant formation. During DNA replication, the DNA molecule duplicates to generate two identical copies. Sometimes, the activity of the replication fork can be interrupted leading to the disengagement of the lagging strand from the initial template. After this stage, the lagging strand attaches itself

to another replication fork through regions of complementary sequences called microhomologies and the DNA synthesis restarts (Zhang et al., 2009). Depending on the priming orientation of the new fork (direct or inverted) and location (upstream or downstream) this can result in deletion, duplication and inversion, as well as complex rearrangements resulting from multiple nested rounds of FoSTeS (Zhang et al., 2009) (Figure 5).

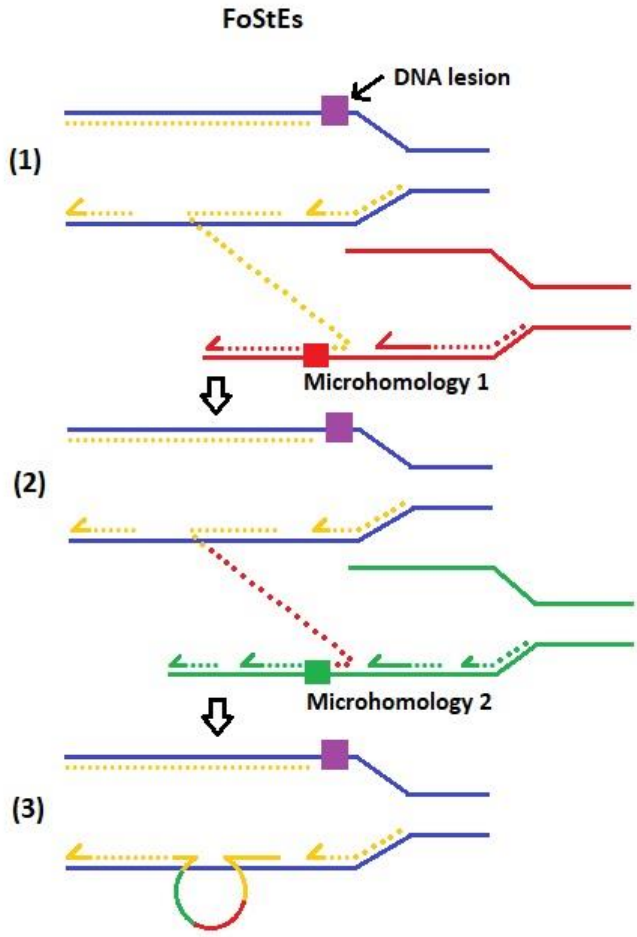


Figure 5 - Fork stalling and template switching schematic representation. (1) After the replication fork (dark blue, solid line) stalls at a DNA lesion (purple square), the lagging strand (yellow, dotted line) disengages and anneals to replication fork (red, dotted line) through microhomology 1, which primes DNA synthesis. (2) The lagging strand disengages once more and invades a second fork (green, dotted line) where it anneals to another microhomologous region (green square) to restart synthesis. The invasion can occur several times before the lagging strand returns to the original replication fork (3).

1.9 Copy number measurement

Accurate measurement of copy number variants has proved to be challenging, especially in regions with high copy numbers (Khan et al., 2013). Several methods have been used to measure copy numbers at different multiallelic loci, each presenting their own advantages and disadvantages. Initially, the most commonly used techniques were fluorescence *in situ* hybridisation (FISH) and Southern Blotting but they were time consuming and FISH lacked the capacity of detecting small fragments (Aldred et al., 2005; Stuppia et al., 2012). Alternatively, comparative genome hybridisation (CGH) has been used in detection of copy numbers. CGH works by comparing a reference sample, usually a healthy control, with a test samples on a microarray (Carter et al., 2007). The downside of using CGH is that it can be prone to CNV measurement error for test samples that have copy numbers close to the reference ones (Aldred, 2004).

Multiplex Amplifiable Probe Hybridisation (MAPH) is a reliable technique used for measuring copy number in genomic DNA. The MAPH technique has been used by Aldred et al in 2005 to measure copy number variants at *DEFA1A3* locus. MAPH is based on the quantification and amplification of specifically bound probes (Armour et al., 2000). Ranging between 100 and 600 bp, MAPH probes are designed to recognise a unique region of the newly denatured DNA and to be amplified by a single primer pair making it a cost-effective technique. The downside of using the MAPH method is that it can only detect small copy number fragments, needs a large DNA sample and

because is a gel-based technique only a certain number of probes can be analysed (Patsalis et al., 2007).

High-throughput methods such as Multiplex Ligation-dependent Probe Amplification (MLPA), Real-Time PCR and Paralogue Ratio Test (PRT) have also been used to study the structure and phenotypic consequences of CNV at different genomic regions (Stuppia et al., 2012, Armour et al., 2000). Real-Time PCR is probably the most common and easy to perform high-throughput method. The disadvantage of using Real-Time PCR is that of losing the accuracy of measuring higher copy number (Stuppia et al., 2012).

MLPA is a PCR-based method that relies on the ligation of two adjacently annealed oligonucleotides to a target sequence. This method has been most extensively used to screen DNA samples for specific copy number variation (Stuppia et al., 2012). Using MLPA offers numerous advantages in comparison with other techniques as it can be used for up to 50 DNA targets (all in one PCR-based reaction). In a subsequent PCR only the ligated oligonucleotides can serve as template (it does not need a purification step in order to remove unbound oligonucleotides), all probes have similar end sequences (a single primer pair is enough to perform the PCR) and can distinguish between sequences that differ by one nucleotide (Stuppia et al., 2012). Even if, out of the methods described above, the MLPA method is expected to achieve a high level of accuracy in determining the correct copy number measurements even for samples with higher copy

numbers (Khan et al., 2013; Stuppia el al., 2012). for a larger set of samples, methods involving Paralogue Ratio Test (PRTs) have been proven to be more successful than any other methods described, especially for more challenging CNV regions (Table 1) (Armour et al., 2000; Khan et al., 2013).

Method	Applicable to all CNVs	High-throughput	Accuracy
PRT	No	Yes	High
aCGH	No	No	Medium
FISH	No	No	High
MAPH	Yes	No	High
MLPA	No	Yes	High
RT-PCR	Yes	Yes	Low

Table 1 – Different feature of various CNV measuring methods.

In brief, the PRT method involves a single pair of primers in the amplification of two different genomic regions: one is a single copy (non-variable) region known as the reference locus and the other one is the region of interest. Then the copy number of the target locus is estimated from the ratio of test locus to reference locus. There are many more advantages of using a PRT method instead of any other method. For example, it is less expensive, uses only a small amount of DNA and it can be used for a large set of samples (Armour et al., 2000).

1.10 Phenotypic consequences of CNV

Gene expression is known to be easily affected by copy number variation; depending on where they occur, some CNVs can have a wide range of effects on human physiological functions (Stranger et al., 2007).

The expectation is that gene copy number will be directly proportional to gene expression, but this is not always the case. In 2007, Stranger and colleagues, showed that the gene copy number and gene expression can be negatively correlated in lymphoblastoid cell lines. Although the association between copy number variants and disease development needs further confirmations, few recent studies have validated this observation.

Copy number variable regions are believed to be involved in some human complex diseases. In recent years, copy number variants have been associated with numerous sporadic and Mendelian diseases. In 1992, Lupski and colleagues have found a correlation between *PMP22* gene duplication and Charcot-Marie-Tooth Type 1A disease. Similar to *PMP22* is *RAI1,* a dosage-sensitive gene which causes Potocki-Lupski syndrome when duplicated and Smith-Magenis syndrome when deleted (Stankiewicz and Lupski, 2010). Inaccuracy of copy number measurement studies led to a dispute concerning the association of *CCL3L1* gene and increased susceptibility to AIDS (low copy number) or to rheumatoid arthritis (high copy number) (Gonzalez

et al., 2005; McKinney et al., 2008; Field et al., 2009; Aklillu et al., 2013; Carpenter et al., 2011).

A study conducted by Ai and colleagues in 2016, confirmed a strong correlation between low copy number of *DEFA1A3* and IgAN /renal disfunction in IgA nephropathy patients of Chinese and Caucasian origin. Other studies have come to the same conclusion on the basis of SNP data that the *DEFA1A3* locus is involved in IgA nephropathy and renal disease progression (Yu et al., 2011; Li et al., 2015; Qi et al., 2016).

Over the years, many studies failed to combine accurate copy number measurements and changes in gene expression in their search for associations between complex human disease and CNV loci, resulting in the few reported reproducible associations.

1.11 Aims

In this study, the main aim was to type and analyse a set of 744 European DNA samples, with different stages of periodontitis, in order to identify their *DEFA1A3* copy numbers, and their compositions in terms of DEFA1 and DEFA 3 repeat unit variants, using a newly established method of CNV measurement involving fluorescent PCR measurement of allelic ratios. The samples were used in a recent study that established an association of periodontitis with SNPs in the *DEFA1A3* (Munz et al., 2017); typing these samples for *DEFA1A3* copy number would allow us to test whether there was also an association of the CNV with periodontitis. Because this precise format of typing DEFA1A3 CNV has not been used on a large scale before, we particularly wanted to test the new PRT format and to evaluate the usefulness of this typing methodology and its accuracy. Providing a robust method for typing copy numbers, will increase our chances of understanding CNV functional consequences.

2. MATERIALS AND METHODS

2.1 DNA Samples

744 DNA samples from unrelated patients of European origin, with different stages of periodontitis were used for copy number measurement. The sample were obtained from Prof. Arne Shäfer (Charité Hospital, Berlin).

2.2 PRT methods

In this study, two new duplex paralogue ratio tests combined with four established allelic ratio measurements were used to establish the copy number of *DEFA1A3*, as well as the individual copy number of *DEFA1* and *DEFA3* variants (and other repeat variants) for each sample. Before this project started, these methods had already been validated on CEPH samples with known copy numbers, that have been subject to previous PRT based *DEFA1A3* typing and 1000 Genomes Project read depth measurement of copy numbers (John Armour, unpublished data). For each PCR, 0.5 µl (25ng) of genomic DNA was added to 9.5 µl of a master mix with the following components: 0.1 µl of 0.5 U *Taq* DNA polymerase (NEB), 7.4 µl of H_2O, 0.5 µl of each of the 10 µM Primers and 1 µl of 10x LD Mix (1xLD mix has a final concentration of 50 mM Tris-HCl pH8.8, 12.5mM $(NH_4)_2SO_4$, 125 µg/ml BSA, 200 µM of each dNTP, 1.4mM $MgCl_2$, 7.5mM 2-mercaptoethanol and H_2O (Khan et al. 2013).

The two PRT systems were designed to measure the same *DEFA1A3* copy number using reference copies at either centromeric

(CEN1) or telomeric (TEL2) regions, with this dual design expected to confer a greater accuracy in the designation of a copy number. This depended on the telomeric or centromeric reference copy having a characteristic, reliable product size of 260bp (TEL2) and 444bp (CEN1) and all other copies having a size of 263bp (TEL2) and 333bp (CEN1) (Figure 6). Previous work performed by Khan and colleagues in 2013 (Khan et al., 2013) allowed us to validate the use of the 260bp product from the telomeric copy and 444bp from the centromeric copy as a reference for 2 copies. Both TEL2 and CEN1 PRTs were carried out in duplicate using two similar PCR assays, one with a HEX labelled forward primer and the other one with a FAM labelled

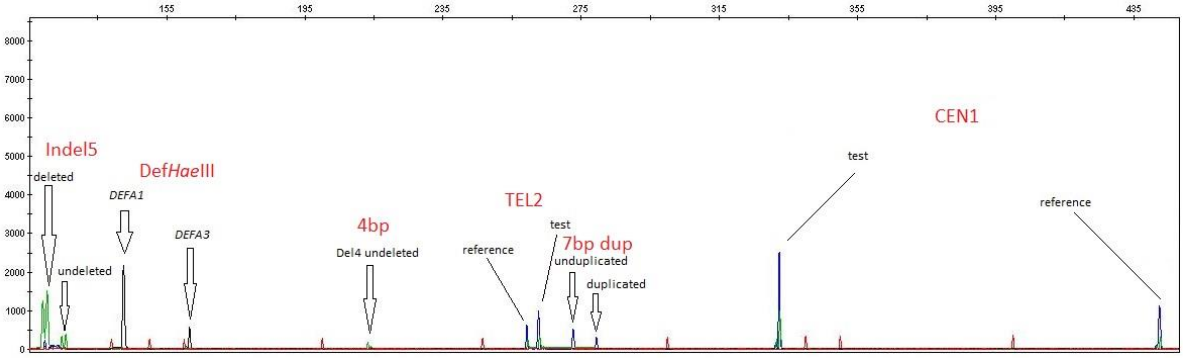


Figure 6 - GeneMapper profile of a random sample with a total copy number of 7. This profile includes all the PRT and allelic ratio measurements. Both CEN1 and TEL2 ratios show a ratio of test to reference of about 2.5 (5:2). The indel5 ratio is the ratio of deleted to undeleted (5:2). The DefHae3 ratio is the ratio of DEFA1 to DEFA3 (6:1). The 4bp ratio is the ratio of deleted (0) to undeleted (7). The 7bpdup ratio is the ratio of unduplicated to duplicated (5:2)

forward primer, using the average to determine a single unrounded estimated of the copy number. By using two separate dyes for each PRT, we were able to analyse both PCR products for each sample in a single electrophoretic capillary, thus reducing the cost and time.

At the telomeric end, the TEL2 PCR products were amplified in two parallel reactions with 1 µM FAM labelled forward primer FAM-

DEFAtel2F (5'-AGAGCAGCCGTGCACAAAC-3') or HEX labelled forward primer HEX-DEFAtel2F (5'-AGAGCAGCCGTGCACAAAC-3'), both using the same DEFAtel2R reverse primer (5'-GCATCTYGGGGTCCATTGT3'), for 25 cycles of 95 ºC for 30 seconds, 66 ºC for 30 seconds, 63 ºC for 2 minutes, followed by a 72 ºC hold for 40 minutes. Subsequently, the TEL2 PRT primers delivers two products: one of 260bp for the two reference copies in the telomeric position and 263bp for *DEFA1A3* copies in other parts of the CNV array (Figure 7). For example, we would predict that in a 6-copy sample there would be exactly two telomeric (260bp) copies and four non-telomeric(263bp) copies, giving a ratio of 4:2 (263bp:260bp).

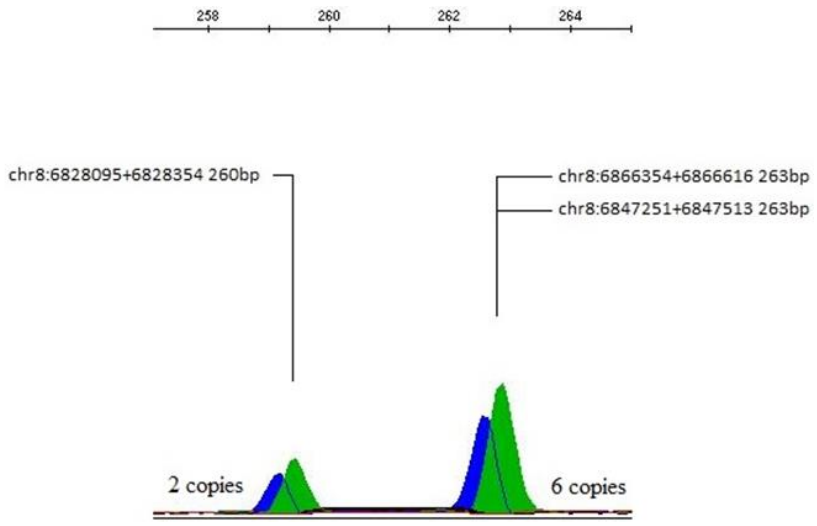


Figure 7 - 4 Predicted PCR products from TEL2 primers and actual peaks obtained from capillary electrophoresis. This is the result of a duplicate test in both FAM (blue) and HEX (green) and both measurements agree with each other. This individual was determined to have 8 copies of DEFA1A3. We can therefore interpret this profile as being 2 telomeric copies and 6 non-telomeric copies. Genomic coordinates are given according to GRCh37/hg19 Feb. 2009.

At the centromeric end, products were amplified in duplicate reactions using 1 µM FAM labelled forward primer DEEFAcen1F (5'-TCCTAGAAAGCTGGTTGCTC-3') or HEX labelled forward primer HEX-

DEFAcen1F (5'-CCCAGAGAGCTCCTTCATT-3') with DEFAcen1R primer (5'-TCCTAGAAAGCTGGTTGCTC-3'), for 25 cycles of 95 ºC for 30 seconds, 58ºC for 30 seconds, 68ºC for 2 minutes, followed by a 4 minutes final extension at 72 ºC. The CEN1 PRT giving products of 332bp for non-centromeric copies of *DEFA1A3* and 444bp for the two reference centromeric copies (Figure 8). For example, a 7-copy sample would have 5 non-centromeric copies (332bp) and two centromeric copies (444bp), giving a ratio of 5:2 (332bp:444bp).

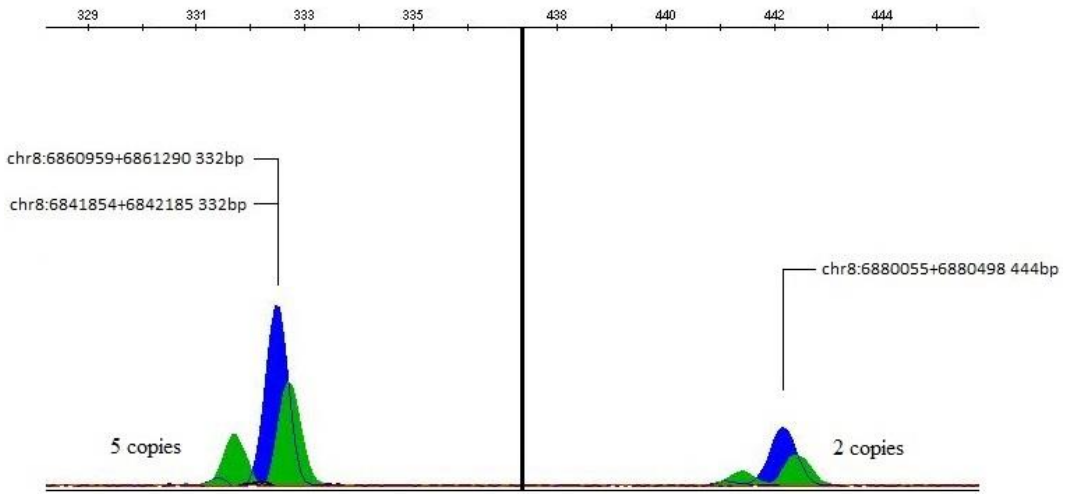


Figure 8 - 4 Predicted PCR products from CEN1 primers and actual peaks obtained from capillary electrophoresis. This is the result of a duplicate test in both FAM (blue) and HEX (green) and both measurements agree with each other. In the case of Hex-labelled products, each product is represented as a double peak, differing by one base pair, presumably due to incomplete dA addition of the 3' end of the product. This individual was determined to have 7 copies of DEFA1A3. We can therefore interpret this profile as being 2 centromeric copies and 5 non-centromeric copies. Genomic coordinates are given according to GRCh37/hg19 Feb. 2009.

2.2.1 Del4 and CEN1 Multiplex PCR

In order to cut down on the numbers of PCRs, we designed a multiplex PCR involving one of the allele ratio assay (Del4) and one of the new PRTs described above - CEN1. Del4 is a 4bp deletion variant

(TATC) found within the second intron of DEFA1A3 which is believed to have a significant role in the HNP1-3 protein expression (Ai et al., 2016). The Indel4/CEN1 assay was performed following the same PCR mix as described in the PRT assays and by using 1 μM of each primer and 25 cycles at 95ºC for 30 seconds, 58ºC for 30 seconds, 68ºC for 2 minutes followed by a 40 minutes final extension at 72ºC to generate amplicons of 211bp and 215bp for Del4 and 333bp/444bp amplicons for CEN1 (Figure 9). For the primer sequences used in the 4bp assays see Ai et al., 2016, and the CEN1Hex primers were as described above. While both Indel4 and Cen1 forward primers were Hex labelled, it did not interfere with their separation after capillary electrophoresis, as differ by more than 100bp in length.

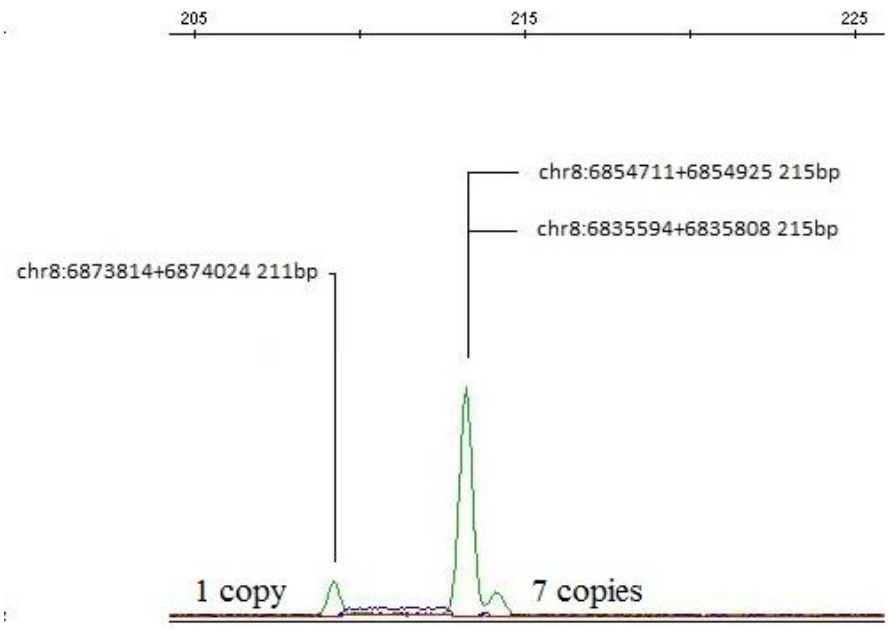


Figure 9 - Predicted PCR products from CEN1-4bp Multiplex PCR primers and actual peaks obtained from capillary electrophoresis. This is an 8-copy individual and we interpret the profile as being 1 undeleted copy and 7 deleted copies. Genomic coordinates are given according to GRCh37/hg19 Feb.2009.

2.3 Allelic ratios

In addition to the PRT method to determine total *DEFA1A3* gene copy number, four allelic-ratio measurements were used to determine the composition of different repeat variants, as well as contribute to measurement of the total copy number at *DEFA1A3* locus: DefHae3 Assay, 4bp, 7bp and 5bp as described by Khan, 2012 and Ai et al., 2016.

2.3.1 DefHae3 Assay

Described by both Aldred in 2005 and Khan in 2013, DefHae3 assay uses the presence or absence of a site for *Hae*III enzyme to distinguish between *DEFA1* and *DEFA3* variants as DEFA1 (GCC) varies from DEFA3 (GAC) by a single base pair change. Therefore, after the enzyme digestion step is completed, the results are two different size products, thus allowing us to determine the ratio between *DEFA1* and *DEFA3* after separation by capillary electrophoresis and quantification by fluorescence detection. To act as a reporter of complete restriction digestion by *Hae*III, the forward primer contains a mismatch that creates an artificial *Hae*III recognition site in all products (Khan et al., 2013). The assay yields products of 144bp for *DEFA1* (*Hae*III present) and 161bp for *DEFA3* (*Hae*III absent) (Figure 10). For primer sequences and PCR conditions see Khan 2013.

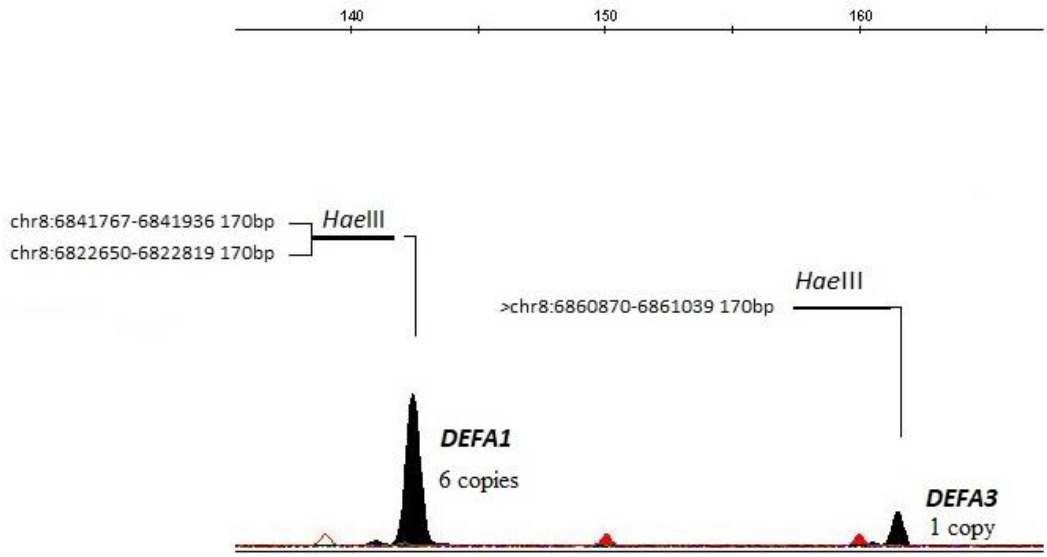


Figure 10 - Predicted PCR products from DefHae3 primers and actual peaks obtained from capillary electrophoresis after *Hae*III digestion. This is a 7-copy individual and we interpret the profile as being 1 copy *of DEFA*3 and 6 copies of *DEFA*1. Genomic coordinates are from GRCh37/hg19 Feb. 2009.

2.3.2Indel5 and 7bpdup Multiplex PCR

The 5bp insertion/deletion ("Indel5") and 7bp duplication ("7bpdup") are two variants well described by Khan in 2013, that have been found in many repeat units within *DEFA1A3* locus (see Figure 2). To reduce the numbers of PCR involved, we did 5bp-7bp indel assays in a multiplex reaction as follows. The Indel5/7bpdup PCR was performed in 10 µl PCR mix which included 1 µl of 25ng genomic DNA, 1 µl of LD MIX, 1 µl of each forward and reverse primers for both 5bp and 7bp, 0.1 µl Taq and 5.9 µl H_2O. PCR cycle conditions were 25 cycles of 95ºC for 30 seconds, 57º for 30 seconds, 70ºC for 30 seconds followed by a 72ºC hold for 40 minutes. The forward primer used in the 5bp assay was HEX labelled while for 7bp assay we used a FAM labelled forward primer. The primers were designed to generate amplicons of 124bp and 129bp for Indel5 and 275bp/282bp for 7bdup. For primers sequences see Khan 2013. The ratio recorded for indel5

was the ratio of 124bp – deleted product to undeleted – 129bp product (Figure 11) while the ratios recorded for 7bp was the ratios of unduplicated - 275bp product to duplicated – 282bp product (Figure 12).

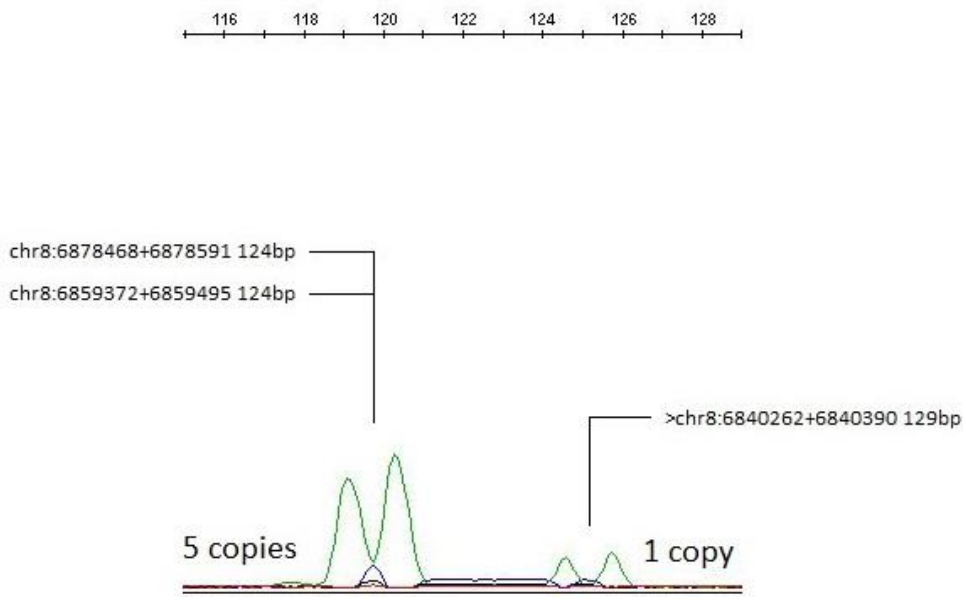


Figure 11 - Predicted PCR products from Indel5 primers and actual peaks obtained from capillary electrophoresis. Each product is represented as a double peak, differing by one base pair, presumably due to incomplete dA addition of the 3' end of the product. This is a 6-copy individual and we can interpret the profile as being 5 copies of undeleted and 1 copy of the deleted. Genomic coordinates are from GRCh37/hg19 Feb. 2009.

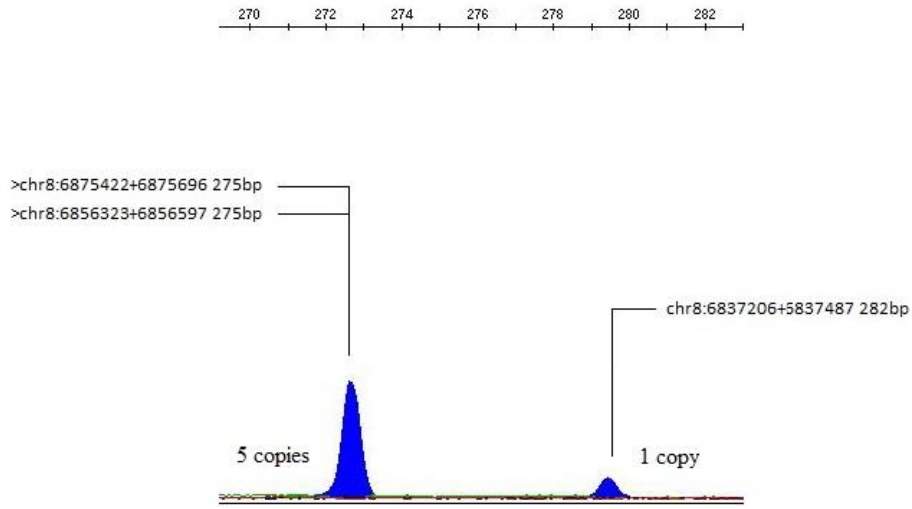


Figure 12 - Predicted PCR products from 7bp duplication primers and actual peaks obtained from capillary electrophoresis. This is a 6-copy individual and we interpret the profile as being 5 copies of unduplicated and 1 copy of the duplicated. Genomic coordinates are from GRCh37/hg19 Feb. 2009.

2.4 Electrophoresis

The PCR products for both PRTs and all the Allele Ratio measurement for each sample were analysed via ABI 3130XL Genetic Analyzer. The ABI process separates the single-stranded products through capillary electrophoresis and uses florescence detection to quantify the amounts. One innovation in this approach was to do electrophoresis not of raw, unpurified PCR products but on AMPure purified products which delivered cleaner results.

The AMPure assay removes unwanted unincorporated primers and other interfering material and offers the opportunity to concentrate or dilute the samples before electrophoresis. It also allows the multiplexing of larger numbers of PCRs into a single electrophoresis capillary. The disadvantages are the extra pipetting steps, and the way the procedures fixes the proportions of products in the mix. For each sample the PCR products were mixed and purified before electrophoresis following the Agencourt AMPure XP PCR Purification protocol. The master mix had the following concentrations: 1 µl from Cen1 - Fam labelled PCR, 1 µl from Cen1 – Hex labelled + Indel4 multiplex PCR, 1 µl from Tel 2 – Fam labelled PCR, 1 µl from Tel 2 – Hex labelled PCR, 1 µl from Indel5 + 7bp multiplex PCR and 3 µl from Defa*Hae*III PCR.

After the AMPure purification step, 1 µl from each fluorescently labelled PCR product mixture was added to 10 µl of HiDi Formamide and ROX-labelled markers. For each group of 16 samples a ratio of 170

µl of HiDi to 2 µl of ROX was used. Then, as a final step the samples were denatured at 96 ºC for 3 minutes and submitted for electrophoresis on an ABI 3130xl Genetic Analyzer (DeepSeq, University of Nottinigham). The same injection time and voltage was used for all the samples: 1kv for 15s. After the capillary electrophoresis step, the ABI data collection program generates an fsa. file that contains all the peaks profiles for which a sample have been tested for.

2.5 Data analysis

Two main programs were used to assign the correct copy number for each sample: GeneMapper (Applied Biosystems and a custom Maximum Likelihood Program developed by Prof. John Armour (Black et al., 2014).

2.5.1 Gene Mapper Analysis

After the ABI capillary electrophoresis step, the fsa files were loaded in the Gene Mapper program. The program allowed us to visualise and analyse all the peaks profiles of each PRT and allelic ratios amplicons involved in the analysis and to quantify all the results in Excel spreadsheets. Therefore, for each peak, analysis parameters including the peak threshold and peak range had to be defined relative to the ROX markers. The peak range template can be seen in Table 2. 10 pts was the minimum peak amplitude threshold used for all samples. Throughout this analysis procedure we made sure that all

the samples that were eligible for further analysis delivered accurate results.

Marker Name	Dye Colour	Min Size	Max Size
5bp deleted	Green	118.5	121.5
5bp undeleted	Green	124.0	126.5
4bp deleted	Green	208.0	210.0
4bp undeleted	Green	212.0	214.9
7bp unduplicated	Blue	271.0	275.0
7bp duplicated	Blue	279.0	282.0
Tel2-A-Fam	Blue	258.0	260.0
Tel2-B-Fam	Blue	261.7	263.5
Tel2-A-Hex	Green	258.5	260.5
Tel2-B-Hex	Green	262.0	264.0
Cen1-A-Fam	Blue	331.0	335.5
Cen1-B-Fam	Blue	441.2	443.2
Cen1-A-Hex	Green	330.0	334.0
Cen1-B-Hex	Green	441.0	443.5
DEFA1	Yellow	142.0	146.0
DEFA3	Yellow	159.0	163.0

Table 2 - All the assays for which a sample has been teste for are displayed under the "Marker Name" column and their minimum and maximum size range input designated for software recognition of the relevant peaks.

All samples have been measured by using "area" in GeneMapper and they have been considered saturated only if they exceeded 8000 fluorescence units (height). In the case of products represented by double peaks, we only accepted the use of a ratio deriving from summing the two peaks if the proportions of the peaks were the same between the two products. Results that did not match our expectations were manually inspected, and any problems found in the analysis rectified manually. For example, the peak threshold had to be modified from plate to plate because there was a natural variation between plates.

2.5.2 Defining Ratios

After all the adjustment were made, the GeneMapper output was analysed using the Excel program. The program was used to determine the raw copy number for all samples in both TEL2 and CEN1, as our measurements were calibrated internally and not against standard reference samples, and to calculate the ratios for every other assay.

Both TEL2 and CEN1 PRTs were carried out in duplicate PCRs, with primers labelled either in HEX (green) or FAM (blue) (see Section 2.2). In order to determine the raw copy number for each sample two ratios were calculated initially: one between the two HEX-labelled peaks and the other one between the two FAM-labelled peaks (Table 3).

	TEL2 -A-FAM	TEL2-B-FAM	TEL2-FAM-Ratio	TEL2-A-HEX	TEL2-B-HEX	TEL2-HEX-Ratio
A01	2456	6084	2.477198697	2889	6325	2.189338872
A02	6550	14991	2.28870229	3598	8080	2.245692051
A03	5632	15375	2.72993608	2878	7797	2.709173037
A04	1979	2843	1.436584133	1964	2803	1.427189409
A05	7511	23159	3.083344428	5207	17246	3.312079892

Table 3 - Example of how the ratios for TEL2 in FAM and HEX are calculated

Then an average was calculated between the two individual ratios (CEN1 FAM Ratio / CEN1 HEX Ratio and TEL2 FAM Ratio / TEL2 Hex Ratio) (Table 4).

Name	CEN1 FAM Ratio	Cen 1 HEX Ratio	CEN1 Average	TEL2 FAM Ratio	Tel2 HEX Ratio	TEL2 AVERAGE
A01	2.99699925	3.186103013	3.091551131	2.853745928	2.423541595	2.638643762
A02	1.915935673	2.060240964	1.988088318	2.01506202	1.789668751	1.902365386
A03	1.696713309	1.752800597	1.724756953	1.551594073	1.359257623	1.455425848
A04	2.243597221	2.964976651	2.604286936	1.975404117	1.698130383	1.83676725
A05	3.60787172	3.689095128	3.648483424	3.079676674	2.586682663	2.833179669

Table 4 - Example of how the average between TEL2 in FAM and HEX Ratios and CEN1 in FAM and HEX are calculated.

PRT averages correlate well with each other and when plotted all together resulting in formation of tight clusters around integer values of 1, 1.5, 2, and 2.5 (Figure 13). We knew from previous work conducted by Tyson and Armour in 2012 that the first cluster is expected to correspond to a copy number of 4, as a copy number of 3 is extremely rare. Therefore, the first cluster (around integer value of 1) corresponds to a copy number of 4, the second one (1.5) corresponds to a copy number of 5 and so on

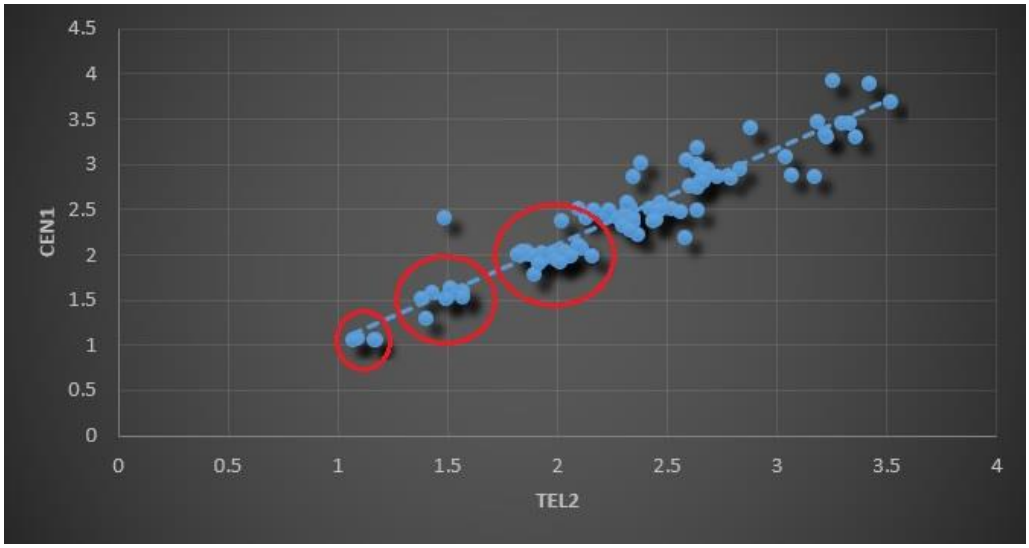


Figure 13 - Copy number measure determined by CEN1 PRT plotted against copy number estimates TEL2 PRT, showing the correlation between the two measures (r^2=?????) and evidence of clustering around integer values of 1, 1.5 and 2.

The average of each cluster was extracted and used in a linear regression formula to determine the unrounded copy number for each sample.

2.5.3 Maximum likelihood analysis.

Using both the unrounded copy numbers and the allelic ratios, a maximum likelihood program such as the one applied by Khan et al in 2013, was then used to provide the Most Likely Copy Number (MLCN) and minimum ratio data for all the samples. The program, written by Prof. John Armour, combines the measurements of each PRT (unrounded copy number) and allelic ratio results from *DEFA1A3*, indel5, 7bp duplication and 4bp deletion assays and evaluates the likelihood of getting the data conditional upon a particular copy number, assuming in turn each possible copy number from 2 to 16. The copy number that maximised the joint probability of all the data recorded for a sample is termed the Maximum Likelihood Copy Number (MLCN). The ratio conventions for *DEFA1A3*, indel5, 7bp and 4bp indels are explained in Khan et al. 2013 and Ai et al. 2016. The highest probability in each group is scaled to 1, and all other values expressed relative to that.

For a more thorough explanation, we have selected sample A05 as an example. Given PRT copy number results of 6.88808 and 7.26329, indel5 deletion/non-deletion ratio of 5.64105, *DEFA1/DEFA3* ratio of 5.49231, 7bp nonduplication/duplication ratio of 5.97692 and a 4bpdeletion/non-duplication ratio of 0.40347, the program gave the

following output (Table 5). In Table 5, N=7 is the most likely copy number. The next highest value in the "Combined" analysis is 5.5 x 10^{-4} for N=8, and so N=7 has a probability that is approximately 10^3 higher than N=8. This is how the minimum ratio of 1813.27 is calculated at the bottom left; the minimum ratio is the factor by which the maximum likelihood value (in this case with N=7) exceeds the next highest value (in this case for N=8) The results from Table 5 are from a single sample, but the analysis can be performed for up to 999 samples.

			best					relative likelihood values for						
			copy no.	analysis	N=2	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10	N=11
Sample ID		A05	7	Combined	0	2.16789e-	1.3E-110	3.31E-39	1.02E-06	1	0.000551	3.54E-08	2.19E-10	1.23E-09
PRT1	6.88808			PRT1	4E-130	1.76E-73	7.02E-38	2.18E-14	0.065932	1	0.102621	0.015653	0.003222	0.00096
PRT2	7.26329			PRT2	1E-150	1.87E-80	1.2E-34	4.42E-14	0.000809	1	0.373025	0.059431	0.011861	0.003279
indel5 rati	5.64105			indel5	0	5.68E-57	4.34E-13	0.004802	0.795928	0.622568	0.087301	0.00752	0.004802	0.184916
DEFA1A3 r	5.49231			DEFA1A3	6.4E-287	1.94E-46	4.02E-12	0.001471	0.358637	1	0.57115	0.185064	0.050668	0.056025
7bp dup ra	5.97692			7bp dup	0	2.38E-60	1.42E-15	0.000207	0.265806	1	0.476543	0.108608	0.019898	0.023762
4bp del rat	0.40347			4bp del	8.03E-07	0.156935	0.37747	0.001459	0.156935	1	0.37747	0.156935	0.738387	0.989323
minimum ratio		1813.27												

Table 5 - -Maximum likelihood output for sample A05 with a copy number of 7. N=7 gives the highest probability value of 1 for PRT1, PRT2, *DEFA1A3*, 7bp, while N=13 (offscreen) is the best copy number for indel5. N=7 is the most likely copy number overall in the combined analysis in the first line. For clarity, the display of the table misses out copy numbers between 12 and 16.

In some cases, depending on the origin of the samples, ratio assays can often be uninformative, and therefore the analysis does not include the ratios at all. We have noticed this trait in the case of Indel4 assay, where the presence of Indel4 is more frequent in Asian populations than in European populations. Nonetheless, the input file will need a value to cover the missing measure, and adding a value of 0, -1 or -10 for everyone will not affect the copy number outcome for the experiment in which the missing value has not been typed. "0" value indicates that there is no signal from the numerator (deleted at indel5. Non-duplicated at 7bp dup, deleted at 4bp del, or *DEFA1*), "-

1" indicates that there is no signal from the denominator (non-deleted at indel5, duplicated at 7bp dup, undeleted at 4bp del, or *DEFA3*, and "-10" indicates that no result was obtained (Table 6).

Plate 1		TEL2	CEN1	5bp	DEFA1A3	7bp	4bp
	A01	5.548838	7.043333	5.5774	2.434826	2.50274	0
	A02	5.925916	5.958909	1.040474	4.303233	0.519326	0
	A03	7.880991	7.954443	5.670377	2.81457	1.62583	0.148836
	A04	7.020209	7.286038	2.565636	5.334732	2.623986	0
	A05	6.888082	7.263287	5.641048	5.492313	5.976915	0.40347
	A06	4.954158	5.093943	3.887375	1.437999	0.696508	0
	A07	6.213758	6.113676	4.837636	-1	2.083191	0

Table 6 - Example of 7 input samples and their ratio assays from Plate 1. The 4bp assays are uninformative for samples A01, A02, A04, A06 and A07 and they all have been marked with 0 because their numerator (4bp deleted) signal is missing. Also, sample A07 missing value from DEFA1A3 measurement has been market with "-1" because the signal from the denominator (DEFA3) was absent.

The ratio assays are evaluated by computing a "normalised ratio", the value of the ratio relative to the expected one. The ratios of products that we observed are not necessarily exactly the same as the ratio of copies that we are measuring. The ratios of the products are different from the ratios of input copies because of the biases in the PCR. A minimum ratio is then generated for each sample which is the ratio between the most likely copy number and the second most likely copy number. Therefore, the minimum ratio result acts like an index of confidence in the assignment of copy numbers (the higher the minimum ratio is, the more likely is that the copy number was assigned with great accuracy).

3. Results

3.1 DEFA1A3 COPY NUMBERS

DEFA1 and *DEFA3* are copy-variable genes present within the alpha-defensin region on chromosome 8, and accurate copy number measurement has encountered difficulties (Khan et al., 2013). In this study, six different measurements were used to measure the copy number at *DEFA1A3* locus – two paralogue ratios test (PRTs) to determine total gene copy number, and four ratio measurements between repeat variants. Each PRT was performed in duplicate, so that each sample was tested using a total of eight PCRs, four PRTs and four variant ratio tests. The four variants used were composed of three indels (5 bp, 7bp, 4bp) followed by the measurement of one base substitution used to differentiate the *DEFA1* and *DEFA3* variants (Khan et al., 2013; Black et al., 2014 and Ai et al., 2016) (see section 2.2;2,3). Eight plates of DNA samples were processed following a standard operation procedure and analysed using a Maximum Likelihood method similar to the one employed by Khan et al.2013 (see section 2.5). The ABI profile of a randomly selected sample, showing the arrangement of the two PRT and four allelic ratios at *DEFA1A3* locus is displayed in Figure 14.

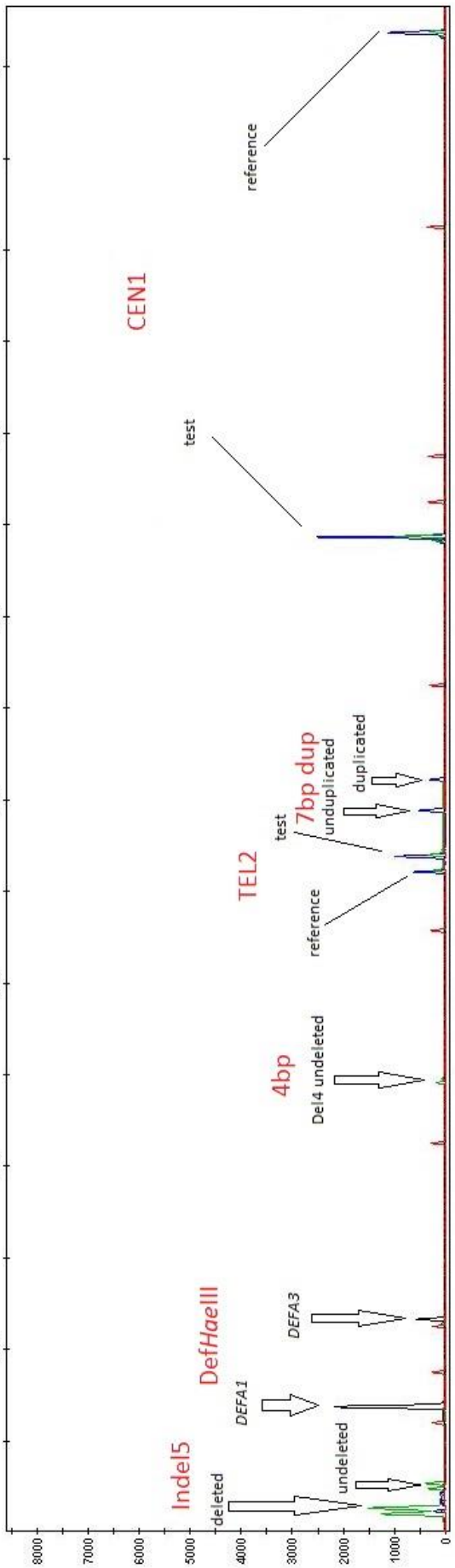


Figure 14 - Representation of an ABI profile from a random sample which includes all the PRT and allelic ratio measurement. Displayed on the y axis is the full range of fluorescent units up to 8000; anything above this will saturate the detector. The red peaks come from ROX markers. Cen1 and Tel2 were analysed in both HEX (green) and FAM (blue). This is a 7-copy number sample.

For greater accuracy, the ratio measurements of TEL2 and CEN1 were each done in duplicate using two differently-labelled primers (HEX/green and FAM/blue). The ratio is calculated by comparing the amount of PCR product in both HEX and FAM from the 263/444 (non-reference product) with the amount from 260/333 (reference product), assuming that the reference locus always has two-copies per diploid genome. The same principle of deriving a ratio of two differently-sized peaks applies to the measurement of the four allelic ratios and they were calculated as follows: Def*Hae*III ratio is the ratio of *DEFA1* to *DEFA3*, the Indel5 ratio is the ratio of deleted (124bp) to undeleted (129bp), the 7bpdup ratio is the ratio of unduplicated (275bp) to duplicated (282) and the 4bp ratio is the ratio of deleted (211bp) to undeleted (215bp).

The six measurements were then used to determine *DEFA1A3* copy number in 744 unrelated DNA samples from Europeans from a study of periodontitis (Munz et al., 2017). Of these 744 samples, 3 failed (0.40%) to produce useful results and 31 other targeted samples for which all the procedures have been successful were further analysed for reasons discussed below. Similar to Khan et al., 2013, PRT measurements and Aldred et al's MAPH measurement, we have found that in our analysis the deduced integer *DEFA1A3* copy numbers

ranged from 4 to 12. The distribution and frequency of copy number measurement for all 741 samples is displayed in Figure 15.

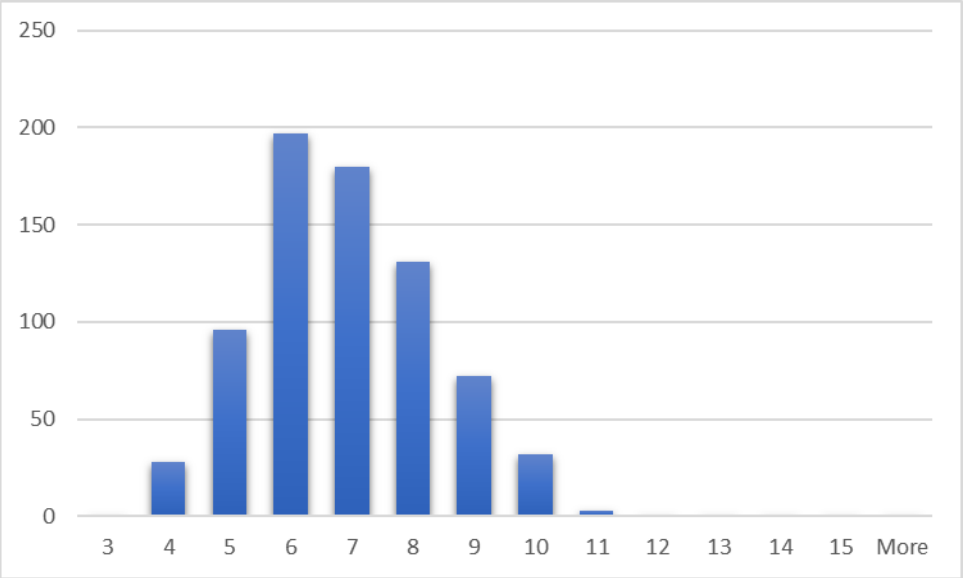


Figure 15 - The distribution and frequency of measured copy numbers in 741 unrelated samples from Europeans.

3.2 Quality control

We wanted to validate the consistency of measurements between different plates, and in particular to evaluate any evidence for batch effects, in which the values returned may depend on the particular typing experiment performed. There might be small but important differences in exactly how the experiment went from one plate to another. We pursued this in initial data by comparing allelic ratios between plates; as a simple check, we had a close look at the distribution of uninformative (zero) ratios between plates, to reassure ourselves that our data is consistent with the one described in the literature for a European population. In the case of the allelic ratios,

the frequency of having a ratio of 0 (for example, samples with no copies of *DEFA3*) should be approximately the same in all plates, and it seems that our data was mostly consistent with the expectation. With the exception of two allelic ratios that stood out in the overall analysis, *DEFA1A3* in the second plate with a percentage of 40% and Indel4 in the fourth plate with 79.5%, our data was consistent between all eight plates (Table 7). We specifically inspected the results from those plates and saw no evidence of technical problems. Furthermore, to test if the results for 4bp in plate 4 and DEFA1A3 in plate 2 are different from all the others we did a Chi-Square test of independence and found that there is a significant difference between them and all the other plates. With a p value of 1.71×10^{-3} for 4bp and 1.84×10^{-5} for DEFA1A3. We expect some variation between the plates because we sample a different set of people randomly. The results deduced from the analysis of the overall allelic ratios equal with zero (Table 8) proved that our data strongly agrees with the literature and matched with previous results from a European population (Khan et al., 2013, Black et al., 2014).

	Plate 1	Plate 2	Plate 3	Plate 4	Plate 5	Plate 6	Plate 7	Plate 8
4bp %	63.8	57.7	67.4	79.5	59.3	58.8	50.5	58.1
7bp %	5.6	5.5	7.8	4.5	5.4	6.6	4.4	3.4
DEFA1A3 %	23.6	40	13.4	20.4	19.7	14.4	15.7	15.1
5bp %	13.8	8.8	14.6	10	9.8	11.1	11.2	8.1

Table 7 - The percentage of samples with a ratio of 0 between plates.

Percentage	Plates
4bp %	61.9
7bp %	5.4
DEFA1A3 %	20.3
5bp %	10.9

Table 8 - Allelic ratio percentage of samples with a ratio of 0 for all eight plates.

3.3 PRT agreement

Each ABI profile was individually analysed to ensure that there are no oversaturated or uneven PRT and allelic ratio peak measurements that may affect the quality of the results. All eight plates delivered good results, with some repeat testing needed for some plates due to some inconsistency in the case of the CEN1-4bp assay. We have observed that there seemed to be a correlation between freshly diluted primers and the success of the experiment, especially in the case of the CEN1-4bp assay; this is presumably because the multiplex PCR (see Methods) may be more sensitive to the condition of the primers.

The full set of PRT measurements in both TEL2 and CEN1 for all 741 samples was then used to determine the raw copy number. Both TEL2 and CEN1 PRTs were carried out in duplicate PCRs using two similar PRT assays: one assay with HEX-labelled primers and the second one with FAM-labelled primers. Firstly, two ratios are

calculated, one between the two HEX-labelled peaks and the other one between the two FAM-labelled peaks.

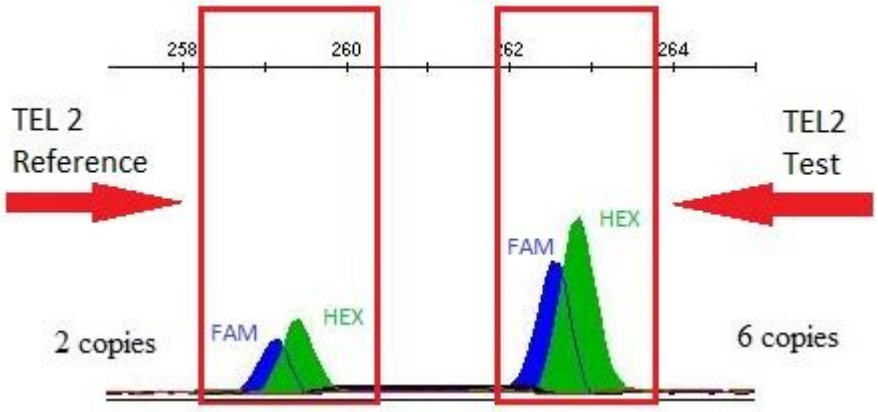


Figure 16 -Predicted PCR products from TEL2 primers and actual peaks obtained from capillary electrophoresis. This is the result of a duplicate test in both FAM (blue) and HEX (green). This individual was determined to have 8 copies of DEFA1A3. We can therefore interpret this profile as being 2 telomeric (reference) copies and 6 non-telomeric (test).

Secondly, the final ratios in TEL2 and CEN1 are calculated as an average of the ratios from two individual tests. PRT measurements correlate well with each other both individually in each plate (Table 9) and when plotted all together resulting in formation of tight clusters around integer values of 1, 1.5, 2, and 2.5 (Figure 17). We knew from the literature that the first cluster is expected to correspond to a copy number of 4, as a copy number of 3 is extremely rare, with just one example found by Tyson and Armour in 2012. Therefore, the first cluster (around integer value of 1) corresponds to a copy number of 4, the second one (1.5) corresponds to a copy number of 5, the third (2) to a copy number of 6 and the forth (2.5) with a copy number of 7 while the clusters for high copy numbers are more dispersed. TEL2 and CEN1 ratios for the lowest copy number cluster (corresponding to a copy number of 4) are around values of 1

for both TEL2 and CEN1 PRTs, which can be explained as 2 non-reference copies and 2 reference copies. However, the clusters for a copy number of 6 centre on values slightly less than 2 for TEL2 and more than 2 for CEN1. This is completely explicable in terms of the increased efficiency of amplification of shorter fragments. The average of each formed cluster for copy numbers of 4,5,6 and 7 was recorded and used in a linear regression formula to determine the raw copy number for all samples in both TEL2 and CEN1; our measurements were calibrated internally and not against standard reference samples.

Plate	1	2	3	4	5	6	7	8
TEL2/CEN1 R^2	0.872	0.716	0.870	0.914	0.894	0.847	0.877	0.817

Table9 – R^2 showing the agreement between TEL2 and CEN1 measurements in all eight plates. The highest agreement is seen in plate 4 with a R^2 of 0.914, while the lowest agreement with a R^2 of 0.716 is present in plate 2. Although, there is some random variation between plates, the R^2 results are consistent among plates.

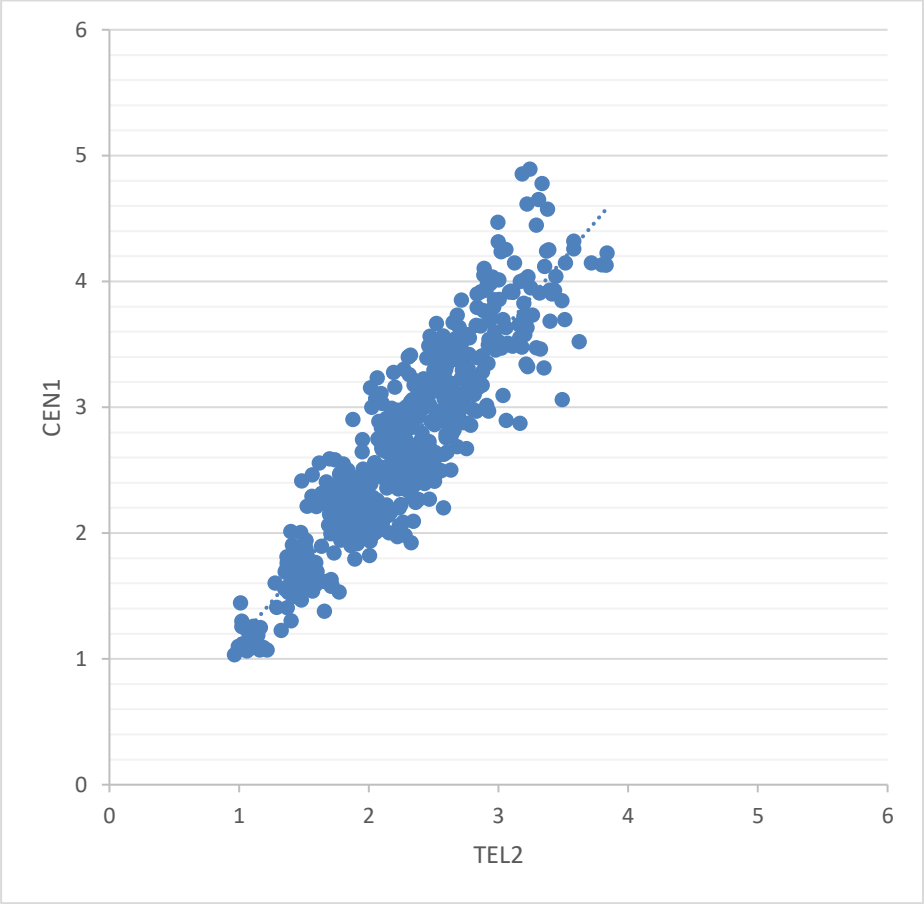


Figure 17 - PRT measurements plotted against each other resulting in the formation of clusters of raw copy numbers. This display shows samples for which the two PRT measurements were concordant but there are exceptions (see Figure 26). The correlation between the two measurements is R2=0.8319. The PRT results are from 741 European samples with periodontitis.

3.4 Copy Number Assignment Quality.

Having the inferred copy number and the allelic ratios we were able analyse the data with a Maximum Likelihood approach. The Maximum Likelihood approach assumes that all the measurements on a single sample are consistent with a single true copy number, and the results are evaluated for their consistency with each of the possible copy numbers from 2 to 20. The copy number that maximises the probability of the full set of measurements is defined as the maximum

likelihood copy number (MLCN). We started by looking at the normalised ratios for PRTs; the normalised ratio for PRTs is the ratio of the observed copy number estimated from PRT to the MLCN. Thus, a PRT measurement of 6.6 for an MLCN of 6 would yield a normalised ratio of 1.1; ideally, the distribution of normalised ratios for well-calibrated and accurate measure should have a mean of about 1, and a low standard deviation. The maximum likelihood model used assumes that each copy number measurement is drawn from a normal (Gaussian) distribution, and if the data does not conform to those expectations the maximum likelihood calculation will not be accurate. We observed that both the mean value of the normalised ratio for the TEL2 PRT (TEL2= 0.984) and for the CEN1 PRT (CEN1= 0.997) were close to 1, indicating that our measurements were accurate and around the same mean values with a standard deviation of 0.067404 for TEL2 and 0.058809 for CEN1. The distributions of the normalised ratios for both PRTs are shown in Figure 18 and Figure 19 and conform well to the expectation of a normal distribution.

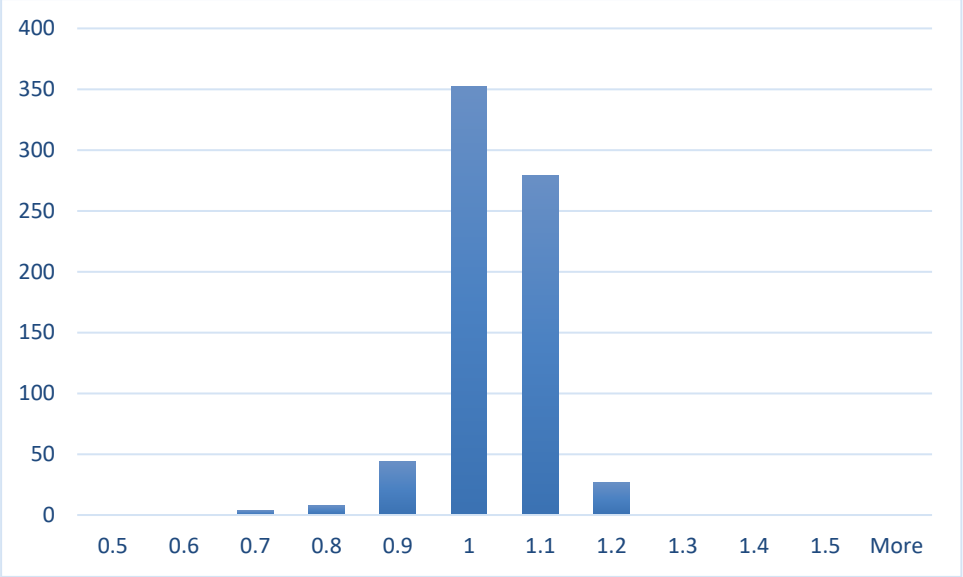


Figure 18 - Normalised ratio distribution for TEL2.

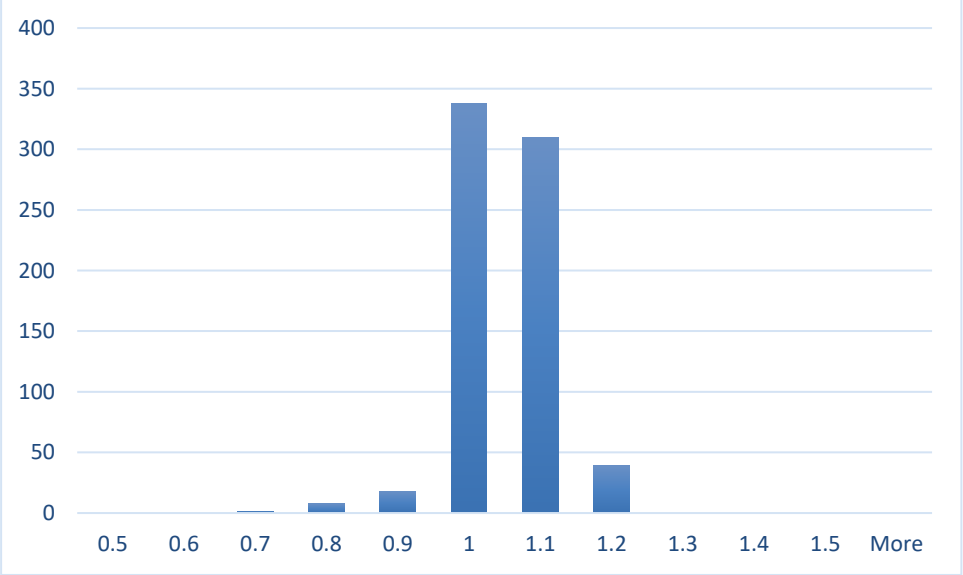


Figure 19 - Normalised ratio distribution for CEN1.

In addition to the PRT measurements, the mean and standard deviation values observed for allelic ratios (Table 10) were strongly correlated with those observed in previous work (Khan et al. 2013).

Ratio Assays	Mean Observed	Mean Assumed	SD Observed	SD Assumed
Indel5	0.98	1.05	0.084	0.11
Del4	1.15	1.04	0.19	0.12
7bpdup	1.02	1	0.09	0.12
Def*Hae*III	0.99	0.91	0.127	0.126

Table 10 - Assumed mean and standard deviation from previous work compared with observed measurement for all ratio assays deducted after Maximum likelihood analysis of 741 individuals.

It is important to reassure ourselves that we obtained accurate results from all allelic ratio assays in all 741 samples, as they help us validate the unrounded total copy number determined from the PRTs. Furthermore, the ratios obtained will help us to establish the correct copy numbers for either the deleted or the undeleted variants in the case of Indel5, *DEFA1* or *DEFA3*, deleted or undeleted variants of 4bp and the duplicate or the unduplicated variants of 7bpdup.

We were particularly concerned to evaluate how the Def*Hae*III reflected copy numbers. From previous work (Khan, 2012) we knew that for *DEFA1A3* assay there is a bias towards the undigested (*DEFA1*) products arising from incomplete digestion and this copy number measurement will require some further adjustments. Because this PCR assay may form single-stranded products or heteroduplexes, *Hae*III enzyme will fail to distinguish the digestion site leading to incomplete

digestion and a ratio shift towards more *DEFA*3. For example, if we have a ratio of 1:1 (*DEFA1:DEFA3*), in the products we will always get more of the undigested (*DEFA3*-type) product if the digestion is incomplete, shifting the ratio to (for example) 1:1.5. Therefore, we used the observed ratios to predict the number of copies per person for *DEFA*3 based on the MLCN and the observed ratio (Figure 20). We observed that the first non-zero peak was not at 1 but at about 1.5. This suggested that the degree of incomplete digestion is leading to a distortion and we readjusted the data specifically for DefaHaeIII by a factor of 1.5 and included the new ratio in the analysis (Figure 21).

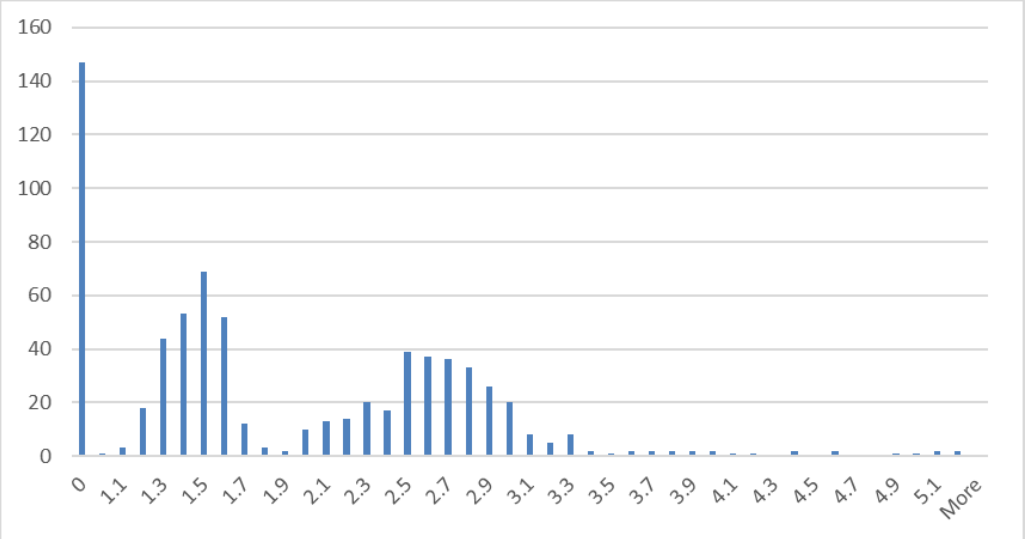


Figure 20- DEFA 3 copy number frequency before calibration. Absolute numbers of copies of DEFA3 estimated based on the DEFA3 to DEFA1 ratio and the total copy number of each sample. We would expect cluster around 0,1 and 2 if the measurements are properly calibrated.The majority of the samples formed clusters around integers with a value of 1.4 and 2.8 suggesting a miscalibration of DEFA1A3 by a factor of about 1.5.

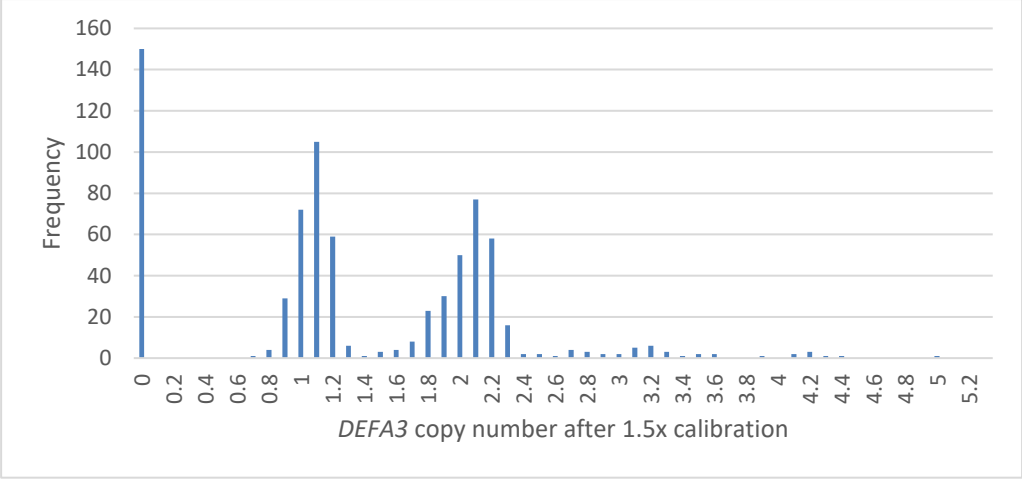


Figure 21 - DEFA 3 copy number frequency after calibration. Absolute numbers of copies of DEFA3 estimated based on the DEFA3 to DEFA1 ratio and the total copy number of each sample. The majority of the samples formed clusters around integers with a value of 1 and 2, suggesting that our calibration lead to better copy number assignment.

Because the copy number measurements of *DEFA1A3* are whole numbers, all the ratio measurements should be centred values expected from on whole numbers as well. In the case of Del4, Indel5 and 7bpdup assays, the ratios showed no evidence of bias, and therefore further calibration was not required (see figures 22, 23, 24).

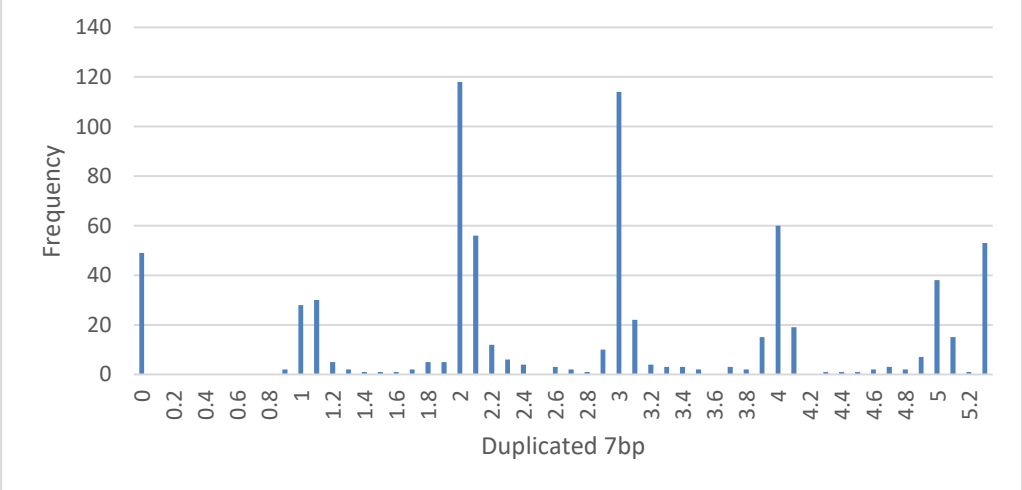


Figure 22 – Frequency of 7bp duplicated copy numbers deducted from the MLCN and the 7bpdup ratio in 741 European samples.

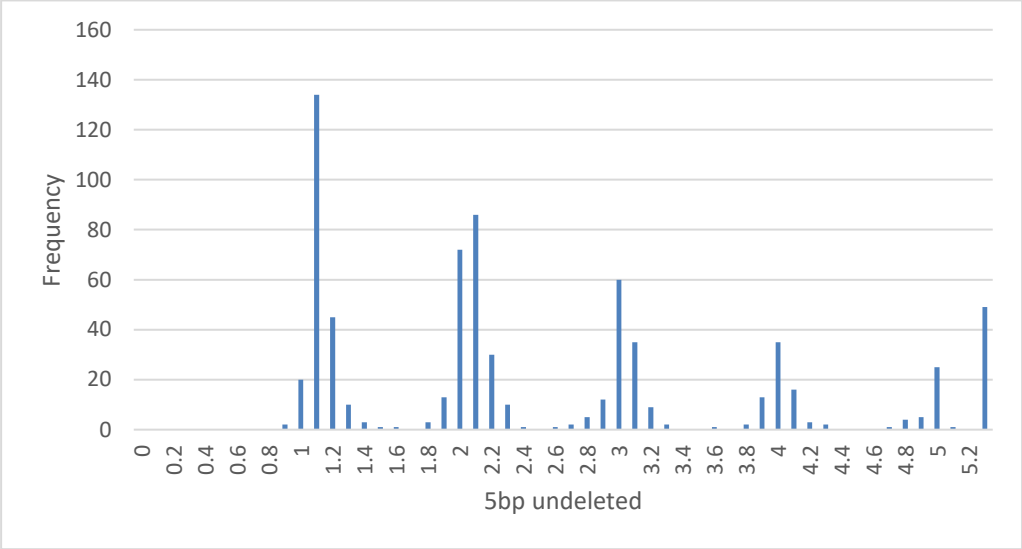


Figure 23- Frequency of 5bp undeleted copy numbers, estimated by combining the observed ratio with the MLCN, in 741 European samples.

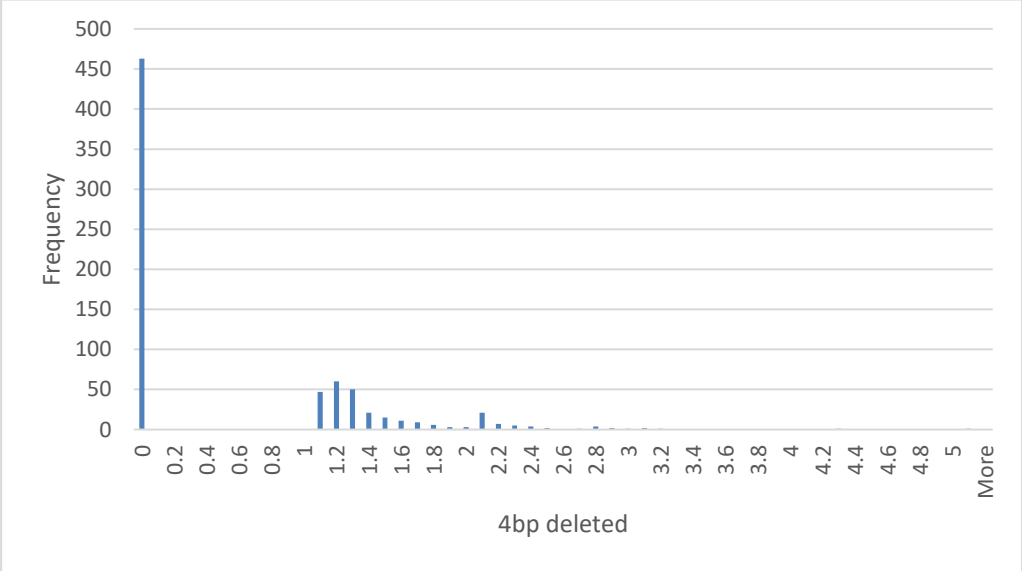


Figure 24 - Frequency of 4bp deleted copy numbers, estimated by combining the observed ratio with the MLCN, in 741 European samples.

After we reassured ourselves that the means and SD fitted for each measure were as expected for a European distribution, and that we have taken into the account and calibrated all the PRT biases, we proceeded by looking at the distribution of the minimum ratios. The

MLCN program assumed that all the PRTs and allelic ratios measure the same copy number and as a measurement of confidence it generates a minimum ratio (MR) value. The minimum ratio is the ratio of the likelihoods between the most likely copy number and the second most likely copy number. Therefore, the higher the minimum ratio is, the more likely it that we have correctly assigned the copy number.

The MR results very between just above the unity, up to several millions. A MR between 2 and 10 was recorded for 125 samples (16.9%) (Figure 25). The median MR value was 844.038, and the interquartile range was 19.42-47357.8. Most of the samples that delivered a minimum ratio of below 10 were high copy number samples or had missing data.

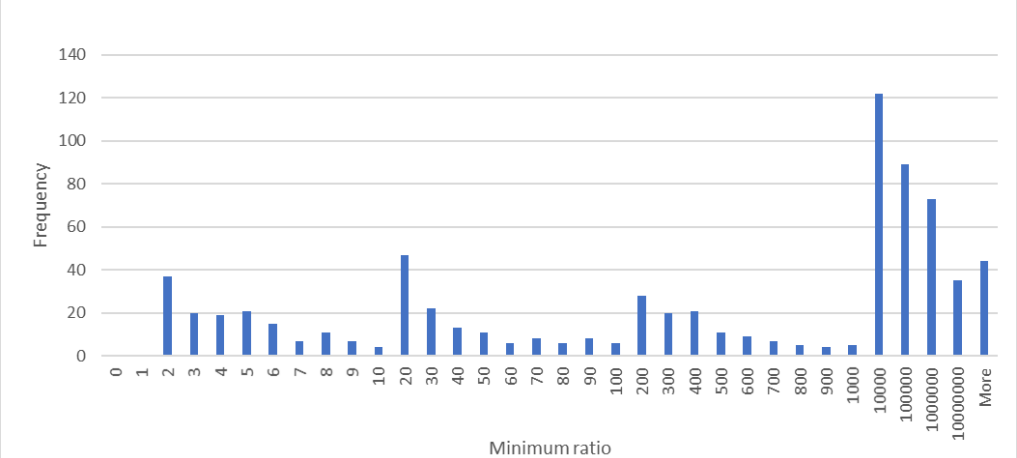


Figure 25 - Minimum ratio distribution in 741 unrelated samples. 83.08% of the samples had a minimum ratio greater than 10.

3.5 Concordance between PRT measurements

Among the 741 samples we noticed that for 31 of them there was a different relationship between CEN1 and TEL2 measurements. The CEN1 ratio measurements were between 1.5 to 2 times higher than TEL2 measurements for 27 samples, and we have found four cases in which TEL2 ratio measurements were between 1.5 and 2 times higher than CEN1 measurements. For this reason, the samples were excluded from our internal calibration, but they were analysed as normal through the Maximum Likelihood test and we therefore expected that those samples should have more indications of inconsistency between measurements. The overlay display of the 31 samples on the general distribution is shown in Figure 26.

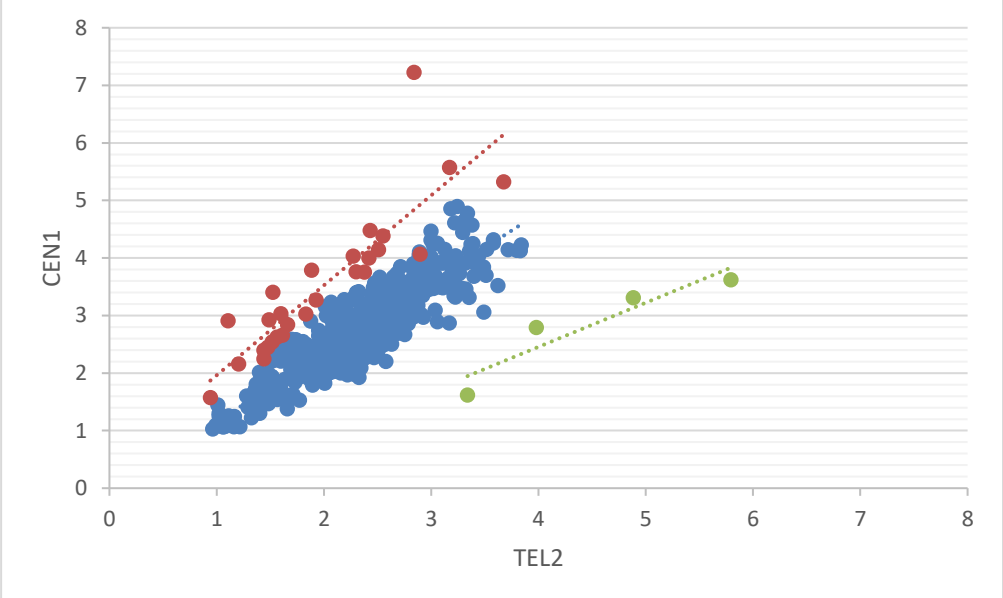


Figure 26 Marked with blue is the distribution of the 712 consistent samples overlaid in red by the distribution of the 27 samples were CEN1 measurements are higher than TEL2 measurement and marked in green the distribution of the 4 samples where TEL2 measurements are higher than CEN1 measurement.

This phenomenon is reproducible as we have repeated 15 random samples out of 31 and found the same measurements in both CEN1 and TEL2. It is important to mention that the PRT assays relies, for example, on CEN1 444 bp product always coming from the centromeric position and the 332 bp product always coming from non-centromeric position. If either the 444bp sequence was transferred to a non-centromeric position in an allele or a 332 bp sequence was transferred to the centromeric position, we will get the wrong answer. If there was a rearrangement of the locus such that a 332 bp product was derived from a centromeric (reference) position on one of the haplotypes, that will approximate double the ratio and will in turn give rise to an apparent copy number that is twice what it should be. For example, in Figure 26, the points shown in red could arise in some cases because the centromeric regions have been replaced by the non-centromeric regions, giving rise to a ratio approximately double the one expected for that copy number. Also, the 4 examples found in green could arise because a telomeric sequence is replaced by a sequence usually seen in non-telomeric regions. Agreement of measurements from both PRT assays gives further reassurance that the estimated copy number is indeed accurate as the independent tests are measuring the same number of repeat units.

We proceeded by looking at the distribution of the minimum ratio and found a MR median of just 7.992 and a quartile range from 2.110 to 37.645. This is an indication that CEN1 measurements in the samples might positively mislead the determination of the copy

numbers, but further analysis needs to be done in order to fully understand the cause of it. The most plausible mechanism for this kind of discrepancy is failure to amplify one copy of the reference because of sequence mismatches under the PRT primers, so that the apparent ratio is more likely to be correct in each case. Nevertheless, more work is needed to determine the true copy numbers for these samples before the nature of the discrepancies can be defined.

3.6 CNV Association

A CNV association analysis can be performed using different methods. In our case, the initial approach was to use all our results and to assign to every sample the best integer copy number and then use the copy number to do the association test. The problem in using a standard association test was that a simple association test will treat the copy numbers as if they are known without error, and it would ignore all the uncertainties associated with it. For example, it can make a just as strong a conclusion for samples that have been derived with very low confidence as for samples derived with high confidence. Therefore, we used an analysis method that will take into the account all the uncertainties present when deriving a copy number measurement and communicate them to the association analysis part. Even if the PRT method has been proven to be one of the most precise methods used in the determination of copy number, for a copy number of 6 and a minimum ratio value of (say) 43, there is still a chance that 1 time in 44 the copy number will be incorrectly measured.

For our purposes and because we worked with nonstandard data such as likelihood ratios from PRT based tests, an R package called CNVassoc was the best approach used in analysing the relationship between our cases and controls. In comparison with a standard association test the CNVassoc package allowed us to integrate the uncertainties from our measurements into the analysis (Subirana et al., 2011).

Based on our data, CNVassoc interpreted how many different copy number classes there were among cases and controls and then using uncertainties about the assignment derived from our likelihood analysis it did the association test. Our maximum likelihood analysis of PRT data was used to produce an intermediate file in which each sample is not associated with just one copy number but is associated with probabilities of each copy number, very like the likelihoods produced in our initial analysis. The Maximum likelihood output generates 15 copy number states ranging from 2 copies to 16 copies for each sample. Because the first three copy numbers (from 2 copies to 4 copies) and the last five (from 11 copies to 15 copies) gave a low probability, we combined the copy numbers probabilities ranging from 2-4 and 11-16, resulting in an 8 possible copy number states. To conform to the CNVassoc input format these are (confusingly) labelled 1-8. (Table 11). In comparison with another R package called CNVtools (Barnes et al., 2008), CNVassoc allowed us to do the CNV association testing based on the copy numbers and probabilities matrix that we provided from our own analysis.

	[1,]	[2,]	[3,]	[4,]	[5,]	[6,]	[7,]	[8,]
[1,]	6.76908E-15	0.037457623	0.962235711	0.000297211	2.87259E-07	6.98424E-07	5.83833E-06	2.63165E-06
[2,]	4.31312E-65	5.15241E-14	0.067455353	0.826041228	0.018752127	0.077444256	0.009320141	0.000986895
[3,]	8.24686E-33	5.47805E-07	0.999979114	1.66294E-05	3.50761E-06	1.17276E-07	4.51409E-08	3.83385E-08
[4,]	4.88805E-27	0.000159213	0.999628089	3.6548E-05	0.000163963	9.87383E-06	2.09339E-06	2.20146E-07
[5,]	7.74913E-26	8.02823E-06	0.998804313	0.001133713	4.75844E-05	4.94059E-06	1.18778E-06	2.33092E-07

Table 11 - We have 5 samples and 8 copy number states and each of these copy number states is associated with a probability. These probabilities have been obtained from the Maximum likelihood output.

We compared our data with 589 unrelated Europeans samples from unaffected individuals that act as controls for the periodontitis samples. The data for these 589 samples was obtained from Khan et al., 2013, Additional File 3. Through this method we wanted to test whether the distribution of our values matches a null hypothesis that there is no difference in the frequencies of copy numbers between our cases and the controls. The CNVassoc analysis gave us a CNV Likelihood Ratio Test output with a Chi-square value of 35.983 and a P-value equal to 7.30×10^{-6}. This showed that there is a statistically significant difference between the cases and the controls.

The comparison of the copy number distribution from our results, with the copy number distribution from Khan et al., 2013 is shown in Figure 27.

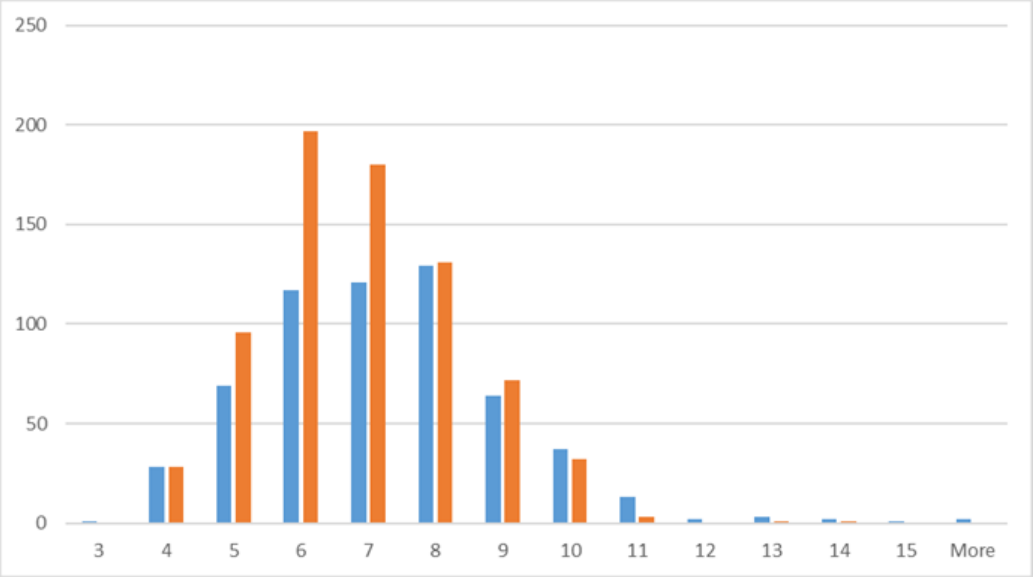


Figure 27 - The copy number distribution comparison between our periodontitis samples (marked in orange) and controls (marked in blue). There seems to be a significant over-representation of copy numbers of 5,6 and 7 among cases (p = 7.30x10-6 likelihood ratio test)

The two copy number distributions in Figure 27 showed clear differences, with an overrepresentation of central (N= 6 or 7) copy numbers among periodontitis samples. There is no apparent simple additive association with copy number, and the overrepresentation appears to apply to specific copy number states. Even if there might be an influence from typing artefacts, the fact that for a copy number of 6 in the case of periodontitis samples there are nearly 100 extra samples indicates that the distribution is not likely to be due to random sampling effects.

Because pattern is clearly non-additive, and because it seems to have an excess of average copy numbers (6, 7) we undertook some further analysis to examine the accuracy of the data extracted from Khan. To test if there are any artefacts between cases and controls

we looked at the distribution of estimated copy number for the 7bp duplication variant (Figure 28).

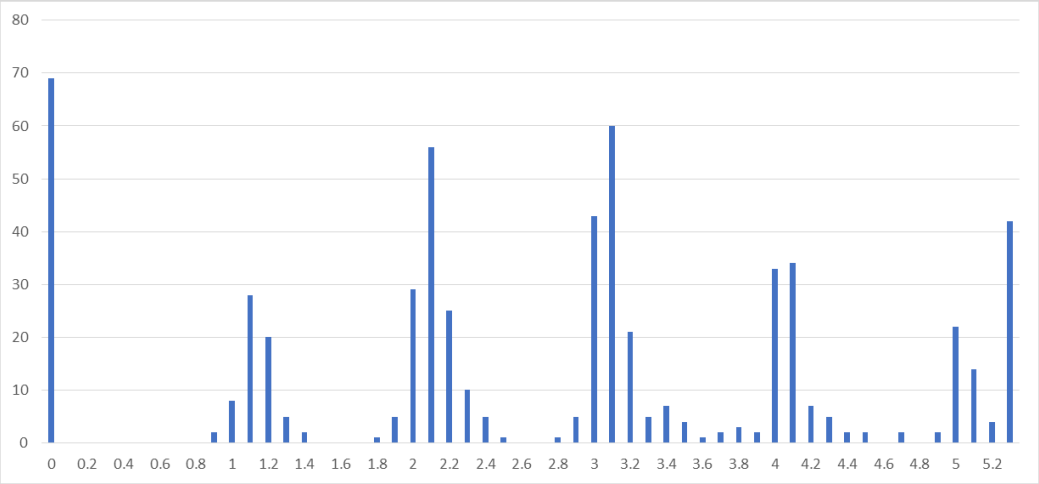


Figure 28 - Frequency of 7bp duplicated copy numbers deducted from the MLCN and the 7bpdup ratio in 589 samples from Khan et al., 2013.

The estimates cluster around integer values, and we would predict that pervasive error in this dataset would lead to erosion of the distinctions between these clusters. The internal consistency of the data from Khan gave us confidence that those copy numbers are accurate.

The overrepresentation of average copy numbers implies that there may be specific CNV alleles that are important in the case of samples with periodontitis. A specific overrepresentation of 3-copy alleles will tend to be associated with a few additional 5 copy samples but would predominantly lead to an excess of 6 and 7 and a few higher copy numbers samples. 3 is the average copy number haplotype and will not shift the overall distribution. Therefore, we can conclude that the important outcome of our analysis is not the overall copy number distribution but the distribution of a particular allele.

As a very simple confirmation of CNV Likelihood Association Test, we did a Cochran-Armitage trend test between 3 groups of copy numbers: low copy numbers (from 3 to 5), medium copy numbers (from 6 to 8) and high copy numbers (from 9 and higher). This might not be as sensitive as a Likelihood Ratio Test but nevertheless we should expect to see a shift in the overall distribution of copy numbers. The Cochran-Armitage test gave a P value of 0.02305, suggesting that there are significant differences between our cases and controls, confirming the outcome of the CNV Likelihood Ratio Test.

As a secondary analysis we looked at the distribution of the individual repeat unit types between cases and controls to see if there might be a significant difference in the representation of *DEFA3*, 4bp or 7bp. We divided the copy number states for each of our individual repeat variant types in 3 categories: 0 copy number, 1 copy number and 2 copy number or more. Using a Cochran-Armitage test we have found that in case of *DEFA3*, a P value of 0.2650 was recorded between cases and controls for copy number splits of 0, 1 and 2 or more, whereas for 4bp and 7bp a P value of 0.2939 and 0.222 was recorded. This suggested that there seems to be an equal distribution of copy numbers for the individual repeat unit types in our cases and controls. In the context of the outcome of the association with total copy number distribution, the Cochran-Armitage test is not a satisfactory test because it assumes additivity. Using the same three categories, we did a Chi-Square test to see if there are significant

difference between the distributions. We expect a chi-square test to be more sensitive to non-progressive shifts in distribution than the Cochran-Armitage test. In the case of *DEFA3* we obtained a chi-square statistic of 0.396 ($P > 0.05$). The 7bbpdup Chi-square statistic value was 7.814 with a p value of 0.020. For the 4bp deletion a Chi-square statistic of 6.828 was obtained and a p value of 0.032. We had particular reservation in the case of the 4bp control data because the data was not typed in Khan et el., 2013 and there is a good reason to be concerned that the control data is not comparable with the cases data. Using only a very small set of data (57 samples) for which the typing was clear, we concluded that there is a nominal significant difference between cases and controls (Figure 29).

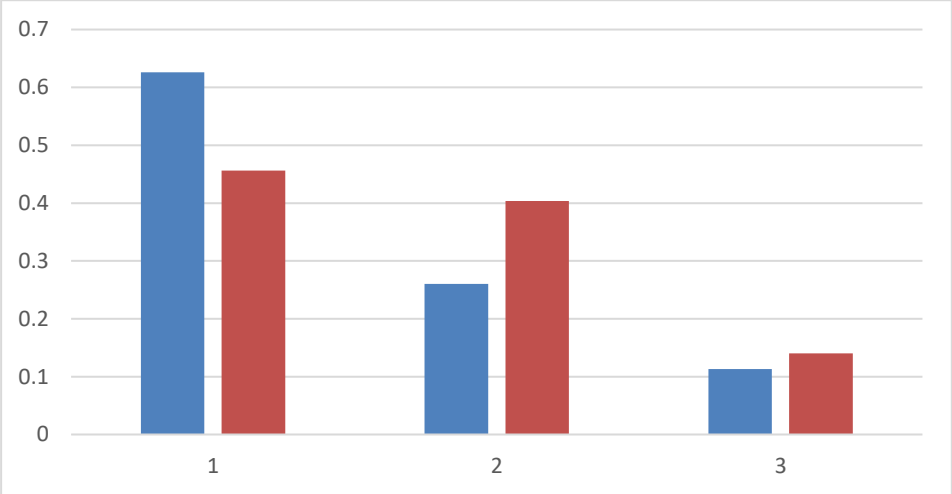


Figure 29 - The 4bp copy number states distribution, grouped in 3 categories. Our periodontitis samples are market in blue and the controls are marked in red. The first column represents the 0-copy number distribution, the second represents the 1-copy number distribution and third is the distribution of a 2-copy number or more copy numbers.

4. DISCUSSION

4.1 Measurement of CNVs

As discussed in subchapter 1.9, there are many methods that can be used to measure CNVs at multiallelic loci. This study focused not only in undertaking copy number measurement in 744 European samples with periodontitis but also to further demonstrate that the PRT method is a reliable and cost-effective method of CNV typing. We only had time to do CNV typing, but further studies can be done, where CNV data can be integrated with SNP data. The advantage is that the SNP data have already been obtained for these samples (Munz et al., 2017) and as a model for how to approach the further analysis, Ai et al., 2016 is a good example. They have found that the CNV association with IgA nephropathy was stronger than SNP association and it remained significant even after conditioning on the SNP genotypes. Also, SNP data can be integrated with diseases status and ultimately combined SNP and CNV data can be associated with disease status.

We were particularly concerned about the accuracy of the experiment, especially because our measurements were calibrated internally, in the absence of standard reference samples, and because (after promising initial testing) the TEL2 and CEN1 PRTs were being used for the first time in the systematic typing of a large set of samples. The agreement between the PRT assays combined with the results from four allelic ratio measurements gave us confidence in the

correct assignment of copy numbers. Furthermore, the recorded minimum ratio values have proven that our copy number assignment was supported by multiple consistent independent measurements and is likely to have a low frequency of error.

As mentioned in Chapter 3, 31 samples out of 744 were targeted to be further analysed. We have observed that is seems to be a discrepancy between TEL2 and CEN1 ratio measurements. Usually, the two PRT measurements should have about the same ratio as they are measuring the same copy number (Armour et al., 2000). In 27 of these 31 samples, CEN1 ratio measurement is from 1.5 to 2 times higher than TEL2 ratio, and in four samples the TEL2 ratio was between 1.5 to 2 times higher. It is possible that the CEN1 and TEL2, values might influence the copy number calling and this explanation is consistent with their low minimum ratio results. Our assumption is that there is something unusual in the sequence of these haplotypes that gives a higher CEN1 or TEL2 value, most likely a SNP affecting amplification of the reference amplicon. We will expect that this phenomenon to exist on a particular haplotype background and to share a particular SNP haplotype. Unfortunately, we did not have the time and the SNP data resources to investigate further in the context of this project.

Given the CNV Likelihood Ratio Test output, the P value from the Cochran-Armitage test and the total copy number distribution plot between the two sets of data, we can conclude that further analysis

need to be done in order to firmly establish if there is a true association between *DEFA1A3* copy number and periodontitis that is not influenced by measurement error.

4.2 CNV Association

Before we began our association study, we had to make sure that the copy number measurements from control samples we are using to compare our periodontitis data with are as accurate as possible. The data we have used was extracted from Khan et al., 2013 Additional file 3. Because the study was done five years ago, and our data was not typed using the same methods described in the article we tried to make sure that their method of measurement does not introduce some biases in the analysis, which can generate significant differences.

To do so, we compared the distribution of the 7bpdup individual repeat units. We have chosen this individual repeat, because among other repeats this is the most informative and accurate in terms of forming clusters around the integers (the corresponding data generated in this project can be seen in Figure 22). We therefore used the 7bpdup ratios and total copy numbers from Khan et al to generate a distribution and to compare it with our distribution. As a result, the two distributions shared the same characteristic of forming clusters around the same integers, reassuring us that the data extracted from Khan et al was likely to have a low frequency of measurement error.

As the results we obtained from the CNV Likelihood Test and the confirmation from Cochran-Armitage test, we conclude that there is an overall association between copy number distribution for *DEFA1A3* periodontitis in this sample. The pattern observed is consistent with overrepresentation of a 3-copy allele, but further work needs to be done. Among others, the integration of our CNV data for these periodontitis samples with existing SNP data would be a helpful next step in understanding the basis of this apparent association.

5. BIBLIOGRAPHY

Ai, Z., Li, M., Liu, W., Foo, J.-N., Mansouri, O., Yin, P., Zhou, Q., Tang, X., Dong, X., Feng, S., Xu, R., Zhong, Z., Chen, J., Wan, J., Lou, T., Yu, J., Zhou, Q., Fan, J., Mao, H., Gale, D., Barratt, J., Armour, J.A.L., Liu, J., Yu, X., 2016. Low α-defensin gene copy number increases the risk for IgA nephropathy and renal dysfunction. Science Translational Medicine 8, 345ra88.

Aldred, P.M.R., 2004. Variation in Human Immune Response Genes (PhD Thesis). University of Nottingham.

Aldred, P.M.R., Hollox, E.J., Armour, J.A.L., 2005. Copy number polymorphism and expression level variation of the human α-defensin genes DEFA1 and DEFA3. Human Molecular Genetics 14, 2045–2052.

Almal, S.H., Padh, H., 2012. Implications of gene copy-number variation in health and diseases. Journal of Human Genetics 57, 6–13.

Armour, J.A.L., Palla, R., Zeeuwen, P.L.J.M., Heijer, M. d., Schalkwijk, J., Hollox, E.J., 2007. Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. Nucleic Acids Research 35, e19–e19.

Armour, J.A.L., Sismani, C., Patsalis, P.C., Cross, G., 2000. Measurement of locus copy number by hybridisation with amplifiable probes. Nucleic Acids Res 28, 605–609

Bals, R., 2000. Epithelial antimicrobial peptides in host defense against infection. Respiratory research 1, 141.

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

Black, H.A., 2014. Allelic structures and mechanisms of copy number change at the human DEFA1A3 copy number variable locus (PhD Thesis). University of Nottingham.

Black, H.A., Khan, F.F., Tyson, J., Armour, J.A., 2014. Inferring mechanisms of copy number change from haplotype structures at the human DEFA1A3 locus. BMC genomics 15, 614.

Carter, N.P., 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. Nature Genetics 39, S16–S21.

Cekici, A., Kantarci, A., Hasturk, H., & Van Dyke, T. E. (2014). Inflammatory and immune pathways in the pathogenesis of periodontal disease. Periodontology 2000, 64(1), 57–80.

Chung, W.O., Dommisch, H., Yin, L., Dale, B.A., 2007. Expression of defensins in gingiva and their role in periodontal health and disease. Current pharmaceutical design 13, 3073–3083.

Cunliffe, R., Rose, F., Keyte J., Abberley, L., Chan, W., Mahida, Y., 2001. Human defensin 5 is stored in precursor form in normal Paneth cells and is expressed by some villous epithelial cells and by metaplastic Paneth cells in the colon in inflammatory bowel disease. Gut 48, 176-185.

Dale, B. A., & Fredericks, L. P. (2005). Antimicrobial Peptides in the Oral Environment: Expression and Function in Health and Disease. Current Issues in Molecular Biology, 7(2), 119–133.

de Smet, K., Contreras, R., 2005. Human Antimicrobial Peptides: Defensins, Cathelicidins and Histatins. Biotechnology Letters 27, 1337–1347.

Dolińska, E., Skurska, A., Pietruska, M., Dymicka-Piekarska, V., Milewski, R., Pietruski, J., & Sculean, A. (2017). The Effect of Nonsurgical Periodontal Therapy on HNP1-3 Level in Gingival Crevicular Fluid of Chronic Periodontitis Patients. Archivum Immunologiae et Therapiae Experimentalis, 65(4), 355–361.

Elliott, D.E., Siddique, S.S., Weinstock, J.V., 2014. Innate Immunity in Disease. Clinical Gastroenterology and Hepatology 12, 749–755.

Ericksen, B., Wu, Z., Lu, W., Lehrer, R.I., 2005. Antibacterial Activity and Specificity of the Six Human α-Defensins. Antimicrob Agents Chemother 49, 269–275.

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

Freeman, J.L., 2006. Copy number variation: New insights in genome diversity. Genome Research 16, 949–961.

Fujii, G., Selsted, M.E., Eisenberg, D., 1993. Defensins promote fusion and lysis of negatively charged membranes. Protein Sci 2, 1301–1312.

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

Ganz, T., 1999. Defensins and Host Defense. Science 286, 420.

Gonzalez, E., 2005. The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. Science 307, 1434–1440.

Greer, A., Zenobia, C., Darveau, R.P., 2013. Defensins and LL-37: a review of function in the gingival epithelium. Periodontology 2000 63, 67–79.

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

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., Ira, G., 2009. Mechanisms of change in gene copy number. Nature Reviews Genetics 10, 551–564.

Hazlett, L., Wu, M., 2011. Defensins in innate immunity. Cell and Tissue Research 343, 175–188

Hegele, R.A., 2007. Copy-number variations and human disease. The American Journal of Human Genetics 81, 414–415.

Hollox, E.J., 2017. Analysis of Copy Number Variation Using the Paralogue Ratio Test (PRT). Genotyping: Methods and Protocols 127–146.

Ionita-Laza, I., Rogers, A.J., Lange, C., Raby, B.A., Lee, C., 2009. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics 93, 22–26.

Ji, S., Choi, Y., 2013. Innate immune response to oral bacteria and the immune evasive characteristics of periodontal pathogens. Journal of Periodontal & Implant Science 43, 3.

Khan, F.F., 2012. Characterization of the alpha defensin copy number variation in humans (PhD Thesis). University of Nottingham.

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

Khurshid, Z., Naseem, M., Yahya I. Asiri, F., Mali, M., Sannam Khan, R., Sahibzada, H., Zafar, M., Faraz Moin, S., Khan, E., 2017. Significance and Diagnostic Role of Antimicrobial Cathelicidins (LL-37) Peptides in Oral Health. Biomolecules 7, 80.

Laine, M.L., Crielaard, W., Loos, B.G., 2012. Genetic susceptibility to periodontitis. Periodontology 2000 58, 37–68.

Lehrer, R.I., Ganz, T., 2002. Cathelicidins: a family of endogenous antimicrobial peptides. Current Opinion in Hematology 9.

Levy, J.A., Scott, I.,Mackewicz, C., 2003. Protection from HIV/AIDS: the importance of innate immunity, Clinical Immunology, Volume 108, Issue 3, Pages 167-174, ISSN 1521-6616

Li, M., Foo, J.-N., Wang, J.-Q., Low, H.-Q., Tang, X.-Q., Toh, K.-Y., Yin, P.-R., Khor, C.-C., Goh, Y.-F., Irwan, I.D., Xu, R.-C., Andiappan, A.K., Bei, J.-X., Rotzschke, O., Chen, M.-H., Cheng, C.-Y., Sun, L.-D., Jiang, G.-R., Wong, T.-Y., Lin, H.-L., Aung, T., Liao, Y.-H., Saw, S.-M., Ye, K., Ebstein, R.P., Chen, Q.-K., Shi, W., Chew, S.-H., Chen, J., Zhang, F.-R., Li, S.-P., Xu, G., Shyong Tai, E., Wang, L., Chen, N., Zhang, X.-J., Zeng, Y.-X., Zhang, H., Liu, Z.-H., Yu, X.-Q., Liu, J.-J., 2015. Identification of new susceptibility loci for IgA nephropathy in Han Chinese. Nature Communications 6, 7270.

Lillard, J.W., Boyaka, P.N., Chertov, O., Oppenheim, J.J., McGhee, J.R., 1999. Mechanisms for induction of acquired host immunity by neutrophil peptide defensins. Proc Natl Acad Sci U S A 96, 651–656.

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

Linzmeier, R.M., Ganz, T., 2005. Human defensin gene copy number polymorphisms: Comprehensive analysis of independent variation in α- and β-defensin regions at 8p22–p23. Genomics 86, 423–430.

López, S., García, I., Smith, I., Sevilla, A., Izagirre, N., de la Rúa, C., Alonso, S., 2015. Discovery of copy number variants by multiplex amplifiable probe hybridization (MAPH) in candidate pigmentation genes. Annals of Human Biology 42, 485–493.

Lupski, J.R., Wise, C.A., Kuwano, A., Pentao, L., Parke, J.T., Glaze, D.G., Ledbetter, D.H., Greenberg, F., Patel, P.I., 1992. Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A. Nature Genetics 1, 29.

Mathews, M., Jia, H.P., Guthmiller, J.M., Losh, G., Graham, S., Johnson, G.K., Tack, B.F., McCray, P.B., 1999. Production of β-defensin antimicrobial peptides by the oral mucosa and salivary glands. Infection and immunity 67, 2740–2745.

McCormick, T.S., Weinberg, A., 2010. Epithelial cell-derived antimicrobial peptides are multifunctional agents that bridge innate and adaptive immunity: Epithelial cell-derived antimicrobial peptides in immunity. Periodontology 2000 54, 195–206.

McKinney, C., Merriman, M.E., Chapman, P.T., Gow, P.J., Harrison, A.A., Highton, J., Jones, P.B., McLean, L., O'Donnell, J.L., Pokorny, V., 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, 409–413.

Mullaney, J.M., Mills, R.E., Pittard, W.S., Devine, S.E., 2010. Small insertions and deletions (INDELs) in human genomes. Human Molecular Genetics 19, R131–R136.

Munz, M., Willenborg, C., Richter, G.M., Jockel-Schneider, Y., Graetz, C., Staufenbiel, I., Wellmann, J., Berger, K., Krone, B., Hoffmann, P., 2017. A genome-wide association study identifies nucleotide variants at SIGLEC5 and DEFA1A3 as risk loci for periodontitis. Human Molecular Genetics 26, 2577–2588.

Ouhara, K., Komatsuzawa, H., Yamada, S., Shiba, H., Fujiwara, T., Ohara, M., Sayama, K., Hashimoto, K., Kurihara, H., Sugai, M., 2005. Susceptibilities of periodontopathogenic and cariogenic bacteria to antibacterial peptides, β-defensins and LL37, produced by human epithelial cells. Journal of Antimicrobial Chemotherapy 55, 888–896.

Page, R.C., 1986. Gingivitis. Journal of Clinical Periodontology 13, 345–355.

Page, R. C. (1991), The role of inflammatory mediators in the pathogenesis of periodontal disease. Journal of Periodontal Research, 26: 230–242.

Patsalis, P.C., Kousoulidou, L., Männik, K., Sismani, C., Žilina, O., Parkel, S., Puusepp, H., Tõnisson, N., Palta, P., Remm, M., 2007. Detection of small genomic imbalances using microarray-based multiplex amplifiable probe hybridization. European Journal of Human Genetics 15, 162–172.

Perry, G.H., 2008. The evolutionary significance of copy number variation in the human and chimpanzee genomes. Arizona State University.

Preshaw, P.M., Taylor, J.J., 2011. How has research into cytokine interactions and their role in driving immune responses impacted our understanding of periodontitis?: Cytokine interactions and immune responses in periodontitis. Journal of Clinical Periodontology 38, 60–84.

Qi, Y., Zhou, X., Cheng, F., Zhang, H., 2016. Elevated Plasma α -Defensins (HNP1–3) Levels Correlated with IgA1 Glycosylation and Susceptibility to IgA Nephropathy. Disease Markers 2016, 1–7.

Ranzani, V., Rossetti, G., Panzeri, I., Arrigoni, A., Bonnal, R.J.P., Curti, S., Gruarin, P., Provasi, E., Sugliano, E., Marconi, M., De Francesco, R., Geginat, J., Bodega, B., Abrignani, S., Pagani, M., 2015. The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. Nature Immunology 16, 318–325.

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

Rice, W., Ganz, T., Kinkade, J. J., Selsted, M., Lehrer, R., & Parmley, R., 1987. Defensin-rich dense granules of human neutrophils. Blood, 70(3), 757-765

Schneider, J.J., Unholzer, A., Schaller, M., Schäfer-Korting, M., Korting, H.C., 2005. Human defensins. Journal of Molecular Medicine 83, 587–595.

Schutte, B.C., McCray, P.B., 2002. β-Defensins in Lung Host Defense. Annual Review of Physiology 64, 709–748. Selders, G.S., Fetz, A.E., Radic, M.Z., Bowlin, G.L., 2017. An overview of the role of neutrophils in innate immunity, inflammation and host-biomaterial integration. Regen Biomater 4, 55–68.

Selsted, M.E., Ouellette, A.J., 2005. Mammalian defensins in the antimicrobial immune response. Nature Immunology 6, 551–557. Shi, H.-Z., 2004.

Eosinophils function as antigen-presenting cells. Journal of Leukocyte Biology 76, 520–527.

Selsted, M.E., Harwig, S.S., 1989, Determination of the disulfide array in the human defensin HNP-2. A covalently cyclized peptide. J. Biol. Chem. 264: 4003-7.

Shastry, B.S., 2002. SNP alleles in human disease and evolution. Journal of human genetics 47, 561–566.

Shwan, N.A.A., Louzada, S., Yang, F., Armour, J.A.L., 2017. Recurrent Rearrangements of Human Amylase Genes Create Multiple Independent CNV Series: HUMAN MUTATION. Human Mutation 38, 532–539.

Stankiewicz, P., Lupski, J.R., 2010. Structural Variation in the Human Genome and its Role in Disease. Annual Review of Medicine 61, 437–455.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S.W., Tavare, S., Deloukas, P., Hurles, M.E., Dermitzakis, E.T., 2007. Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. Science 315, 848–853.

Stuppia, L., Antonucci, I., Palka, G., Gatta, V., 2012. Use of the MLPA Assay in the Molecular Diagnosis of Gene Copy Number Alterations in Human Genetic Diseases. International Journal of Molecular Sciences 13, 3245–3276.

Subirana, I., Diaz-Uriarte, R., Lucas, G., Gonzalez, J.R., 2011. CNVassoc: association analysis of CNV data using R. BMC medical genomics 4, 47.

Tang, Y.Q., Selsted, M.E., 1993, Characterization of the disulfide motif in BNBD-12, an antimicrobial beta-defensin peptide from bovine neutrophils. J. Biol. Chem. 268: 6649-.

Tang, Y.-Q., Yuan, J., Miller, C.J., Selsted, M.E., 1999. Isolation, Characterization, cDNA Cloning, and Antimicrobial Properties of Two Distinct Subfamilies of α-Defensins from Rhesus Macaque Leukocytes. Infection and Immunity 67, 6139–6144.

The 1000 Genomes Project Consortium, A global reference for human genetic variation, Nature 526, 68-74

The International HapMap Project, 2003, Nature 426, 789–796.

Walker, S., Janyakhantikul, S., Armour, J.A.L., 2009. Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs. Genomics 93, 98–103.

Weischenfeldt, J., Symmons, O., Spitz, F., Korbel, J.O., 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. Nature Reviews Genetics 14, 125–138.

Wimley, W. C., Selsted, M. E. and White, S. H. (1994), Interactions between human defensins and lipid bilayers: Evidence for formation of multimeric pores. Protein Science, 3: 1362–1373.

Yang, D., Biragyn, A., Hoover, D.M., Lubkowski, J., Oppenheim, J.J., 2004. Multiple Roles of Antimicrobial Defensins, Cathelicidins, and Eosinophil-Derived Neurotoxin in Host Defense. Annual Review of Immunology 22, 181–215.

Yang, D., Chen, Q., Chertov, O., Oppenheim, J., 2000. Human neutrophil defensins selectively chemoattract naive T and immature dendritic cells. Journal of leukocyte biology 68, 9–14.

Yu, X.-Q., Li, M., Zhang, H., Low, H.-Q., Wei, X., Wang, J.-Q., Sun, L.-D., Sim, K.-S., Li, Y., Foo, J.-N., Wang, W., Li, Z.-J., Yin, X.-Y., Tang, X.-Q., Fan, L., Chen, J., Li, R.-S., Wan, J.-X., Liu, Z.-S., Lou, T.-Q., Zhu, L., Huang, X.-J., Zhang, X.-J., Liu, Z.-H., Liu, J.-J., 2011. A genome-wide association study in Han Chinese identifies multiple susceptibility loci for IgA nephropathy. Nature Genetics 44, 178–182.

Zhang, X., Ding, L., Sandford, A.J., 2005. Selection of reference genes for gene expression studies in human neutrophils by real-time PCR. BMC molecular biology 6, 4.

Zhang, L.J., Gallo, R.L., 2016, Antimicrobial peptides, Current Biology, Volume 26, Issue 1, Pages R14-R19, ISSN 0960-9822,

Zhang, F., Gu, W., Hurles, M.E., Lupski, J.R., 2009. Copy Number Variation in Human Health, Disease, and Evolution. Annual Review of Genomics and Human Genetics 10, 451–481.