Phenotypic Consequences of β-Defensin Copy Number Variation in Humans

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Razan Abujaber MPH

Department of Genetics

University of Leicester

2016

Abstract

Phenotypic Consequences of β-Defensin Copy Number Variation in Humans

Razan Abujaber

Beta defensins (DEFB) at the 8p23.1 genomic location are multifunctional secreted short peptides that have antibacterial and antiviral action in many species and possess immune cell signal activity, constituting a link between innate and adaptive immunity. In humans, the β-defensin region is known to be copy number variable (CNV) and contains seven genes repeated as a block, with a diploid copy number between 1 and 12.

This thesis shall explore the structural variability of the β-defensin CNV region; compare and contrast the different methods used for calling DEFB CNVs and investigate the role of CNVs of DEFB in various diseases. One of its aims is to also develop a model system to investigate if DEFB expression levels differ with CN in response to treatment with Pneumolysin by using Normal Human Bronchial Epithelial (NHBE) cells.

Results from this thesis confirm that the DEFB CNV region is 322kb in length, with a polymorphic inversion that occurs at a prevalence of 30% at the 8p23.1 genomic location that is independent of the DEFB CN. Paralogue Ratio Test (PRT) proved to be the best method of genotyping DEFB CNV especially in larger cohorts. In addition, work from this thesis also founded the basis of developing an *in vitro* model system to investigate whether DEFB expression levels differ with CN in response to treatment with pneumolysin by using Normal Human Bronchial Epithelial (NHBE) cells. As far as case/control and cohort studies are concerned, results from this thesis show that DEFB CN is not associated with lung function in the general population and has no effect on patients with COPD and Asthma, nor does it support previous results that present an association between HIV viral load and DEFB CN. DEFB CN was also found not to be associated with recurrent UTIs in VUR patients, nor with hypertension. Data suggested that DEFB CN might be associated with BMI but this has not been reproduced in a smaller cohort.

Acknowledgments

I gratefully acknowledge the funding received towards my PhD from The Royal Hashemite Court. Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without their financial support and the guidance I received from many people.

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Edward Hollox for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor.

Thanks to all the members in lab G3; Linda, Jenny, Shamik, Ezgi, Ade and Lee for the support, useful advice and lab fun times! I would also like to thank my APG thesis committee Prof. Yuri Dubrova and Dr. Richard Badge for going carefully through my first, second and third year reports giving useful suggestions regarding my project.

A big thank you goes out to my dear friends Pille, Chiara, Pierpaolo, Raheleh, Dan, Laura, Giorgio, Barbara, Mika, Angelica, Gurdeep, Mahdiyeh, Amin, Carmen and Basma for all the love, continuous support, coffee breaks, and wonderful nights out. A great appreciation goes to Firas who pushed me towards a PhD and throughout supported me in the good, bad and ugly. His help and constant motivation made me realise the sky is the limit!

Last but definitely not least is a special thanks to my parents; Mutaz and Suzanne who made all this possible with their great emotional and financial support, positivity, visits, believing in me and attempting to understand what I do…☺ . I will always be grateful to them for everything they have done for me. Thanks my lovely siblings Faisal and Sarah for being you and there whenever I felt homesick.

Table of Contents

List of Tables

List of Figures

Abbreviations

µl	microlitre
aCGH	array-Comparative Genomic Hybridisation
AMPs	Antimicrobial peptides
APCs	Antigenic presenting cells
BAC	Bacterial Artificial Chromosome
bp	Base pairs
BSA	Bovine Serum Albumin
cDNA	complementary DNA
CEPH	Centre de'Etude du Polymorphisme Humain
CF	Cystic Fibrosis
CHAVI	The Centre for HIV-AIDS Vaccine Immunology
CHD	Congenital heart disease
CMT1A	Charcot-Marie-Tooth disease type 1A
CN	Copy number
CNV	Copy Number Variation
COPD	Chronic obstructive pulmonary disease
Cp	Crossing point
DC	Dendritic cells
ddNTPs	Dideoxy Nucleotides Triphosphates
ddPCR	Digital droplet polymerase chain reaction
DEFB	human β-defensin
DNA	Deoxyribonucleic Acid
dNTPs	Deoxynucleotides
ECACC	The European Collection of Authenticated Cell Cultures
ELF	Bronchial epithelial lining fluid
FEV1	Forced expired volume in 1 second
FISH	Fluorescent in situ hybridization
FoSTeS	Fork-stalling and template switching
FVC	Forced vital capacity
gSTRip	Genome Structure In Population
GWAS	Genome wide association studies
HAART	Highly active antiretroviral therapy
hBD-1	human β-defensin 1
hBD-2	human β-defensin 2
hBD-3	human β-defensin 3
HERV-K	Human endogenous retrovirus - K family
HIV	Human Immunodeficiency Virus
HNPP	Hereditary neuropathy pressure palsy
IAVI	International AIDS Vaccine Initiative
IL-1β	Interleukin 1 Beta

Kb	Kilobase
LCR	Low copy repeats
LoH	Loss of heterozygosity
LPS	Lipopolysaccharide
LRCs	Low copy repeats
LSVs	Large-scale structural variations
LTR	Long terminal repeat
MAPH	Multiplex amplifiable probe hybridization
Mb	Megabase
mCNV	Multiallelic copy number variation
MLPA	Multiplex ligation-dependent probe amplification
NAHR	Non-allelic homologous recombination
NGS	Next-generation sequencing
NHEJ	Non-homologous end joining
OvDLP	Ornithorhynchus venom defensin-like peptide
PBS	Phosphate buffered saline
PC	Principal component
PCR	Polymerase Chain Reaction
PFIDO	Phase Free Inversion Detection Operator
PnCells	Pneumolysin treated cells
PRT	Paralogue ratio test
qPCR	Quantitative polymerase chain reaction
REPD	Repeat distal
REPP	Repeat proximal
RIVUR	Randomized Intervention for Children with Vesicoureteral Reflux
RNA	Ribonucleic Acid
ROMA	Representational Oligonucleotide Microarray Analysis
SCS	Silesian Cardiovascular Study
SD	Segmental duplications
SHCS	Swiss HIV Cohort Study
SNP	Single nucleotide polymorphism
SRD	Sequence read depth
TAP	Tracheal antimicrobial peptide
TNF-α	Tumour Necrosis Factor alpha
UTI	Urinary Tract Infection
VL	Viral Load
VUR	Vesicoureteral Reflux
YMCA	Young Men Cardiovascular Association

1 Introduction

1.1 Copy Number Variation

The human genome consists of 6 billion nucleotides of DNA that are packed into 23 pairs of chromosomes, one of each pair inherited from each parent. There are about 20,000 human protein-coding genes. The estimate of the number of genes has been repeatedly revised down from initial predictions of 100,000 or more as genome sequence quality and gene finding methods have improved, and could continue to decrease further (Pennisi, 2012; International Human Genome Sequencing Consortium, 2004). Protein-coding sequences account for about 1.5% of the genome and the rest is associated with non-coding RNA molecules, regulatory DNA sequences, LINEs, SINEs, introns, and sequences for which as yet no function has been elucidated (Lander et al., 2001).

The diversity in polymorphic genetic variation among humans is majorly accounted for by copy number variants (CNVs) (Sebat et al., 2004) contributing to the differences between individual humans (Hastings et al., 2009). CNVs also play an important role in genetic susceptibility to common diseases. Human genetic variation is the genetic diversity or variation in alleles of genes of humans and represents the total amount of genetic diversity within the human genome at both the individual and the population level (Conrad et al., 2010; Sudmant et al., 2010; Zhang et al., 2009). Recent studies have stated that variations exist in the human genome at diverse levels: large-scale microscopically visible chromosome anomalies; several kilobases to megabase pairs, submicroscopic copy number variation of DNA segments (tens to thousands of kb pairs) and the single base pair (bp).

Deletions, insertions and duplications of DNA segments ranging from several kilobases (Kb) to megabases (Mb) in size at variable number, in comparison with a reference genome are collectively referred to as CNVs (Conrad et al., 2010). A CNV can be simple tandem duplication, or may involve complex gains or losses of homologous sequences at multiple sites in the genome (Figure 1). It has been reported that copy number

varies in different organs and across tissues in the same individual and can arise both meiotically and somatically (Piotrowski et al., 2008).

The simplest form of CNV in the human genome could arise due to deletion or duplication of a gene. A diploid genome consists of two copies of a particular gene, one on each chromosome. Copy number can be categorized into diallelic and multiallelic groups. Diallelic CNVs have two alleles and could produce three different genotypes in both deletion and duplication events. A simple deletion event may change the diploid copy number of a specific gene and hence could result in diploid copy number of two, one or zero. Likewise a diploid genome could thus comprise two, three, or four copies of a gene after a simple duplication event in the genome. However, the deletion and/or duplication events in the genome do not always follow a simple pattern, ultimately causing complex CNVs, known as multiallelic copy number variants (Wain et al., 2009). Multiallelic CNVs (mCNVs) have more than two alleles and could produce more than three genotypes. In general, the size of genomic segments of deletion and duplication regions can vary from a few hundred to several million bp and could contain an entire gene, part of a gene, a region outside of a gene, or several genes in case of larger variants.

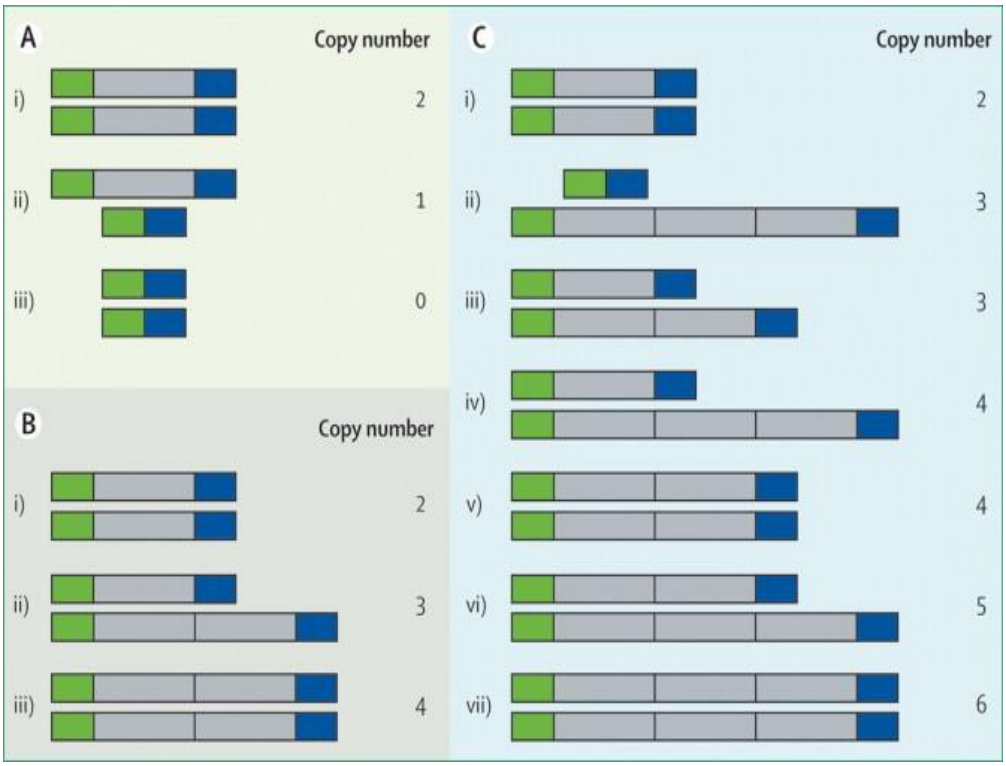


Figure 1: Diallelic locus (grey) and flanking loci (green and blue) with copy number variation caused by (A) deletion and (B) duplication, each showing the locus with (i) normal diploid copy number, (ii) heterozygous state, and (iii) homozygous state. (C) Multiallelic locus showing (i) normal copy number, (ii) multiple rounds of duplication on one chromosome and a deletion on the homologous chromosome, (iii) duplication on one chromosome and no deletion on the homologous chromosome, (iv) multiple rounds of duplication on one chromosome and no deletion on the homologous chromosome, (v) one round of duplication on each chromosome, (vi) one round of duplication on one chromosome and multiple rounds of duplication on the homologous chromosome, and (vii) multiple rounds of duplication on both chromosomes. Multiallelic assays measure total diploid copy number but cannot describe genotypes status of (ii) and (iii), or (iv) and (v). Reproduced from (Wain et al., 2009).

1.1.1 Classes of CNV

CNVs are classified into two classes, based on the mutational origin and molecular mechanism of their formation; “recurrent” and “non-recurrent” CNVs. The mutation rates are thought to be different for recurrent and non-recurrent CNVs (Hollox & Hoh, 2014).

Recurrent CNVs exist in areas within the genome that contain large segmental duplications (SDs) and are mostly created by non-allelic homologous recombination (NAHR) mechanism of CNV formation. 20-40% of normal polymorphic CNVs can be

classified as recurrent CNVs (Conrad et al. 2010). Recurrent CNVs can arise anywhere in the genome but hotspots for these CNVs mainly exist in subtelomeric and pericentromeric regions (Conrad & Hurles, 2007; Redon et al., 2006).

Non-recurrent CNVs on the other hand involve large genomic regions and break-point analysis shows minimal or no-homology is required for them to form (Conrad et al., 2010). Non-recurrent CNVs arise by non-homologous end joining (NHEJ) or fork-stalling and template switching (FoSTeS) mechanisms (Zhang et al., 2009). Non-recurrent CNVs are often large, are more likely to affect genes and hence more likely to have an extremely deleterious phenotypic effect (Arlt et al., 2012). According to Hollox & Hoh (2014), many non-recurrent CNVs are rare because negative selection acts to rapidly remove the deletion from the population.

1.1.2 Prevalence of CNVs in the human genome

The availability of a complete human genome sequence and significant advances in microarray technologies made it possible to obtain genome-wide maps of approximate locations and frequencies of CNVs. In 2004, two independent research groups investigated the human genome for CNVs. Sebat et al., employed Representational Oligonucleotide Microarray Analysis (ROMA) technology in investigating large-scale (>100 kb) CNVs in 20 healthy individuals with 85,000 probes that were 35 kb apart. The result of their studies showed that 221 CNVs were located at 76 CNV loci. Iafrate et al., on the other hand, used Bacterial Artificial Chromosome (BAC) Comparative Genomic Hybridization (CGH) array with approximately 1 Mb resolution where 55 individuals were investigated. 255 clones in the study revealed CNV, 41% of which were present in more than one person, showing that these variations are polymorphic.

In 2006, Redon et al. studied 270 lymphoblastoid cell lines from the International HapMap project that was established from people of African, European, and Asian ancestry. In their study, they realised 1447 CNV regions occupying a sequence of 360 Mb in size. It was through their finding that CNVs was concluded to occupy 15% of the human genome. It was also discovered that an average of 12 CNVs exist in each person compared to a reference genome (Li & Olivier, 2013). According to Li and Olivier, using

a lower limit of 1 kb to define CNVs is discretionary because of the resolution difficulties. Therefore, a change in the threshold setting can radically change the number of CNVs reported.

The controversy of CNV was however summarized by Zhang et al., in 2009 who claimed that approximately 30% (38,406 genomic variants) of the human genome is covered by CNV. In addition to this, they asserted that CNV is a DNA quantitative variation that exceeds 100bp. However, this value may have been over-exaggerated because of the technology's resolution limits in screening for CNVs resulting in high rate of false positives in small sized CNV calls. On the contrary, inexactness in determining CNVs varying between 1 – 20 kb could have contributed to underestimation of the total number of CNVs. Also, the limitations of the current reference genome must be considered when validating the data.

Conrad and colleagues in 2010 designed an experimental strategy to discover CNVs greater than 500 base pairs using a set of 20 NimbleGen arrays, each comprising 2.1 million long oligonucleotide probes covering the assayable portion of the genome across 40 individuals. They identified 51,997 CNV calls, 11,700 CNV loci and 8,599 validated CNVs, 5,238 loci of which were genotyped allowing to distinguish deletions (0, 1 or 2 diploid copy number), duplications (2, 3 or 4 diploid copy number) and mCNVs (greater than 3 possible diploid copy numbers). This data set has been the core scientific resource on common CNVs for years (Conrad et al., 2010).

The reference genome validation and size limitation mentioned above has been tackled in 2015 by Sudmant and colleagues. In an attempt to understand the pattern, selection, and diversity of CNVs in the context of the ancestral human genome, Sudmant et al. (2015) sequenced 236 individual genomes from 125 different human populations and identified 14,467 autosomal CNVs and 545 X-linked CNVs, using a sequence read-depth approach which provided breakpoint resolution to 210 bps. They found that the median size of CNV was 7,396 bp, with 82.2% of events less than 25 kbp. CNVs mapping to SDs were larger on average (median = 14.4 kbp) than CNVs mapping to the unique parts of the genome (median = 6.2 kbp). Around 50% of CNV base pairs mapped within previously annotated SDs. In total, 7 % of the human genome is variable because of CNVs, in contrast to around 1 % resulting from single-

nucleotide variations. Duplications were more common (4.4 % of the genome) compared with deletions (2.8% of the genome). When comparing their data set to Conrad et al. (2010), 67 - 73 % of calls were exclusive to their study, whereas they captured 68 – 77 % of formerly identified CNVs (Sudmant et al., 2015).

1.1.3 Functional consequences of CNVs

The functional importance of many CNVs is relatively clear; reduced copy number of a gene can be correlated with reduced expression level, while duplicated copies of a gene can lead to increase expression level (McCarroll & Altshuler, 2007; Stranger et al., 2007). 85% - 95% of CNVs in human and mice were reported to be associated with a change in expression of the affected genes (Stranger et al., 2007).

The majority of the genes with CNVs play a part in the immune system, brain development and brain functioning. CNVs were firstly linked to human diseases in the 1980s. However; their population incidence was assumed to be not only small but also directly related to certain genomic disorders (Freeman et al., 2006; Ghanem et al., 1988). For instance, CNV at the α-globin locus was claimed to be the causing agent of α-Thalassaemia (Goossens et al., 1980). CNVs may also impair the performance of adjacent regulatory signals that trigger or mute genes without directly influencing the copy number of that gene itself (Cahan et al., 2009; Henrichsen et al., 2009).

CNVs seem to be critical for evolution; some CNV copies sustain their original function whereas paralogues undergo rapid adaptive evolution to specialize in their functional niche (Inoue & Lupski, 2002). In certain instances, more copies of certain CNV genes offer a selective advantage such as, the *AMY1* gene. *AMY1* gene encodes the salivary amylase enzyme. Populations whose diet constitutes high levels of starch have significantly higher average copy number of *AMY1* (Perry et al., 2007). An example of a reduction in copy number being important has been proposed for the α-globin locus. The disorders of α -globin gene deletion in homozygotes, for example α thalassemia, might be stabilized by resistance to malaria for heterozygotes (Higgs et al., 1989).

Although most CNVs are benign in nature, some may have an effect on gene expression, affecting the phenotype through not only disrupting genes, but also changing gene dosage (Dereli-Oz et al., 2011). Copy number changes are also involved in the formation as well as progression of cancer (Shlien & Malkin, 2009; Volik et al., 2006). In support of this, Frank et al., (2007) argued that copy number contributes to cancer proneness. Apart from causing cancer, CNVs increase susceptibility to schizophrenia (Ahn et al., 2014), epilepsy (Bassuk et al., 2013; Mefford et al., 2010), autism (Polan et al., 2014; Marshall & Scherer, 2012), Psoriasis (Hollox et al., 2008), and HIV (Hardwick et al., 2012; Larsen et al., 2012; Liu et al., 2010). Several CNV genes are involved in some known metabolising enzymes, such as *CYP2D6* and *GSTM1*. Others are widely studied such as the β-defensins at the 8p23.1 genomic location due to their potential clinical relevance for innate immunity, inflammation, and cancer (Groth et al., 2008). While others are potential drug targets such as *CCL3L1*, which may also make significant contributions to pharmacogenomic studies (Ouahchi et al., 2006).

In 2015, Handsaker et al. sought to use whole-genome sequence data to deepen their understanding regarding mCNVs. mCNVs are loci that exist in more states that can be explained by the segregation of just two structural alleles (Handsaker et al., 2015). They analysed 849 genomes sequenced by the 1000 Genomes project to identify most large (>5 kb) mCNVs, including 3878 duplications, of which 1356 appeared to have 3 or more segregating alleles. Handsaker at al., 2015 discovered that mCNVs give rise to most human variation in gene dosage – 7 times the combined contribution of deletions and bi-allelic duplications, and this in turn generates abundant variation in gene expression.

According to recent CNV studies, the human genome has CNV distribution hotspots (Sudmant et al., 2015). The hotspots are structural variant-rich such as the immunity and cell-cell signalling genes and genes that encode proteins which are involved in interaction with the environment. Another hotspot includes genes that code for retroviruses as well as transposition related proteins (Li & Olivier, 2013).

1.2 CNV detection methods

As it was introduced above, CNVs result in several human genetic disorders. Each person's genome has several copy number polymorphisms of different sizes that are believed to contribute to phenotypic variation, as well as vulnerability to multifactorial disease. Thus, it is not surprising that a wide variety of laboratory methods have been developed to aid in the identification of copy number changes. To fully understand the role of genetic variation in disease, accurate and complete CNV measurements are needed. The precise detection of a gene's copy number faces many challenges such as detection of CNV regions and ability to distinguish between the different classes of CNVs. Other challenges include having robust detection algorithms for precise detection and full understanding of the mechanisms that create CNVs (Li & Olivier, 2013). However, various methods have been developed and optimized to study CNVs. They include fluorescence in situ hybridization (FISH), Southern Blotting, conventional karyotyping, microarray-based copy number screening, PCR-based methods, multiplex ligation-dependent probe amplification (MLPA), and comparative genomic hybridization (CGH) (Wain & Tobin, 2011). Each method is associated with specific advantages and disadvantages that determine the choice of which technique to use in a study. Conventional karyotyping enhances detection of structural variations in the entire genome, but its resolution is low (>5–10 Mb). FISH analysis, on the other hand, requires metaphase chromosomes or interphase nuclei with a resolution of approximately 100kb.

1.2.1 Quantitative PCR (qPCR)

In 1983, Polymerase Chain Reaction (PCR) was invented and used extensively within three years after its invention (Oswald, 2007). The principle of PCR technology was to increase a target from minute amount of starting material. At the end of amplification, the product was run on a gel for detection of the specific product. The introduction of Quantitative Real-Time PCR (qPCR) reduced all these processes because the technology had the potential to combine the DNA amplification step and enhancing immediate detection of the desired products in one tube. qPCR is a technique that is

extensively used currently in microbial ecology to describe gene or transcript numbers that exist in environmental samples. This detection method is based on changes in fluorescence which correlates with the increase of target. Fluorescence is monitored throughout each PCR cycle giving an amplification plot that allows the user to track the reactions in real time. The build-up of PCR product which is calculated in the real time results is shown as a sigmoidal amplification curve. Many detection chemistries exist to measure product accumulation. The chemistries include hydrolysis probes, dual hybridization probes, molecular beacons and double-stranded DNA specific binding dyes.

The target certainty of qPCR is evaluated by an internal probe that allows quantification of functional gene markers that exist in an individual. Through the use of the qPCR technique, the gene of interest is instantaneously amplified as well as quantified in real time. qPCR primarily involves using fluorescent techniques such as TaqMan where the threshold cycles (Ct) between the target gene and a reference gene is compared. The generated ΔCt values are then used to determine the copy number. qPCR is normally used as a justification technique for computationally identified loci (Li & Olivier, 2013). There is a connection between the period that the fluorescent PCR signal rises above the background and the earlier amount of input DNA. A large quantity of input material leads to lower crossing point (Cp) values. The Cp value symbolizes the fractional PCR cycle that is typical for the magnification curve or at which the fluorescence exceeds a certain threshold.

The advantages of qPCR are many over the majority of the alternative methods. For instance its instrumentation costs are minimal (Karlen et al., 2007). The qPCR approaches also bring together the identification of target prototype with quantification by recording the magnification of a PCR product through a conforming rise in the fluorescent signal linked to product formation during every cycle in the PCR. In relation to this description, qPCR that employs fluorescence-based detection in a study provides greater sensitivity. It also allows discrimination of gene numbers along a wide dynamic range (Karlen et al., 2007; Rutledge & Cote, 2003). For instance, the use of qPCR enables discrimination of twofold changes in a given target concentration. Apart from being rapid and straightforward, qPCR allows for the detection of tiny

duplications and deletions using as little as 5ng of DNA (Fernandez-Jimenez et al., 2011). The homogeneous format of the products in single tubes is very beneficial since it does away with the significant contamination risk associated with opening tubes for carrying out post-PCR manipulation (Karlen et al., 2007). This is in tandem with the fact that the reaction tubes can be observed and determined without opening the tubes. However, the number of samples that can be studied at any one time is limited by the number of available fluorophores and the instrument's detection capabilities.

The use of qPCR has been less frequently utilised in the determination of copy number. The overall accuracy of using the qPCR method in determination of CNVs depends on the number of qPCR assays. Determination of CNVs involves selection of genes or intergenic regions in which the number of the selected genes influences the number of assays in each gene (Rutledge & Cote, 2003). When a large number of genes need to be screened, the number of assays on each gene is often limited to one or two because of practical and financial demands. If a CNV in a given gene is associated with a certain disorder or phenotype, screening much more is advocated to ensure the identification of small deletions. qPCR is considered one of the keys to success after considering the cost as well as its resolution per sample (D'haene et al., 2010).

In their study; Haridan et al. (2015) suggested that qPCR could potentially introduce false-positive calls, therefore CNV association studies based on qPCR should be counter validated. Indeed, many studies showed contradictory findings on copy number and association with disease development, for instance *CCL3L1* in HIV (Walker, Janyakhantikul & Armour, 2009), and *DEFB4* in Crohn's disease (Aldhous et al., 2010; Fellermann et al., 2006). The disadvantage of qPCR though is common across copy number assays and this is expected in view of the principle of CNV and the chemistry of qPCR amplification. The absolute values also vary usually upon repetition since qPCR is technically demanding. This again is a common phenomenon observed in most qPCR assays (Willcocks et al., 2008).

1.2.2 Multiplex Ligation-dependent Probe Amplification (MLPA)

Multiplex ligation-dependent probe amplification (MLPA) was first introduced in 2002 by Schouten et al. (2002) and has been extensively applied in a variety of clinical and research situations (Schouten et al., 2002). The technique has proven to be efficient and reliable for copy number variation detection and validation (Hills et al., 2010; Pedersen et al., 2010; Janssen et al., 2005).

In MLPA, two sequence-tagged half probes are annealed to adjacent sites on the genomic target sequence and ligated using a thermostable DNA ligase. The ligated probes are then amplified with universal PCR primers, one of which is fluorescently labelled, and quantified using electrophoresis. Each amplicon has a different size which allows identification of specific DNA fragments. The amount of ligated probe is proportional to copy number of target gene and can be quantified after running the ligated PCR products on capillary electrophoresis (Figure 2b). In each MLPA experiment, reference probes are also put in probe mixes to determine unknown copy number. The reference probes are assumed to have a copy number (n=2) in both test samples and control samples. The reference probes are designed from chromosomal regions which are non-variable. Groth and co-workers estimated beta-defensin copy number in 44 different samples using MLPA technique and a noticeable correlation was observed with other techniques, such as qPCR and PRT (Groth et al., 2008). MLPA detects copy number variation of maximum 45 separate genomic sequences in a single reaction using relatively small amounts of starting DNA (20 ng) and does not require cells for chromosome spreads. MLPA assay can be used to target any genomic sequences for copy number analysis, irrespective of their size or proximity to each other. MLPA allows more accurate determination of the size of deletions or duplications in comparison to FISH or qPCR (Janssen et al., 2005). MLPA is high throughput and results can be obtained within 20 hours.

There are difficulties in custom probe design for regions not yet available commercially as kits. A list of criteria (probe length, melting temperature, secondary structure, GC content, nucleotide composition at the ligation, site, sequence uniqueness, avoidance of known SNPs, etc.) is a pre-requisite to increase the probability of a successful MLPA

assay. Unknown SNP in the probe binding regions may affect MLPA results and appear as exon deletions.

1.2.3 Multiplex Amplifiable Probe Hybridization (MAPH)

Multiplex amplifiable probe hybridization (MAPH) is a PCR-based method of quantifying multiple genomic loci in a single reaction (Armour et al., 2000; Hollox et al., 2002). The technique is based on the quantitative amplification of multiple probes that have been hybridized to immobilized genomic DNA (Figure 2A). All the probes have universal primers at the ends to be amplified by a single PCR. MAPH probes are generated by cloning the target sequences into a plasmid vector, followed by PCR amplification of cloned sequence using primers directed to the vector, to have similar flanking sequence in all PCR products. Probes have different length and identical tails facilitating PCR amplification with a single primer pair. Around 0.5–1 μg denatured genomic DNA is spotted onto a nylon filter and hybridized with a set of probes corresponding to the target sequences. The membranes are then washed rigorously to remove unbound probe, and the remaining specifically bound probe is proportional to its target copy number. The probes are stripped from the membrane and amplified simultaneously with the universal primer pair and separated by electrophoresis. A relative comparison is made between the test and control probes based on band intensities, peak area/peak heights depending on the detection method. Reduced band intensities or peak area/peak heights compared to internal control probes indicate deletion and increased band intensities or peak area/peak heights indicates duplication. Armour et al., multiplexed up to 40 probes in one single reaction and resolved by gel electrophoresis simultaneously (Armour et al., 2000). MAPH was used to measure β-defensin copy number (Armour et al., 2000; Hollox et al., 2002).

The designing and establishing of probes for MAPH is far simpler than the MLPA probe generation. MAPH works with double-stranded DNA probes that are obtained from cloning or PCR. SNPs in the probe binding regions are unlikely to affect MAPH but if part of a region targeted by a MAPH probe is deleted, the probe may still hybridise and the target will be scored as being present. The washing steps in the MAPH technique which is essential to remove unbound probe, may also introduce a contamination risk.

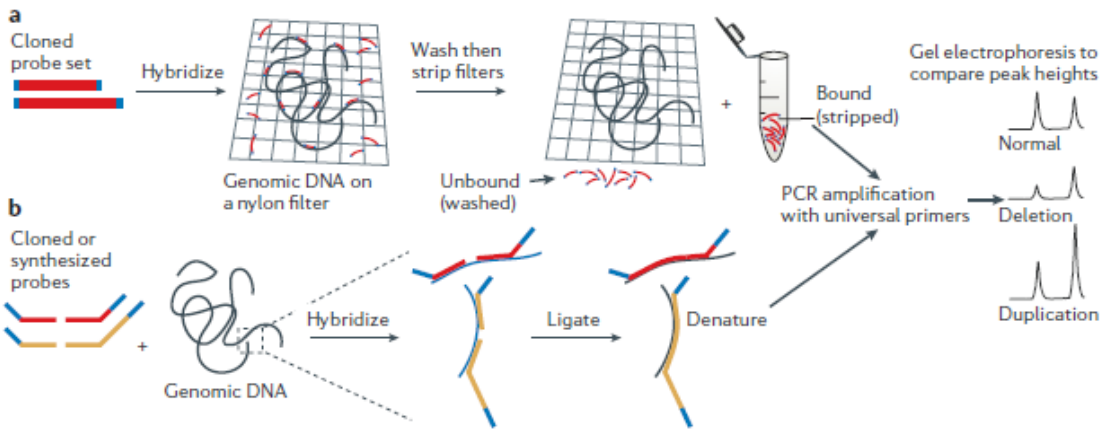


Figure 2: Multiplex PCR-based methods for the identification of copy-number variants. A) In multiplex amplifiable probe hybridization (MAPH), the probes (red) of different sizes are normally cloned into vectors and amplified by PCR such that each end is flanked by the same sequence site (blue). The genomic DNA is fixed to a membrane and probes are hybridized to it. Unbound probes are removed after rigorous washing and the probes are stripped from the membranes. The amount of probe present at this stage is proportional to its copy number in the target genomic DNA. Probes are amplified by a universal primer pair and size-separated by gel electrophoresis. Changes in peak heights relative to controls DNA (non-CNV), indicate the copy number. B) Multiplex ligation-dependent probe amplification (MLPA) for each target region 2 probes are designed, which hybridize adjacent to each other (probes for 2 regions are shown in red and yellow). Like MAPH, all probe pairs are flanked by universal primer sites (blue). The probes are hybridized to genomic DNA and adjacent probes are ligated to join the two primers together. The number of ligated primers is proportional to the target copy number. After denaturation, the ligated probes amplify with PCR amplification. Sometimes 'stuffer' sequence is added with one of these probes as having a universal primer site, which allows each probe set to produce fragments of a different size. Size separation by gel electrophoresis is carried out as with MAPH to detect deletions and duplications. Reproduced from (Feuk et al., 2006).

1.2.4 Paralogue Ratio Test (PRT)

The paralogue ratio test (PRT) is a comparative PCR-based technique that was developed by Armour et al. in 2007 (Armour et al., 2007) . In this technique, one set of a primer pair co-amplifies a 'test' region which is copy number variable and a 'reference' region which is not variable in copy number using PCR. It is regarded an accurate, besides a relatively high-throughput method, for identifying gene copy number at a single loci. The use of an identical primer pair makes the amplification efficiency for both the test and reference loci similar enhancing reproducibility. Since the resulting amplicons differ slightly in terms of size, they can be easily differentiated by capillary electrophoresis. The strategy of PRT reduces the problems that occur as a result of comparison of dissimilar amplicons with different amplification efficiencies. Experiments can be performed in duplicate by using two different fluorescent dyes to label the same primer, and then run both products on the same capillary.

One of the main challenges of this method is primer design. The primer must anneal to only the reference locus and copy number variable locus. This can be achieved with the help of an algorithm, which can quickly design couple of primers, suitable for PRT methods. The algorithm blasts the region of interest with the entire genome sequence, masking repeated regions, in order to find specific and unique paralogous regions and in combination with primer design software, selects the oligos annealing only for those (Veal et al., 2013).

An estimation of the integer copy number is achieved by combining all raw values for each assay and is calculated using maximum-likelihood approach. An associated significance value portraying the confidence of the typed copy number compared with all other copy numbers; between 1 and 10 is also given (Aldhous et al., 2010; Abu Bakar et al., 2009; Armour et al., 2007). Alongside the DEFB CN estimated using the ML approach, an alternative method to confer DEFB CN was employed by Walker et al., 2009; and adopted in this thesis; the weighted mean integer value (2.6.5)

The precision and accuracy of the PRT assay are equivalent to MLPA and MAPH. But PRT has the advantage of high-throughput analysis for CNV typing of large cohort effectively using small amount (10 ng) of DNA (Armour et al. 2007; Hardwick et al.

2014; Machado et al. 2013; Aklillu et al. 2013; Abu Bakar et al. 2009; Hollox et al. 2008; Wain et al. 2014; Hardwick et al. 2012; Hollox et al. 2008).

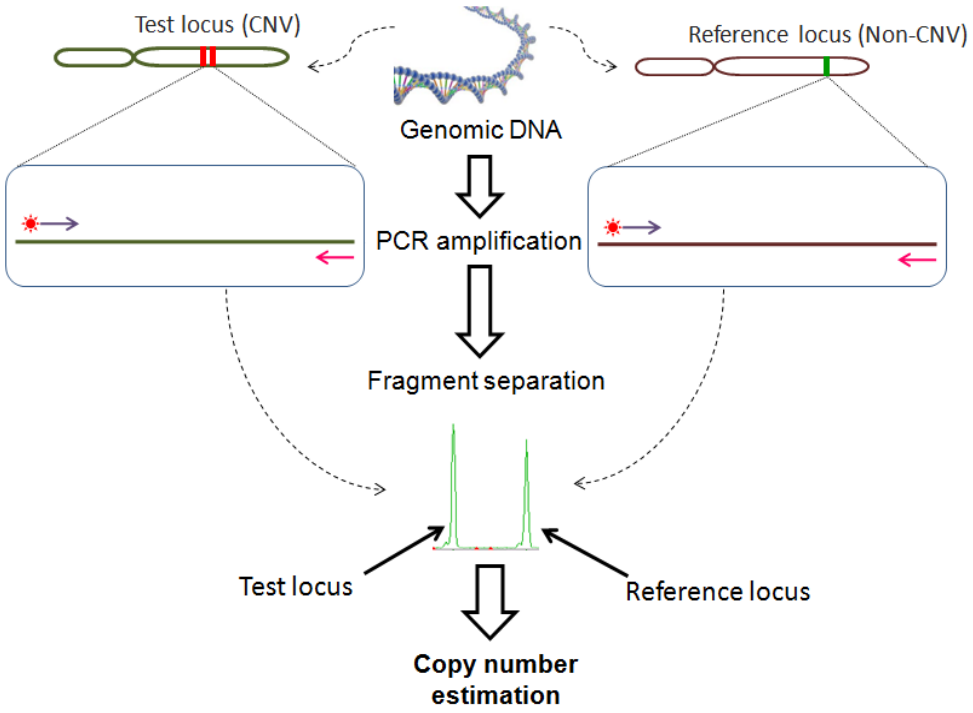


Figure 3: Schematic picture describing different steps of PRT. Both test (CNV) and reference regions are PCR amplified using florescent labelled primers. The amount of PCR products are quantified with capillary electrophoresis and copy number is estimated by comparing amount of test products with reference products.

1.2.5 Array Comparative Genomic Hybridization (aCGH)

Comparative genomic hybridisation (CGH) is a technique that allows the detection of changes in chromosomal copy number without the need for culturing cells. It gives a global overview of chromosomal gains and losses throughout the whole genome of a tumour. Thus, CGH is a comparatively fast screening technique that can point at specific chromosomal regions that might play a role in the pathogenesis or progression of tumours (Weiss et al., 1999). The development of commercial comparative genomic hybridization arrays (aCGH) platforms for calling copy number was firstly done by Agilent and NimbleGen. Basically, equivalent quantities of both the target and reference DNA are fluorescent labelled differentially, usually using Cyanine 3 (Cy3) and

Cyanine 5 (Cy5) and co-hybridized on a probe array. The created slide is scanned using a microarray scanner and the spot intensities are measured and analysed for copy number analysis (Ahn et al., 2013; Jaillard et al., 2010; Baris et al., 2007; Feuk et al., 2006) In cases where the concentrations of the fluorescent dyes correlate with one probe, the region of the patient's genome is claimed to have equal quantity of DNA in the test as well as the reference samples. If the ratio is altered, it indicates relative losses or gains in a target sample.

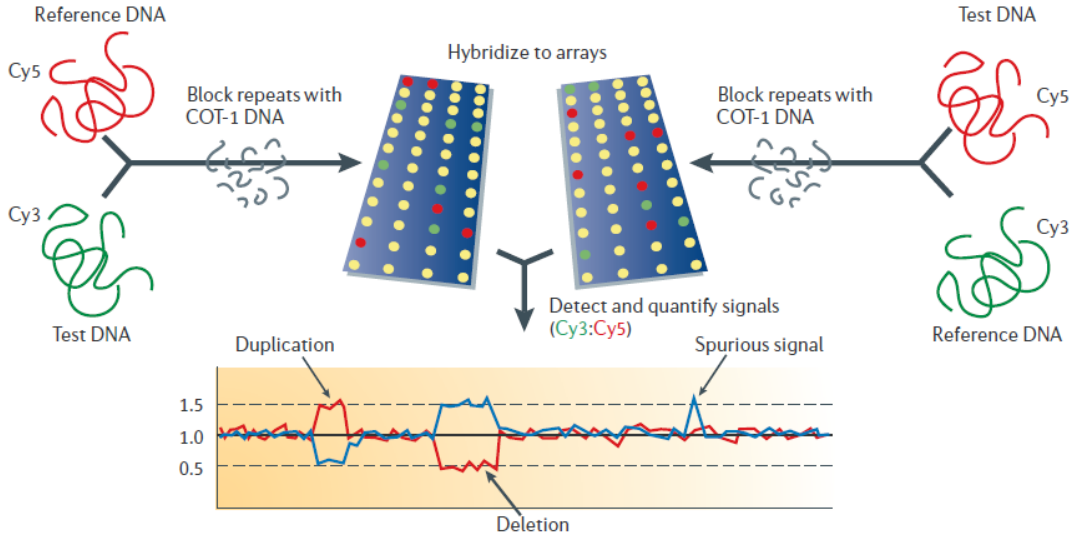


Figure 4: Schematic picture of array-based comparative genome hybridisation (array-CGH) adopted from (Feuk et al., 2006). The reference and test DNA samples are differentially labelled with fluorescent tags (Cy5 and Cy3, respectively), repetitive-elements are blocked using COT-1 DNA and then hybridised to genomic arrays. After hybridisation, the fluorescence ratio (Cy3:Cy5) reveals copy-number differences between the two DNA samples. Typically, in array-CGH, the initial labelling of the reference and test DNA samples reversed for a second hybridisation ('dye-swap') (left and right sides of the panel) to detect spurious signals. The red line represents the original hybridisation and the blue line represents the reciprocal hybridisation.

The construction of probe arrays have included large insert clones like bacterial artificial chromosomes (BACs) that range in size from 40-200 kb, small insert clones such as cosmids (38-45 kb) along with cDNA clones (0.5-2 kb). The construction also includes not only genomic PCR products (100bp-1.5kb), but also oligonucleotides (25-80bp) (Carter, 2007). The probe can be selected to represent the genome or for a

unique region allowing highest sensitivity and specificity, as summarized in the Table 1 below.

In addition to the length of the DNA probe, the resolution of the different aCGH platforms is determined by and gaps between them. The ability to identify copy number changes also depends on the signal-to-noise ratio (Carter, 2007).

Table 1: A comparison between the different probes used to construct CGH arrays for copy number calling

Probe type	Size (kb)	Advantage	Disadvantage
BACs	80-200	High signal-to noise-ratio	Only large differences are detected (>50 kb)
Fosmid/Cosmid	40	High signal-to-noise ratio	Relatively large differences are detected (~30kb)
cDNA clones	0.5-2	High resolution	-Variable resolution cross the genome due to uneven distribution of genes -Signal-to-noise ratio is reduced due to mismatch hybridization between genomic DNA and cDNA
PCR products	Single/part gene	Complete coverage	-Poor signal-to-noise ratio -Probe generation is expensive
Oligonucleotides	25-80bp	Highest potential resolutions	Poor signal-to-noise ratio

1.2.6 nCounter

The nCounter Analysis System uses NanoString digital detection technology that detects and counts different types of molecules in one tube. This technological device uses colour-coded molecular barcodes to hybridise directly given set of molecules to different type of expected molecules (NanoString Technologies, 2013).

The nCounter Custom CNV Assay essentially requires the DNA to be fragmented as well as denatured so that it can yield single-stranded targets to be hybridised with nCounter probe pairs. nCounter probe pairs are comprised of a 'reporter probe' which carries the signal, and a 'capture probe' which allows the complex to be inactivated for data collection. After hybridization, samples are moved to the nCounter prep station where surfeit probes are removed and target complexes aligned and immobilized in the nCounter cartridge. The cartridges are then placed in the nCounter digital analyser

for collection of data. Every CNV probe pair is detected by the "colour code" produced by six ordered fluorescent spots that are present on the reporter probe. The reporter probes on the cartridge's surface are then counted and tabulated as shown in Figure 5 below:

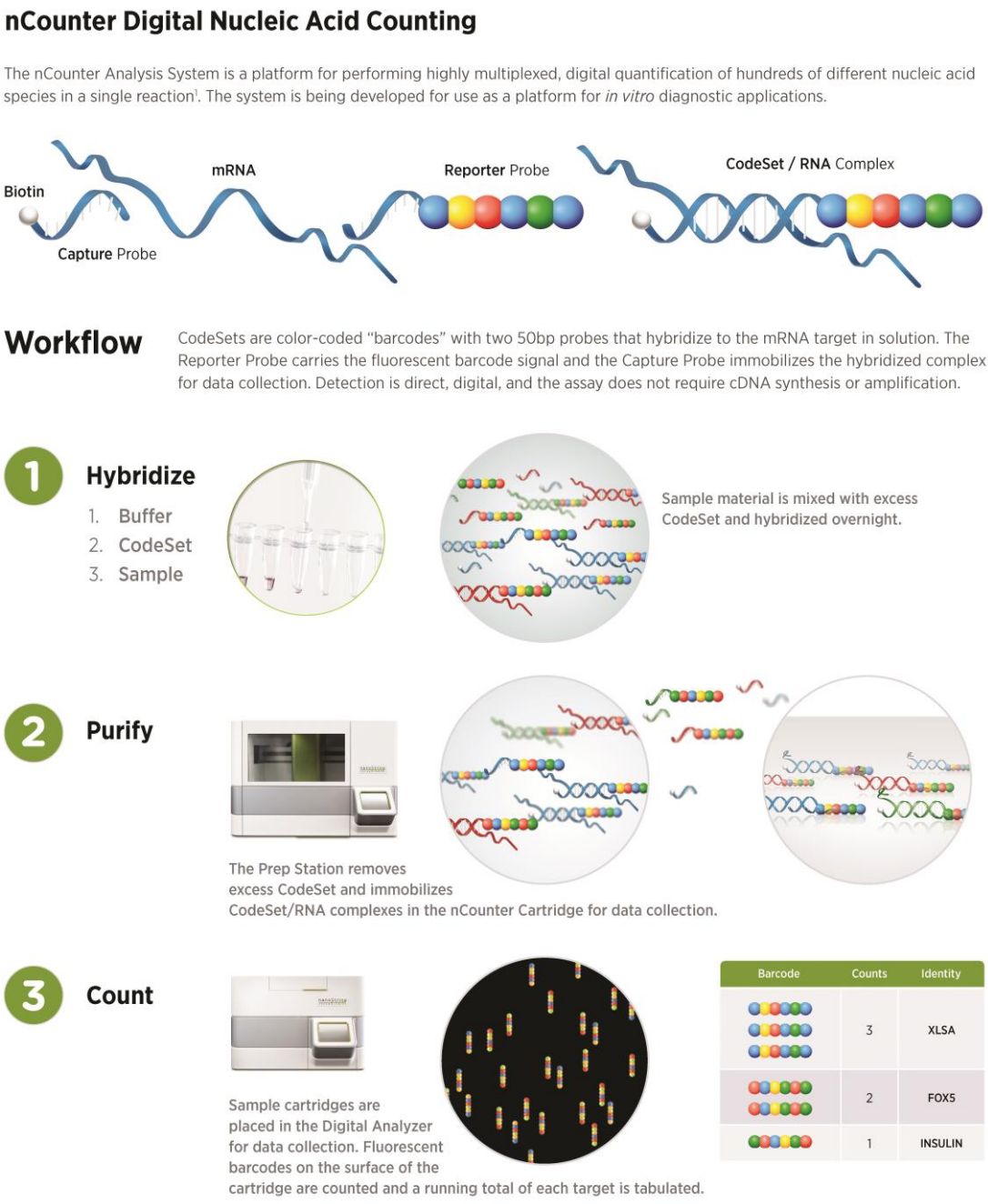


Figure 5: nCounter Custom CNV Assay workflow. Reproduced from (NanoString Technologies, 2013).

The nCounter system has many advantages that make it a viable method for detection of CNVs. Apart from not requiring amplification, it can analyse approximately 800

genes concurrently (NanoString Technologies, 2013). The flexible format of this system also allows interrogation of actual gene number. In addition to having high accuracy for CNVs, nCounter system can analyse 12 samples in 25 minutes (NanoString Technologies, 2013).

The applications of the NanoString nCounter technology are diverse; one example in which nCounter was used in CNV studies was to verify the detected risk of CNVs in a replication study which implied increased risk of congenital heart defects in Down syndrome patients (Sailani et al., 2013).

1.2.7 Digital Droplet PCR (ddPCR)

The main difference between ddPCR and traditional PCR lie in the method of quantifying nucleic acid amounts. Use of ddPCR also limits some challenges encountered when using qPCR. For instance, the technical limitation of qPCR includes the requirement for assay calibration similar in terms of quality standards with samples to be evaluated. Multiplexing assays sometimes is not straightforward because of the competition between assays. It also has a theoretical limit of quantification that is not enough for CNVs applications with heterogeneous materials (Karlen et al., 2007; Rutledge & Cote, 2003). Digital PCR was first described by Vogelstein & Kinzler in 1999. Single molecules were separated by dilution and each amplified by PCR. Each product was then analysed separately for either presence or absence of mutation in the *ras* oncogene using fluorescent probes (Vogelstein & Kinzler, 1999). At that time, ddPCR was used in pinpointing base substitution mutations, chromosomal translocations, and gene amplifications. It was also used to determine spliced products, changes in gene expression, allelic discrimination in addition to allelic imbalances (Vogelstein & Kinzler, 1999).

A newer version of ddPCR carries out a single multiplex reaction within a sample through use of differently labelled TaqMan probes for target and reference. 20µl of the sample studied is pipetted into the sample well of the 'droplet generator cartridge'. Droplet generation oil is then added into the specified well before a vacuum is applied to the outlet wells. The vacuum draws the sample and oil via a flow-focusing

junction in which 20,000 droplets are produced and the reaction proceeds in each droplet individually.

After partitioning, the sample is moved into a 96 well plate and wrapped with heated foil before thermal cycling it into an endpoint. Subsequently, the plate is shifted to a droplet reader where every droplet from every well is aspirated and gushed to the detector. On the way to the detector, an injection of a spacer fluid divides as well as aligns the droplets for single-file simultaneous two-colour detection (Hindson et al., 2011).

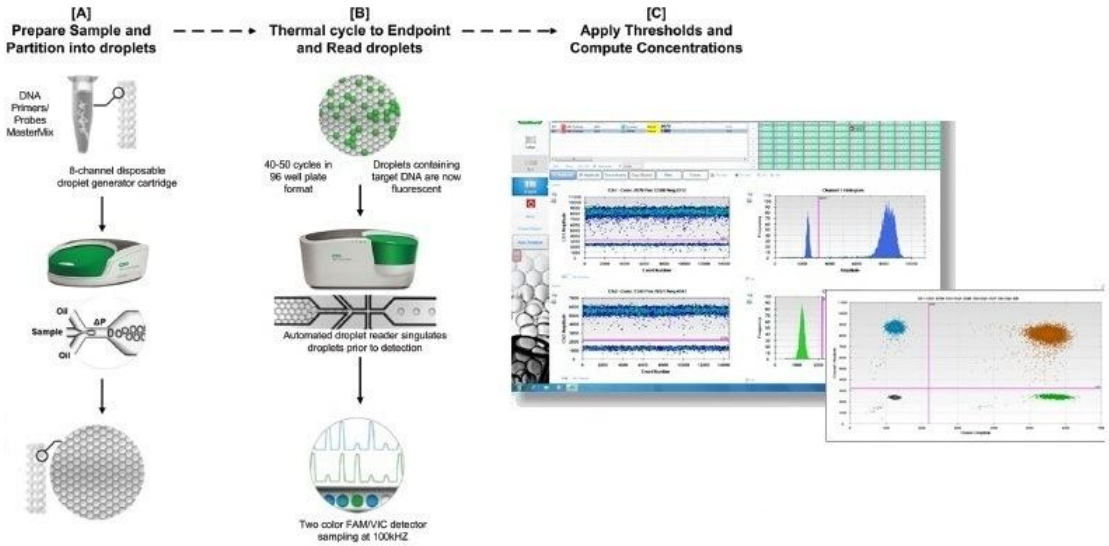


Figure 6: Schematic showing the ddPCR workflow. (A) Each 20 µL sample containing the Master Mix, primers, TaqMan probes, and DNA target is loaded in the middle wells of a disposable eight channel droplet generator cartridge. Droplet generating oil comprising the emulsion stabilizing surfactant is then loaded into the left-hand wells of the droplet generator cartridge. A vacuum is applied to the outlet well (right) creating a pressure difference that converts the aqueous sample into stable water-in-oil droplet emulsions which concentrate due to density differences from the oil phase and accumulate in the droplet collection wells of the cartridge. The droplets from each well are then transferred to a 96-well plate, foil sealed and thermal-cycled to the end-point. (B) After amplification, the plate is then loaded to a droplet reader where an auto sampler aspirates the droplets and, using a microfluidic singulator, streams them single file (~1500 droplets/s) past a FAM/VIC two colour fluorescence detector. (C) Threshold is applied and reaction mixture in units of copies/µl is calculated after being fitted to a Poisson distribution. Reproduced from (Pinheiro et al., 2012).

ddPCR was used by Handsaker and co-workers in their recent publication to evaluate their genotypes for CNVs with higher copy numbers (Handsaker et al., 2015), also Marques et al. (2014) reached their conclusion that CNVs typed via ddPCR, within regions identified in previous GWAS, may play a role in human essential hypertension.

1.2.8 Sequencing based methods

1.2.8.1 Next Generation Sequencing

Next-generation sequencing (NGS) represents a category of sequencing methods developed since 2005. These can be sub-classified as second-generation methods, including the most commonly used platforms such as Illumina (HiSeq and MiSeq), and Ion Torrent (PGM and Proton), and third-generation methods, referring to platforms developed by Pacific Biosciences and Oxford Nanopore.

The second-generation sequencing approach relies on short fragments (reads) of 35-400 bp, each of which can be considered as an independent experiment. Due to this, every nucleotide is sampled several times by several reads spanning that specific position in the genome (coverage). This strategy increases the amount of data produced at each sequencing run since multiple experiments are running in parallel at the same time. Third-generation sequencing platforms have been optimised in order to maximise the read length based on the length of DNA fragments. So far, Pacific Biosciences was reported to produce 7 kb read length data (English et al., 2012), while Nanopore devices promise to generate read lengths of ~50 kb (Schneider & Dekker, 2012). With this approach, unamplified DNA samples can be directly sequenced avoiding any PCR step.

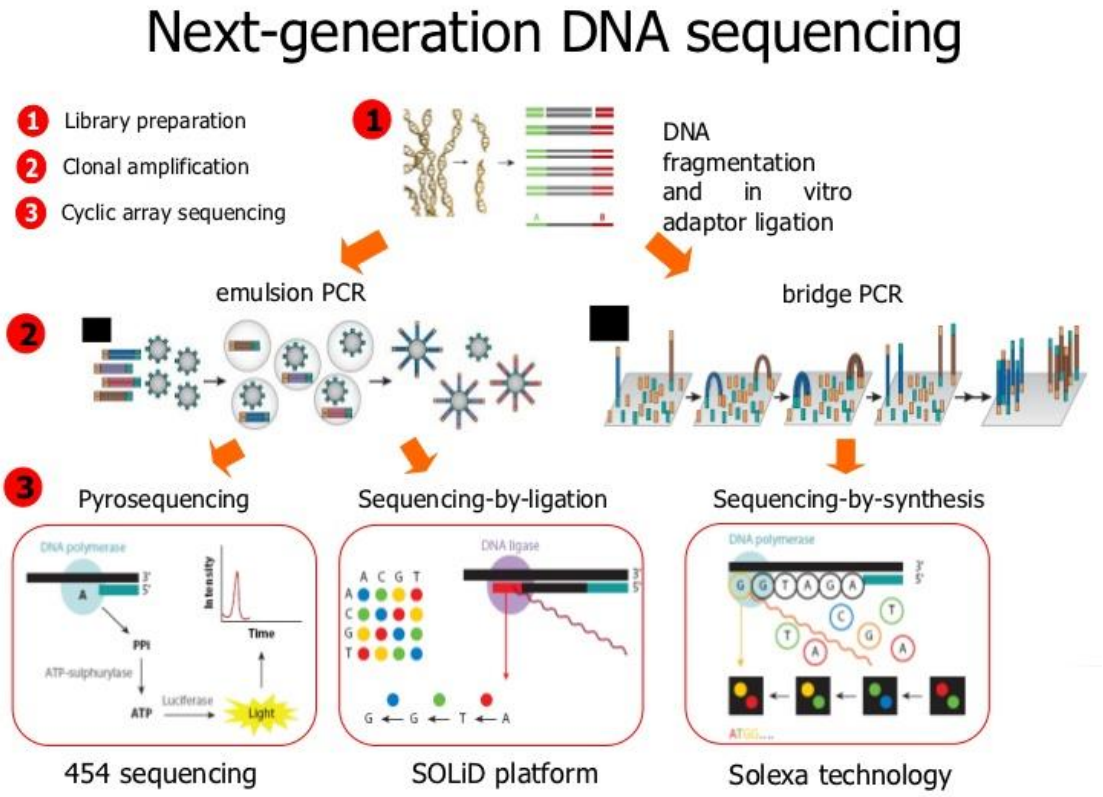


Figure 7: Schematic diagram showing the workflow of NGS. The first step of this method is to fragment genomic DNA to a uniform size. Sequence enrichment could be performed for targeted or exome sequencing; whole genome sequencing does not require enrichment. To make the library sequencing-ready, adapters are ligated to both ends of the DNA. The different coloured adapters (green and red) reflect a sequencing adapter and a barcoding adapter. The library is then immobilised to an array where bridge amplification occurs to generate clusters (clonal libraries). Sequencing is then done by the various available methods as per the manufacturer's orders.

1.2.8.2 Read depth methods

Read depth methodology is a useful measure for determining absolute CN as essentially, the number of sequencing reads that map to a specific region is proportional to the number of copies that the region is present in the genome (Medvedev et al., 2009). This basis for read depth methods assumes a Poisson distribution of sequencing reads, thus a region can be assumed to be deleted or duplicated if the region has fewer or more mapped reads than expected. This method has been successfully applied to complex genomic regions containing multiallelic CNPs (Sudmant et al., 2010). Typically, smaller CNV events and those that contain high

genomic copy number require deeper sequence read depth (SRD) to achieve accurate CNV measurement. One limitation of using massively parallel short-read sequencing is the inability to uniquely map short reads to regions such as SDs (Sudmant et al., 2010)

1.2.8.3 Genome Structure in Populations (genome STRiP)

Genome structure in populations (genome STRiP) is a suite of bioinformatics tools for discovering and genotyping structural variations using sequencing data. The methods are designed to detect shared variation using sequence data that are distributed across hundreds or thousands of genomes (Handsaker et al., 2011). Genome STRiP looks both across and within a set of sequenced genomes to detect variation and in order to run discovery or genotyping on a single sequenced genome or a small set of genomes, running the data against a background population, such as a set of genomes from the 1000 Genomes Project is required. Advantageously, the background population does not need to be matched to the target individuals. This method can be used for the discovery of novel structural variations or to genotype known variants in new samples (Broad Institute, 2015).

A schematic diagram to explain the analytical framework for analysing Genome Structure in Populations is shown in figure 8 below:

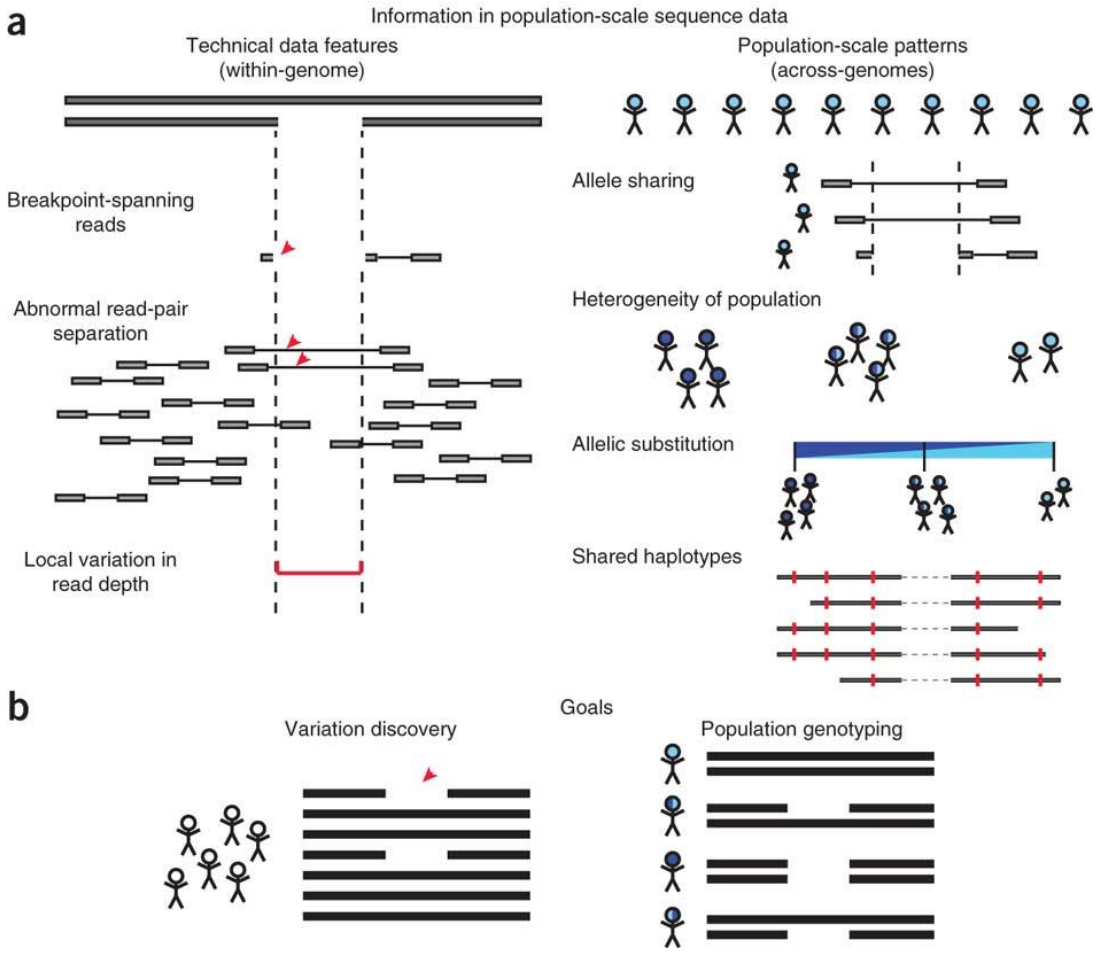


Figure 8: (a) Population-scale sequence data contain two classes of information: technical features of the sequence data within a genome and population-scale patterns that span all the genomes analysed. Technical features include breakpoint-spanning reads, paired-end sequences and local variation in read depth of coverage. Genome STRiP combines these with population-scale patterns that span many genomes, including: the sharing of structural alleles by multiple genomes; the pattern of sequence heterogeneity within a population; the substitution of alternative structural alleles for each other; and the haplotype structure of human genome polymorphism. (b) Goals of structural variation (SV) analysis in Genome STRiP. 'Variation discovery' involves identifying the structural alleles that are segregating in a population. The power to observe a variant in any one genome is only partial, but the evidence defining a segregating site can be derived from many genomes at once. 'Population genotyping' requires accurately determining the allelic state of each variant in every diploid genome in a population. Reproduced from (Handsaker et al., 2011).

Table 2 below summarises, compares and contrasts the above mentioned methods and techniques for typing copy number variations.

Table 2: Methods to measure copy number variations. Adapted from (Cantsilieris et al., 2013)

	qPCR [a]	MLPA [a]	MAPH [a]	PRT [a]	aCGH [b]	nCounter	ddPCR	NGS
Detection	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Absolute copy number	Absolute copy number	Absolute copy number
Sample	5-10ng DNA	100-200ng DNA	0.5-1µg DNA	5-10ng DNA	0.5-1µg DNA	200ng DNA	5-10ng DNA	1-2µg DNA
Loci	Single	>40	>40	Single	>2 million	Up to 200 regions in each reaction	2	Genome wide
Minimum theoretical resolution	100 bp	100 bp	100 bp	100 bp	5-10 kb	Depends on probe size	100 bp	>1 kb
Cost/sample	Low	Low	Low	Low	Moderate	High	High	High
Time to result	4 h	>24 h	>24 h	4 h	>24 h	8 h	4 h	2-3 days
Labour requirement	Low	Low	Moderate	Low	Moderate	Low [c]	Moderate	High

[a] Minimum resolution is in general the length of a single probe
[b] High resolution aCGH can achieve a minimum resolution of >500 bp
[c] According to Nanostring website (http://www.nanostring.com/products/CNV)

1.3 Defensins

1.3.1 Human defensins: alpha, beta, theta

Antimicrobial peptides (AMPs) are short polypeptides whose size is less than 100 amino acids (Ganz, 2003). They are located in host defence settings, and they play a significant antimicrobial function, especially at physiological concentrations, under specific conditions, present in the tissues they originate in (Ganz, 2003). AMPs have antimicrobial activity against bacteria, fungi and some enveloped viruses (Schneider et al., 2005). In humans as well as other mammals, AMPs are divided into two main families; defensins and cathelicidins. In addition to these, there are other AMPs such as bactenecins. Typically, AMPs are categorized based on features of their secondary structure. For example, cathelicidins have linear α-helical peptides while Defensins have β-strand peptides connected by disulphide bonds. Unlike defensins or cathelicidins, bactenecins have loop peptides (Hazlett & Wu, 2011).

Individual members of these families have been associated with the antimicrobial functioning of phagocytes, epithelial secretions in addition to inflammatory body fluids. In humans, the two principal antimicrobial peptide families are defensins and cathelicidins (Ganz, 2003). Defensins are broadly distributed in epithelial cells of mammals and phagocytes, and they often exist in high concentrations. Cathelicidins, on the other hand, are structurally as well as evolutionarily distinct antimicrobial peptides that correlate with defensins in terms of distribution and abundance. Other mammalian AMPs include histatins, dermcidin plus anionic peptides. Unlike defensins and cathelicidins, these AMPs are restricted not only to a few animal species but also too few tissues (Wang, 2014). This study, however, aims at expounding on defensins, particularly β-defensins (DEFB).

The term defensin was first introduced in 1985 after purification of granule rich sediments from neutrophils collected from human and rabbit. The purification of the sediments led to the characterization of the primary structure of the leading six α-defensins. These early studies concluded that all defensins have a highly conserved subject of 6 cysteine residues that define their antimicrobial function. Consequently, peptides with structures related to those of defensins introduced above were

discovered in bovine, mouse airway, and the human intestinal epithelium in the early 1990s. They were later named β-defensins (Machado & Ottolini, 2015). Defensins, a family of related vertebrate antimicrobial peptides, are characterized by β-sheet-rich fold besides a framework of 6 disulphide-linked cysteines (Ganz, 2003). Apart from being short and cationic, defensins are secreted as highly disulphide-bonded molecules with a very low molecular weight (3-6kDa) (Chen et al., 2006). They were firstly characterized as a family of multifunctional antimicrobial peptides and inflammatory mediators involved in the innate immune response found at epithelial surfaces (Schneider et al., 2005). Initially, defensins were subdivided into two main subfamilies; α- and β- defensins. The two differ in not only the length of peptides segments that are between six cysteines but also the pairing of the cysteines that are linked to each other by disulphide bonds. Recently, another subfamily of defensins referred to as θ- was discovered in the leukocytes of rhesus macaque monkeys (Ganz, 2003). As a result of this identification, defensins are currently subdivided into α-, β-, and θ- on the basis of not only the cysteine residues spatial distribution, but also because of the disulphide bonds connectivity (Chen et al., 2006; Pazgier et al., 2006). However, it is shown by many studies that the θ-defensins are found only in non-human primates like baboons (Li et al., 2014; Cole et al., 2004; Nguyen et al., 2003). The θ- defensins evidently evolved in primates, but they are inactivated in humans because of genetic mutations of the regions with the code for premature stop codons (Cole et al., 2004). Distinctive defensin peptides have been discovered in all mammals including man, chickens, and turkeys that have been examined (Zhao et al., 2001; Brockus et al., 1998; Harwig et al., 1994). Defensin-like peptides have also been identified in snake venom in which they are believed to play the role of representing the snake's adaptation defence peptides against other predators larger than them (Correa & Oguiura, 2013). From this assertion, we can affirm that defensins are present in cells as well as tissues associated with host defence against infections caused by microorganisms. In the majority of the animals, the concentration of defensins is highest in granules that are the storage of leukocytes (Ganz, 2003).

According to Weinberg et al. (2012), α-defensins are produced by not only polymorphonuclear leukocytes, but also intestinal Paneth cells. , β-defensins on the

other hand are produced mainly by epithelial cells and they also form pores in many biological membranes (Weinberg et al., 2012). The finding of β-defensins in these cells shows that their main role is to protect the host (mammals) against microbial pathogenesis at perilous confrontational sites. From this, in addition to immunoregulatory perspective, we can affirm that β-defensins can engage significant numbers of cell surface receptors to enhance chemotaxis. This is shown by the ability of hBD-2 to involve the CCR6 receptor in immature dendritic cells (DC) as well as T cells in a chemokine manner in addition to recruiting these cells to the sites of interest (Weinberg et al., 2012). Additionally, it is shown that the presence of hBD-3 enhances maturation of antigenic presenting cells (APCs) via toll-like receptors (Funderburg et al., 2007).

Besides their antimicrobial activity against Gram-positive and Gram-negative bacteria and fungi (Pazgier et al., 2006), defensins have also been found to have important antiviral activity, especially against HIV-1 (Hollox, 2008). Defensins also prompt some cellular functions especially when they act as ligands by binding to certain receptors. Although different animals have been identified with defensins, it is evident that their functions vary in relation to subfamilies. In other words, defensins do not have one common function. For instance, α-defensins have a wide antimicrobial activity against gram-positive and gram-negative bacteria. They also have antimicrobial activity against enveloped viruses and fungi. Unlike α-defensins, β-defensins mainly act against gram-negative bacteria besides yeast. In addition to this, β-defensins may also act against gram-positive bacteria depending on lipid II binding affinity (Ulm et al., 2012).

β-defensins have also been shown to have immune modulatory activity. An initial important observation was that β-defensins can recruit immature dendritic cells and memory T cells by binding to CCR6 to sites of infection and/or inflammation providing a link between the innate and adaptive arms of the immune system (Yang et al., 1999). The chemoattraction of CCR6-expressing cells was also demonstrated by h-BD3 and interestingly h-BD3 and h-BD4 have also been shown to attract macrophages (Wu et al., 2003) which do not express CCR6, suggesting involvement of a different receptor.

Functional interaction of h-BD3 with CCR6 or macrophages was dependent on both peptide structure and a particular cysteine (cysV) residue (Semple & Dorin, 2012).

Defensins in general play a dynamic role in numerous growth-dependent processes, like proliferation and healing of wounds (Seo et al., 2001). As for other species, one of the β-defensins referred to as *CBD103* has a strong effect on not only pigment type-switching in domestic dogs, but also signalling in transgenic mice through melanocortin receptors (Nix et al., 2013; Candille et al., 2007). In platypus, gene duplication in addition to subsequent functional diversification of β-defensins resulted in Ornithorhynchus venom defensin-like peptide (OvDLP) genes (Whittington et al., 2008).

1.3.2 Molecular structure of defensins

For α-defensins, the cysteine residues form the disulphide bonds with the topology Cys1-Cys6, Cys2-Cys4, and Cys3-Cys5. This varies significantly from that portrayed by β-defensins. The Cys1 links to Cys5 and Cys3 links to Cys6. Only the topology Cys2-Cys4 of α-defensins correlates with that of β-defensins as shown in figure 9 below. Contrary to α- and β, θ-defensins have not only a circular structure, but also the cysteine residues linked as Cys1-Cys6, Cys2-Cys5, and Cys3-Cys4 (Pazgier et al., 2006; Linzmeier et al., 1999). The mature human β-defensin is claimed to contain 41-50 amino acid residues which are all amphipathic molecules (Chen et al., 2006).

Members of β-defensins have poor sequence similarity, and as a result of this, their respective antimicrobial activity is not influenced by their primary structure. The use of nuclear magnetic resonance (NMR) data confirms the close similarity of hBD1, hBD2, and hBD3 tertiary structure, despite having different amino acid sequences. The strands of β-strands are arranged in an antiparallel pattern and held together by three intramolecular disulphide bonds between the six cysteines (Chen et al., 2006). However, the pattern of disulphide bridges may vary; the variation characterizes each family member. The amino-terminal region of β-defensin has a short α-helical loop that does not do not exist in α-defensins.

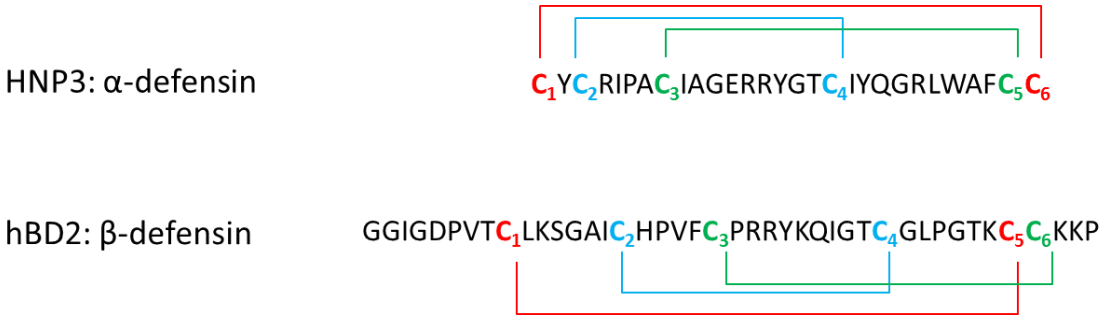


Figure 9: Sequences and the disulphide pairing of cysteines of α- and β-defensins.

1.3.3 β-defensins history and tissue distribution

There is an evolutionary relationship between the defensins of vertebrate and non-vertebrate; however, this evolutionary relationship is not clear. Despite this, phylogeny evidences that a primordial β-defensin is the main ancestor of all vertebrate defensins, and it is from this gene family that all vertebrate defensins evolved. This hypothesis is justified by detection of β-defensin-like genes in other vertebrates such as reptiles (van Hoek, 2014), birds (Zhao et al., 2001), and fishes (Zou et al., 2007). α-defensins are mammalian specific genes, and in humans α-defensin genes and different β-defensin genes are present on adjacent loci on chromosome 8p22–p23. The organization of this cluster is consistent with a model of multiple rounds of duplication and divergence under positive selection from a common ancestral gene that produced a cluster of diversified paralogous (Maxwell et al., 2003). This expansion occurred before the divergence of baboons and humans ~23–63 million years ago (Semple et al., 2003). In relation to this, the present β-defensins might have evolved before mammals separated themselves from birds resulting in α-defensins in rodents, primates as well as lagomorphs after their separation from other mammals. Recent evidence suggests convergent evolution of β-defensin copy number (DEFB CN) in primates, where independent origins have been sponsored by non-allelic homologous recombination between repeat units. For rhesus macaques this resulted in only a 20 kb copy number variation (CNV) region containing the human orthologue of human β-defensin 2 gene (hBD-2). Evidence claims that β-defensins were first reported as the tracheal antimicrobial peptide (TAP) of cow tongue and in chicken leukocytes (GAL-l) (Ganz & Lehrer, 1994). hBD-1, which is the first human β-defensin, was isolated as a 36-residue

peptide in early 1995 from the hemofiltrate of patients suffering from advanced renal failure (Bensch et al., 1995). hBD-1 is constitutively expressed in the epithelial cells of sensitive body parts such as the respiratory and urogenital tract (Janssens et al., 2010). It is also expressed in trachea, uterus, pancreas, and kidneys (Janssens et al., 2010; Zilbauer et al., 2010; Rodríguez-Jiménez et al., 2003). Apart from being expressed in the lung, prostate, placenta, and thymus, hBD-1 is also present in the testis, vagina, ectocervix, endocervix, as well as the fallopian tubes (Hazlett & Wu, 2011). In addition to these body parts, hBD-1 is also expressed in gingival tissue, buccal mucosa and tongue (Weinberg et al., 2012; Yanagi et al., 2005). Other regions that have well expressed hBD-1 include the salivary glands, small intestine, conjunctiva, and cornea (Zilbauer et al., 2010; Schröder & Harder, 1999). Other glands in which you can identify hBD-1 are lacrimal gland and mammary gland. hBD-1 is also expressed in the limb joints, astrocytes, microglia, as well as meningeal fibroblasts (Hollox, Huffmeier, et al., 2008; Pazgier et al., 2006; Quiñones-Mateu et al., 2003).

In 1997, the hBD-2 – a 41 residue peptide which is the second human β-defensin was isolated from psoriatic skin lesions (Schneider et al., 2005). This type of β-defensin (hBD-2) is extensively expressed in epithelia of the skin, oral epithelia, and respiratory tract (Braff & Gallo, 2006; Pazgier et al., 2006; Dunsche et al., 2001; McCray & Bentley, 1997). In addition to being expressed in epithelia of ocular surface and gingival keratinocytes, hBD-2 is expressed in epithelial of gastrointestinal tract (Haynes et al., 1999).

In 2000, hBD-3, which is regarded as the third β-defensin in the human genome, with a 67 amino acid peptide was singly identified by three groups of researchers. Harder et al., (2003) purified it from human psoriatic lesion scales. Garcia et al.,(2003) and Jia et al.,(2003) identified it through the aid of the bioinformatics genomics-based approach. According to Pazgier et al. (2006), hBD-3 is expressed mainly in the skin, oral, and respiratory as well as gastrointestinal tract. It has also been suggested that hBD-3 is expressed in the placenta, testis, and heart. Apart from being detected in the gingival tissues, it is also found in the primary keratinocytes.

Although hBD1, 2 and 3 have antimicrobial activity against yeast and some Gram-positive bacteria, they are claimed to be predominantly active against Gram-negative

bacteria in addition to increased activity strength. Unlike hBD-2 and 3, hBD-1 is constitutively expressed. The two β-defensins; hBD-2, and 3 are expressed under specific conditions. For the two to be expressed, they must be induced by proinflammatory agents like Tumour Necrosis Factor alpha (TNF-α), and Interleukin 1 Beta (IL-1β) (Chen et al., 2006). They are also induced by Lipopolysaccharide (LPS) (Hazlett & Wu, 2011).

As from the year 2000, hBD-4, 5, and 6 have been identified via use of the basic local alignment search technique (BLAST) (Yamaguchi et al., 2002). For instance, hBD-4 is claimed to be expressed in high levels in the testis (Chen et al., 2006) whereas hBD-5 and 6 in the epididymis (Yamaguchi et al., 2002).

The anatomical distribution as well as presentation of β-defensins in various epithelial and mucosal tissues portrays the ability of β-defensins to counteract different pathogens. The presented sites, like epithelial and mucosal sites, are highly prone to microbial infections. For example, it was introduced above that hBD-2 is highly expressed in the lung; hBD-4 strongly expressed in the testes as well as the stomach, and hBD-3 expressed in not only the skin but also tonsillar tissues. The expression of β-defensins in the organs and epithelial tissues that were discussed above is as a result of inducing factors. For instance, expression of hBD-2 in the lung is as a result of lipopolysaccharides or other kinds of bacterial epitopes together with interleukin-1β that is produced by monocyte-derived cells. hBD-3 and hBD-4, on the other hand, are induced by tumour necrosis factor (TNF), interferon (IFN)-γ, and/or toll-like receptor ligands (Yamaguchi et al., 2002). hBD-2 was also found to be involved in bone cell differentiation and enhance mineralization of osteoblast-like cells (Kraus et al., 2012)

Alongside genes *DEFB1* and *DEFB4*; hBD-26 (encoded by *DEFB126)* plays an important function in reproduction. hBD-26 coats the sperm's glycoprotein and takes part in the attachment of the sperm to epithelia of the oviduct (Tollner et al., 2008; Yudin et al., 2005). In support of this, the deletion of a cluster of nine β-defensin genes in a mouse model, resulted in male sterility (Zhou et al., 2013). In human studies, a common mutation in *DEFB126* has been shown to impair sperm function and fertility (Tollner et al., 2011).

1.3.4 Human β-defensins at the 8p23.1 locus

β-defensin genes are arranged in three main clusters in humans. The clusters include 8p23.1, 20p13, and 20q11.1, in addition to another small cluster located on chromosome 6p12. However, this study is concerned specifically with one cluster, 8p23.1.

The 8p23.1 region is known to be a common site of chromosomal rearrangements facilitated by two huge blocks of low copy repeats (LCRs). The region can also be as a result of segmental duplications (SDs). The whole 8p23.1 region, which includes the SDs in addition to containing about 50 genes, can extend in length up to 6.5 Mb. The two large set of SD blocks are referred to as "REPeat Distal" (REPD) and "REPeat Proximal" (REPP). Each of the two SDs constitutes of olfactory receptor gene clusters, defensins and *FAM90A* clusters that are located to copy number variable regions (Hollox et al., 2003).

REPD and REPP are highly homologous. The LCR regions and NAHR between them results in *8p23-inv* which is one of the largest polymorphic inversions in humans. *8p23-inv* encompasses approximately 4.5Mb (Salm et al., 2012). It has been genotyped by FISH (Giglio et al., 2001). However, FISH is not a very suitable method to high throughput analyses since it requires viable cells in metaphase. In addition to this, the size of the inversion's single copy region approaches the resolution limit of FISH techique (Raap, 1998). In 2012, Salm et al. came up with a bioinformatics tool that utilizes SNP data in enabling accurate genotyping of the 8p23.1 inversion known as "phase free inversion detection operator" (PFIDO). In relation to Salm et al.'s study, Giglio et al., found out that the inversion has an estimated frequency of 21% in Japanese populations and a frequency of 52% in African populations (Sugawara et al., 2003; Giglio et al., 2001). The relevance of this inversion variant in the general population is its relationship with flanking segmental duplications (REPD and REPP), which may also be involved with generation of CNVs during crossovers (Small & Warren, 1998).

The inversion is not the only significant chromosomal rearrangement that occurs at this genomic location. In 1999, Devriendt et al., using cytogenetic and FISH techniques associated recurrent deletions at 8p23.1 with congenital heart malformations and

congenital diaphragmatic hernia in 9 patients. Wat et al., in 2009 confirmed these results in 4 patients using cytogenetics, high resolution aCGH and sequencing analysis.

Reciprocal to the 8p23.1 deletion disorder, a genomic disorder known as the 8p23.1 duplication syndrome (8p23.1 DS) has been revealed by aCGH (Barber et al., 2008). 8p23.1 DS and CNV of the 8p23.1 defensin gene cluster are cytogenetically indistinguishable but distinct at the molecular level (Barber et al., 2010). It has an estimated prevalence of 1 in 58,000 and the core 3.68Mb duplication contains 32 genes (Barber et al., 2015). Barber and colleagues (2015) described four patients and five families with eight microduplications of 8p23.1 ranging from 187 to 1082 kb in size and 1 atypical duplication of 4 Mb. They indicated that a minimal region of overlap (MRO) spanning 776 kb in medial 8p23.1 can give rise to features of 8p23.1 DS that included developmental delay, dysmorphism, macrocephaly and otitis media, but not congenital heart disease (CHD) (Barber et al., 2015).

Embedded within REPP and REPD is a large segmental duplication that carries the β-defensin cluster. The history of the reference assembly of this region across several builds of the human genome is shown in Figure 10. It is worth noting the presence of a recalcitrant gap even in the most recent human genome assembly probably due to the polymorphic nature of this region. Both physical and genetic mapping approaches show that the current assembly is only an approximation of the actual situation. Using these approaches, it has been shown that β-defensin repeats can be at REPD, REPP, or both, and can be present at multiple copies at each locus (Abu Bakar et al., 2009; Hollox et al., 2008; Hollox et al., 2003).

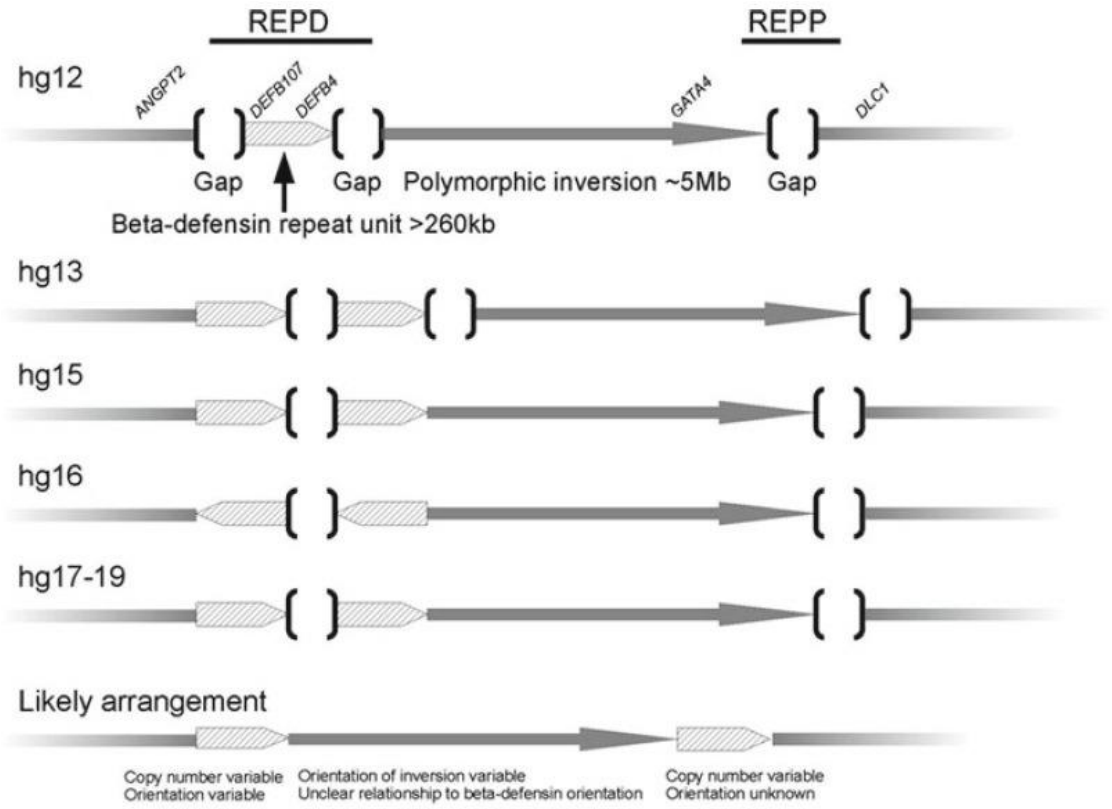


Figure 10: A history of the reference assembly of the beta-defensin region at 8p23.1 showing assembly release from hg12 (released June 2002), with the likely arrangement deduced from both physical and genetic mapping approaches shown at the bottom of the diagram. Not drawn to scale. Reproduced from (Hollox, 2012).

The β-defensin copy number variable cluster constitutes of various annotated genes, 7 of which belong to the β-defensin gene family; *DEFB4, DEFB103, DEFB104, DEFB105, DEFB106, DEFB107, and DEFB109* (Forni et al., 2015). This region varies in copy number as a whole unit with a modal copy number of 4 and a range of copy numbers from 1 to 12 with the majority of the people having 2–7 copies (Hollox et al., 2008; Hollox et al., 2003). Since the region is copy number variable *en bloc*, it was suggested that stating the copy number of the *DEFB4* gene is representative of the entire β-defensin cluster.

The actual size of the β-defensin repeat unit is unclear despite the confirmation from Hollox et al. (2003) that the pulse-field gel analysis showed the unit to be at least 260 kb in size. In 2014, Taudien et al., carried out a study aimed to narrow down the distal border of the DEFB cluster. They established tests for length polymorphisms based on amplification and capillary electrophoresis with laser-induced fluorescence (CE-LIF)

analysis of seven indels containing regions spread over the entire cluster. The tests were carried out on 25 genomic DNAs with different previously determined cluster copy numbers. CNV was demonstrated for six indels between ~1 kb distal of *DEFB108P* and 10 kb proximal of *DEFB107*. Their analysis fixed the minimal length of proven CNV to 157 kb (Taudien et al., 2014).

In addition to the β-defensins, nine other genes mostly expressed in the testes are present in the cluster (Ottolini et al., 2014). One of which is the *SPAG11* gene, associated with encoding for an androgen-dependent molecule explicitly expressed in the epithelium of the male reproductive tract; Epididymis-Specific Secretory Protein (*SPAG11/HE2/EP2*). *SPAG11* portrays strong sequence homology to β-defensins since it is formed by a head to head fusion of β-defensins (Ribeiro et al., 2012; Hollox, Barber, et al., 2008).

Most genomic structures of the β-defensin genes constitute of two exons and one intron. However, *DEFB105* gene is an exception; it has three exons and two introns. The first exon encodes the signal peptide whereas the second conveys information about the mature peptide sequence heralded by a short anionic pro-peptide (Pazgier et al., 2006).

1.3.5 β-defensins CNV and implication in disease

The consequences of CNV as it was explained above are diverse and many. They may include not only increased gene product and production of fusion genes but also the creation of extra coding domains as well as position effect that impair with the expression of gene product. This extensive genome differences in humans can result in diseases where DEFB may be linked to their pathology. This includes a good number of autoimmune and infectious diseases. Given the common variation of β-defensin CN and their important functional role, it has been suggested that β-defensins may be a risk factor in susceptibility to different diseases with both genetics and environmental components to their aetiology. The first study on the association of β-defensin CNV with a clinical phenotype was carried out by Hollox et al., in 2005 on a cohort of Cystic

Fibrosis (CF) patients. No association was found between DEFB CNV and lung disease associated with CF in 355 UK patients.

Using real-time PCR for CNV typing, Fellermann et al., (2006) came to the conclusion that low *DEFB4* gene CN predisposes a person to Crohn disease of the colon. Interestingly, Bentley et al. (2010) found out that elevated *DEFB4* CN enhances the development of Crohn's disease (CD), challenging the data that was published earlier. In relation to Crohn's disease, it was shown that reduced number of Paneth cell in ileum leads to ileal CD. Nonetheless, neither the Welcome Trust Case-Control Consortium (WTCCC) (Craddock et al., 2010) nor the trial to repeat the study using the PRT for typing β-defensin CN succeeded in finding the association between β-defensin CNV and a clinical phenotype (Aldhous et al., 2010).

The studies by Fellermann et al. (2006) and Bentley et al. (2010) failed to successfully explain the association between β-defensin and CNV. The failure of their studies might have been as a result of not fulfilling the requirements of a powerful study which includes a large sample size to minimize the rate of false positives, and an independent replication in a similar sized cohort to give the findings of the studies a higher degree of confidence. For instance, Bentley et al. had a significant sample size with no replication. Fellermann et al. on the other hand had replication with a small sample size. These variations must have contributed to the contradiction of the results. Another reason for their failure could be the choice of method in typing β-defensin CNV. Both Bentley et al.'s and Fellermann et al.'s studies opted for quantitative real-time PCR (qPCR). The ΔCt method depends very much on the amplification efficiency of control and test that are competing in a given single reaction. It has been shown that a 4% change in amplification efficiency often results in an error of up to 400% in ΔCt calculation. Thus, CNV results obtained by qPCR method are often questionable (Fernandez-Jimenez et al., 2011). Apart from this, the data used by Bentley et al. (2010) do not illustrate any clear tendency to cluster around integer values of all sizes including those at low copy numbers of 2, 3 or 4. Thus, the qPCR measurements of β-defensin CN in Bentley et al.'s study may have been influenced by other factors such as the DNA's physicochemical state and adequate degree which occasionally results in incorrect integer call (Aldhous et al., 2010).

In another association study between Dutch and German cohorts, it was found out that the risk of Psoriasis increases with a greater genomic CN of β-defensin genes (Hollox et al., 2008). The same findings have been found by Stuart et al.'s study (2012). Persons with over five β-defensin copies were found to have an increased risk of having psoriasis when compared with individuals with two copies of β-defensin.

Additionally, β-defensin CNV is reported to be involved in susceptibility to Human Immunodeficiency Virus -1 (HIV-1) infection (Hollox et al., 2008; Quiñones-Mateu et al., 2003) as well as progression of Acquired Immunodeficiency Syndrome (AIDS) (Mehlotra, Zimmerman, et al., 2012; Weinberg et al., 2012). In tandem with Mehlotra et al. (2012) and Weinberg et al. (2012) findings, a study that was carried out by Hardwick et al. in 2012, showed that higher β-defensin CN results in increased HIV viral load (VL) prior to highly active antiretroviral therapy (HAART), and poor immune reconstitution following HAART. This claim can be explained by the chemoattractant nature of the β-defensins; hBD-2 which is encoded by *DEFB4* acts via CCR6 and facilitates the seizure of T_h17 cells on the vascular endothelium which is then followed by the penetration of the mucosa. T_h17 cells which are favourably infected by HIV-1 because of the high co-expression of *CCR5*; the HIV co-receptor (Hardwick et al., 2012). Thus, the levels of mucosal β-defensin expression are likely affected by the CNV which affects the pool of T_h17 cells paving way for HIV infection. Additionally, it can be affirmed that β-defensins indirectly act to impair the immune reconstitution. They achieve this by recruiting other cells like dendritic cells to the mucosa, which then affect the cytokine environment (Zeng et al., 2011).

To summarise, table 3 reviews the researches that have been carried out to investigate the role DEFB CN plays – if any – on the onset of the diseases, and the method used in genotyping the DEFB CN.

Table 3: summary of research carried out to study correlation of DEFB CN to the disease in question.

Disease	Method of Typing	Relation	Reference
Cystic Fibrosis	MAPH	No association	(Hollox et al., 2005)
Crohn	aCGH and qPCR	Lower *DEFB4* CN predisposes to colonic CN	(Fellermann et al., 2006)
	qPCR	Higher *DEFB4* CN is a risk factor	(Bentley et al., 2010)
	PRT	No association	(Aldhous et al., 2010)
Psoriasis	MAPH/PRT	High *DEFB4* CN and risk of disease	(Hollox et al., 2008)
Behçet's	PRT	No association	(Park et al. 2011)
Pancreatic Ductal Adenocarcinoma (PDAC) and Chronic Pancreatitis (CP).	MLPA	Higher *DEFB4* protects against disease	(Park et al. 2011)
HIV	PRT	Higher *DEFB4* CN increases HIV load prior to HAART and poor immune reconstitution after HAART	(Hardwick et al., 2012)
Systemic lupus erythematosus (SLE) and ANCA associated small vasculitis (AASV)	PRT	Higher *DEFB4* CN associated with SLE and AASV	(Zhou et al., 2012)
Age of Onset in Huntington Disease	PRT	No association	(Vittori et al., 2013)
Asthma and COPD	PRT	No association	(Wain et al., 2014)
Susceptibility to otitis media	PRT	*DEFB4* CN associated with nasopharyngeal microbiota composition	(Jones et al., 2014)

1.4 Aims of the study

- Further explore the structure of the copy number variable DEFB region via analysing aCGH breakpoint analysis and association with inversion at the 8p23 locus.
- Compare and contrast the different methods used for calling DEFB CNVs.
- Develop a model system to investigate if DEFB expression levels differ with CN in response to treatment with pneumolysin by using Normal Human Bronchial Epithelial (NHBE) cells.
- Explore and replicate the DEFB association study with HIV viral load.
- Investigate if DEFB CN is associated with hypertension, obesity and clinical symptoms that occur more commonly in obese individuals.

2 Materials and Methods

2.1 DNA samples used

2.1.1 HapMap samples

The International HapMap Project was initiated in 2002 to aid in the study of genetic diversity among different populations. A total of 270 DNA samples are present in the project that derived from blood donors from four different populations. The Yoruba people (YRI) of Ibadan, Nigeria provided 30 sets of samples from two parents and an adult child. Each such set is referred to as a trio as is the U.S set (CEU), which were collected in 1980 from U.S. residents with northern and western European ancestry by the Centre d'Etude du Polymorphisme Humain (CEPH). However, 45 samples were collected from unrelated individuals for each of the Japanese set (JPT) from Tokyo and the Han Chinese set (CHB) from Beijing (The International HapMap Consortium, 2003).

The Coriell Institute (http://ccr.coriell.org) provided DNA and cell lines from the above samples for research projects that have appropriate ethical approval.

The DEFB CN for these samples has been previously genotyped using the PRT assay by Dr. Rob Hardwick (Hardwick et al., 2011).

2.1.2 UTI and VUR study subjects

The Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) study is a double-blind, randomized, placebo-controlled trial which recruited 607 children aged 2 to 72 months from 19 paediatric sites across North America assigned to receive daily doses of either antibiotics or placebo for 2 years. Biospecimens and genetic specimens were collected every 6 months throughout the trial to study genetic and biochemical determinants of VUR, recurrent UTI, and renal scarring (Keren et al., 2008). The study was approved by the Nationwide Children's Hospital Institutional Review Board, Nationwide Children's Hospital, Columbus, Ohio, USA.

DNA of 456 individuals from the RIVUR study was sent to us by our collaborator Dr. David Hains (Nationwide Children's Hospital, Ohio, USA).

2.1.3 HIV study subjects

2.1.3.1 International AIDS Vaccine Initiative (IAVI) cohort

IAVI is a global not-for-profit, public-private partnership working to accelerate the development of vaccines to prevent HIV infection and AIDS. This prospective cohort was part of the Protocol C project launched in 2006 to learn more about how HIV progresses and is transmitted. More than 600 volunteers with incident HIV infection have enrolled in Protocol C to date. All volunteers are provided with or referred for routine HIV care, including highly active antiretroviral therapy (HAART) provision. Some 190,000 Protocol C samples have been collected and more than 26,000 shared with researchers around the world. Volunteers were from Kenya, Uganda, Rwanda, Zambia and South Africa (IAVI, 2006). Our collaborators; Dr. Jacques Fellay (ÉPFL, Switzerland) and Dr. David Goldstein (Duke University) provided us with 423 samples.

2.1.3.2 Swiss HIV Cohort Study (SHCS) cohort

The Centre for HIV-AIDS Vaccine Immunology (CHAVI) is founded by the National Institute of Allergy and Infectious Diseases. The Euro-CHAVI Consortium is coordinated by professors in Switzerland and have various cohorts one of which is the Swiss HIV Cohort Study (SHCS) (Fellay et al., 2009). SHCS was established in 1988, as a systematic longitudinal study enrolling HIV-infected individuals in Switzerland. It is a collaboration of all Swiss University Hospital infectious disease outpatient clinics, two large cantonal hospitals, all with affiliated laboratories, and with affiliated smaller hospitals and private physicians looking after HIV patients. Essentially, The SHCS is representative for the Swiss HIV-epidemic (Swiss HIV Cohort Study, 2015). Our collaborators; Dr. Jacques Fellay (ÉPFL, Switzerland) and Dr. David Goldstein (Duke University) sent us 3360 DNA samples. 20µL of 1x TE buffer was added to all the samples before being genotyped for DEFB CN using the PRT.

2.1.4 Cardiovascular cohorts study subjects

2.1.4.1 Young Men Cardiovascular Association (YMCA)

The Young Men Cardiovascular Association Study 1 (YMCA 1) consists of 1,157 biologically unrelated, healthy males recruited in Silesia (Southern Poland) (Tomaszewski et al., 2007; Charchar et al., 2004). The Young Men Cardiovascular Association Study 2 (YMCA 2) - an extension of YMCA 1- recruited an additional sample of unrelated 597 young men in Silesia (Tomaszewski et al., 2007). Clinical and biochemical phenotyping protocols of each study were described in detail in previous publications (Tomaszewski et al., 2007; Charchar et al., 2004). In brief, clinical history was collected using anonymised, coded questionnaires. Recorded anthropometric measurements included height and weight, as well as three consecutive BP measurements (Tomaszewski et al., 2007; Charchar et al., 2004). Only 1.6% and 0.3% men in YMCA 1 and YMCA 2, respectively, were on antihypertensive medication (Tomaszewski et al., 2007; Charchar et al., 2004). BP values from those on antihypertensive treatment were adjusted for BP-lowering effect of therapy using a previously reported method (Tomaszewski et al., 2007; Charchar et al., 2004). The samples were provided to us by our collaborator; Dr. Maciej Tomaszewski (University of Leicester, UK).

2.1.4.2 Silesian Cardiovascular Study (SCS)

The SCS is a cohort of 213 Polish families and 435 singletons recruited through probands with high cardiovascular risk (history of hypertension, coronary artery disease, and/or multiple cardiovascular risk factors), as previously described (Tomaszewski et al., 2009). The samples were generously provided to us by our collaborator Dr. Maciej Tomaszewski (Glenfield General Hospital, UK).

2.2 General Reagents

2.2.1 10x "Low dNTPs" PCR Buffer (10x Ld PCR Buffer)

The 10X Ld PCR buffer was used for the Paralogue Ratio Test method. The buffer contained a final concentration of 50mM Tris-HCl (pH8.8), 12.5mM ammonium sulphate, 1.4mM magnesium chloride, 7.5mM 2-mercaptoethanol, 125µg/mL non-acetylated Bovine Serum Albumin (BSA) (Ambion®, Thermo Scientific) and 200µM of each dNTP.

2.2.2 1x Tris-EDTA Buffer (TE Buffer)

1x TE Buffer was composed of 10mM Tris-HCl (pH8), which maintains the pH of the solution and 1mM EDTA, a chelator of metal ions which helps protect DNA and RNA from enzymatic degradation.

2.2.3 0.5x Tris-Borate-EDTA Buffer (TBE Buffer)

This buffer was used in agarose gel electrophoresis and was made from 40mM Tris-HCl (pH8.3), 4mM Boric acid and 1mM EDTA.

2.2.4 Wash Buffer

This buffer was used in the ELISA protocol and was made from 0.05% Tween-20 (Sigma) in 1X Dulbecco's Phosphate-Buffered Saline (PBS) (Gibco®, Thermo Scientific).

2.2.5 Block Buffer

This buffer was used in the ELISA protocol and it was composed of 1% BSA in 1X PBS.

2.2.6 Diluent

The diluent was used in the ELISA protocol for dilution purposes and it was composed of 0.05% Tween-20, and 0.1% BSA in 1X PBS.

2.3 Primer design

The desired sequence was retrieved from the hg19 human reference genome assembly on UCSC Genome Browser Home (http://genome.ucsc.edu/).

The Primer3 software (http://primer3.ut.ee/) was used, under the default settings, to check and confirm the thermodynamic properties of the PCR primers. The primers retrieved were then checked using the In-silico PCR tool of the UCSC Genome Browser to ensure that primer pairs were unique to the sequence to be amplified and that no common sequence variants were observed.

2.4 Generic PCR optimization

All PCR-based assays were optimised for the following parameters; annealing and extension temperature, extension time and buffer composition in order to maximise the specificity and yield of the reaction for the expected PCR products. For each new primer pair, a standard annealing temperature gradient between 50°C-60°C was routinely performed using the VeriFlex option in the Veriti thermal cycler (Applied Biosystems). Each reaction was first performed in standard Kapa A Buffer (Kapa Biosystems) which includes 1.5 mM magnesium chloride. In presence of non-specific products, smears or low amplification efficiency upon resolving on agarose gel electrophoresis (2% w/v), the 10X low dNTPs PCR buffer was tried. The following standard PCR amplification protocol was used for expected PCR products up to 1kb.

Step	Temperature (°C)	Time	No. of Cycles
Initial Denaturation	94	2 min	1
Denaturation	94	30 sec	35
Annealing	50-60	30 sec	
Extension	70	30 sec	
Final Extension	70	2 min	1

2.5 Agarose gel electrophoresis

PCR products were resolved on 2% agarose gel (2g of agarose was mixed with 100ml of 0.5X TBE buffer and heated in a microwave until the agarose completely dissolved and a transparent solution was formed). Ethidium bromide was added to a final concentration of 0.5µg/ml to the solution to allow visualisation under UV light. The agarose solution was then poured into a casting tray and allowed to solidify, the comb was gently removed and the gel casting tray was submerged in the electrophoresis tank containing 1X TBE buffer. 15µl of DNA was mixed with 3µl of 6X loading dye and ran alongside 3µl of DNA ladder (HyperLadder™ V, Bioline). Electrophoresis was performed at 120V for 50 minutes. The gel was then observed under a UV transilluminator and a soft copy was kept on record.

2.6 Paralogue Ratio Test

The Paralogue ratio test (PRT) is a comparative PCR based approach that uses one pair of primers to simultaneously amplify a copy number variable unit (test locus) and a reference locus with a diploid copy number of two. The products are then separated by size using capillary electrophoresis.

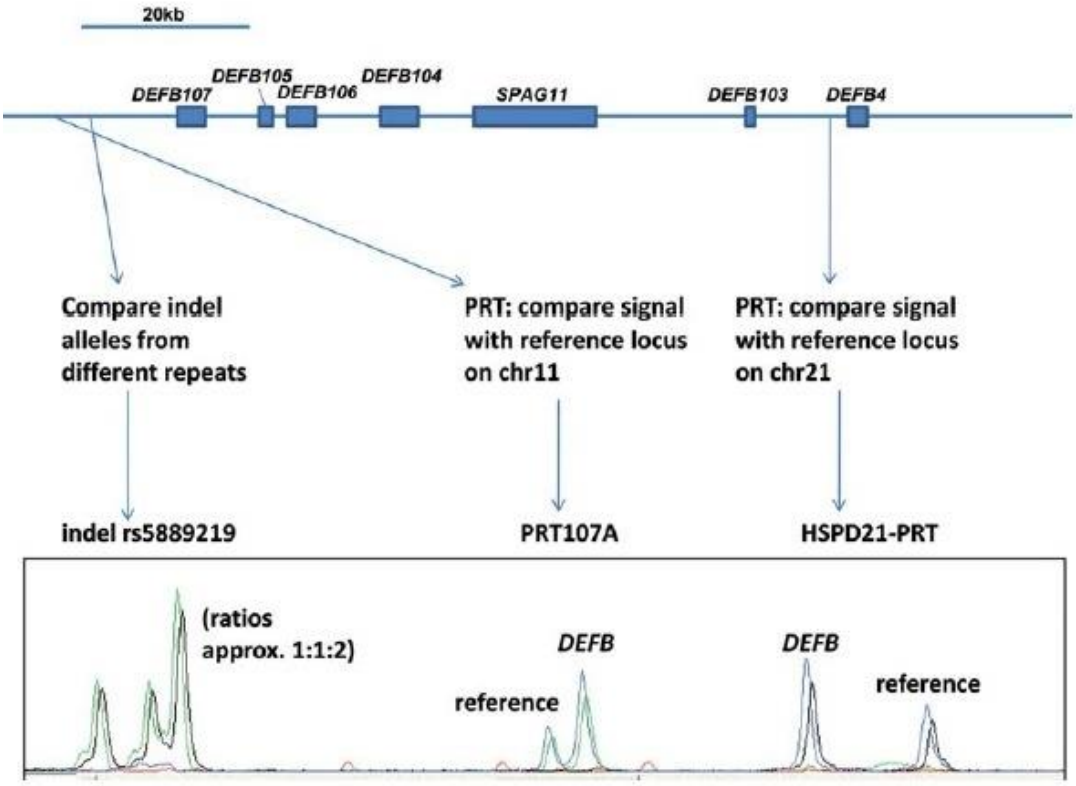


Figure 11: The copy variable β-defensin genomic region showing the locations in the repeat sampled by the PRT-based 'triplex' test. The primers for the PRT assays were designed to specifically amplify two loci in the β-defensins repeat unit and their paralogues elsewhere in the genome, producing different length amplicons. The third assay (rs5889219) examines a triallelic indel polymorphism (123/126/128 bp) that can have different sizes in individual β-defensin repeats. This trace shows a sample with a 4 copy number, reflected in the ratios between the 'test' (β-defensins) and 'reference' peaks in the two PRT systems HSPD21 and PRT107A, and in the approximate 1:1:2 ratios of alleles (consistent with a total of four copies) for the multi-allelic indel rs5889219. Reproduced from (Aldhous et al., 2010).

A modified duplex PRT protocol (Aldhous et al., 2010) was used to measure DEFB diploid copy number. The PRT primers were designed to amplify different locations within the DEFB repeat at 8p23.1 loci and corresponding references (Figure 11), details of which are shown in Table 4 below:

Table 4: PCR Primer sequences for the duplex PRT assay and indel assay. The PRT assays were used to amplify different locations in the DEFB CN variable region whereas the indel assay amplified different polymorphisms in the DEFB CN variable region.

Primer	**Sequence (5' → 3')**	**Size resolved by capillary electrophoresis (bp)**
PRT107A-F PRT107A - R	AGCCTCATTTAACTTTGGTGC GGCTATGAAGCAATGGCCTA	157 for Test on Chromosome 8 155 for Reference on Chromosome 11
HSPD21-F HSPD21-R	GAGGTCACTGTGATCAAAGAT AACCTTCAGCACAGCTACTC	172 for Test on Chromosome 8 180 for Reference on Chromosome 21
5DEL-F 5DEL-R	AAACCAATACCCTTTCCAAG TTCTCTTTTGTTTCAGATTCAGATG	134, 129 and 127

Each PCR was carried out twice using different fluorescently labelled primers (NED- or HEX-labelled PRT107AF, FAM- or HEX-labelled HSPD21R, HEX- or FAM-labelled 5DEL). In total; four master mixes as shown in Figure 12 below were prepared for a given DNA plate, with the last column on each plate reserved for gold-standard control samples (c0088 (cn=4), c0207 (cn=5), c0849 (cn=6), c0913 (cn=3), c0940 (cn=4), c0969 (cn=5)) all from the European Collection of Cell Cultures (ECACC), available from Public Health England, Porton Down, Wiltshire, UK. The samples have not only showed reproducible results with multiple PRT assays but their copy numbers had been also confirmed by other methodologies such as, MAPH (multiplex amplifiable probe hybridization) and REDVR (restriction enzyme digest variant ratio) tests (Armour et al., 2007). The samples were also used as controls in previous studies (Aldhous et al., 2010; Jones et al., 2014).

Master Mix 1

Component	X1 reaction (µl)
10X low dNTP Buffer (200µM)	1
PRT107A – F[NED] (10µM)	0.5
PRT107A – R (10µM)	0.5
HSPD21 – F (10µM)	0.5
HSPD21 – R[FAM] (10µM)	0.5
Taq DNA Polymerase (5U/µl)	0.1
DNA (10ng)	1
water	5.9
Total	10

Master Mix 2

Component	X1 reaction (µl)
10X low dNTP Buffer (200µM)	1
PRT107A – F[HEX] (10µM)	0.5
PRT107A – R (10µM)	0.5
HSPD21 – F (10µM)	0.5
HSPD21 – R[HEX] (10µM)	0.5
Taq DNA Polymerase (5U/µl)	0.1
DNA (10ng)	1
water	5.9
Total	10

Master Mix 3

Component	X1 reaction (µl)
10X low dNTP Buffer (200µM)	1
5DEL – F[FAM] (10µM)	1
5DEL – R (10µM)	1
Taq DNA Polymerase (5U/µl)	0.1
DNA (10ng)	1
water	5.9
Total	10

Master Mix 4

Component	X1 reaction (µl)
10X low dNTP Buffer (200µM)	1
5DEL – F[HEX] (10µM)	1
5DEL – R (10µM)	1
Taq DNA Polymerase (5U/µl)	0.1
DNA (10ng)	1
water	5.9
Total	10

Figure 12: PCR components for the two PRT (PRT107A and HSPD21) and indel (5DEL) assays.

Under the following thermocycling conditions for 107A/HSPD21:

Step	Temperature (°C)	Time	No. of Cycles
Initial Denaturation	95	5 min	1
Denaturation	95	30 sec	22
Annealing	58	30 sec	
Extension	70	1 min	
Repeat Step 3+4			
Annealing	58	30 sec	1
Final Extension	70	40 min	1

And these for 5DEL:

Step	Temperature (°C)	Time	No. of Cycles
Initial Denaturation	95	5 min	1
Denaturation	95	30 sec	25
Annealing	58	30 sec	
Extension	70	1 min	
Repeat Step 3+4			
Annealing	58	1 min	1
Final Extension	70	20 min	1

2.6.1 Capillary electrophoresis:

For each sample, 0.7µL of each PCR product was mixed with 0.1µl of an internal size standard MapMarker® 400 (BioVentures, Inc.) and 10µl of Hi-Di Formamide (Applied Biosystems, UK) in each well of a 0.2mL flat deck detection 96 well plate (Thermo Scientific). The samples were denatured for 3 minutes at 96°C and immediately placed on ice for at least 2 minutes and then resolved using POP-7 polymer with injection time of 30 seconds on 3130xl Genetic Analyser (Applied Biosystems, UK).

2.6.2 Data analysis:

The electrophoresis output files were analysed on GeneMapper® software v.3.7 (Applied Biosystems, UK). The raw data and the images from the output files were sized into fragment length peaks for the triplex PCR amplicons at the positions 157bp/155bp for the PRT107A assay, 180bp/172bp for the HSPD21 assay and any combination of the 134bp/129bp and 127bp for the multi-allelic indel rs5889219.

Only peaks with area size between 500 and 40000 were considered for copy number analysis. Samples that gave no peaks were repeated twice, if still no peaks were detected; the samples were considered ‘failed to type’.

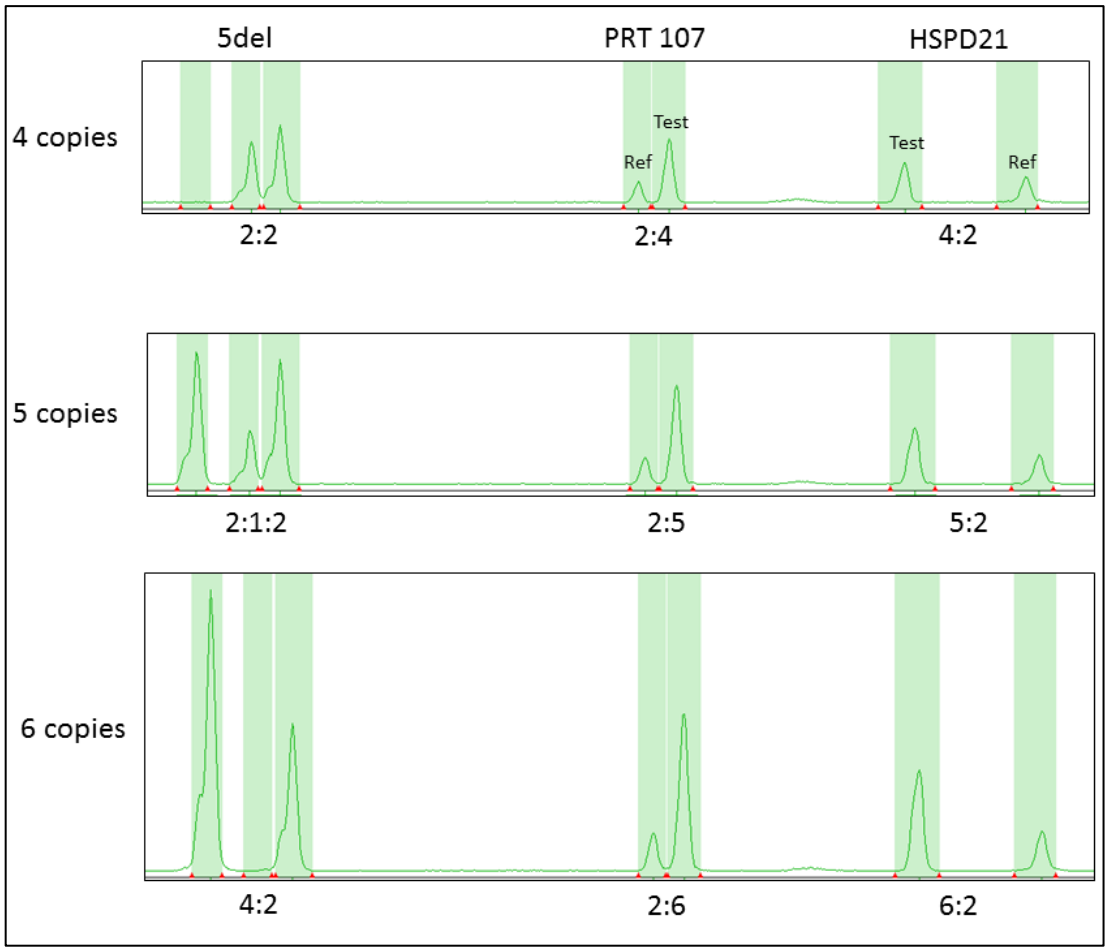


Figure 13: This figure shows the HEX-traces obtained from the multiplex system, (two PRT assays; PRT107A and HSPD21) and one indel assay (rs5889219). Predicted copy number for each individual in these traces from this system is shown to the left. Inside the box, there are two peaks visible for each PRT; one from the reference locus and one from the variable (test) locus. The genomic ratio of test: reference copies for each PRT are shown below the peaks. For the indel, there are up to three distinct alleles (123, 125 and 127bp) detected in this experiment. The genomic ratio from each peak to the smallest peak is also shown here below the peaks.

2.6.3 Data normalization:

The raw paralogue ratios (test/reference) calculated from each dye for the duplex PRT assays were corrected for PCR plate batch effects. The internal controls included on each PCR run of the assay served as gold-standards of known DEFB CN to calibrate the rest of the plate.

Data normalisation was performed in Excel by linear regression of the control sample (2.6) ratios against their expected integer copy numbers as shown in Figure 14 below.

The resulting regression equation was used to adjust all other PRT data for each of the two dyes for the 107A and the HSPD21 assays. The 5DEL assay was not calibrated as this assay does not infer an absolute copy number, rather it represents a factor of the true integer copy number (i.e. a 5DEL ratio of 2:1:0 could be any copy number multiple of 3).

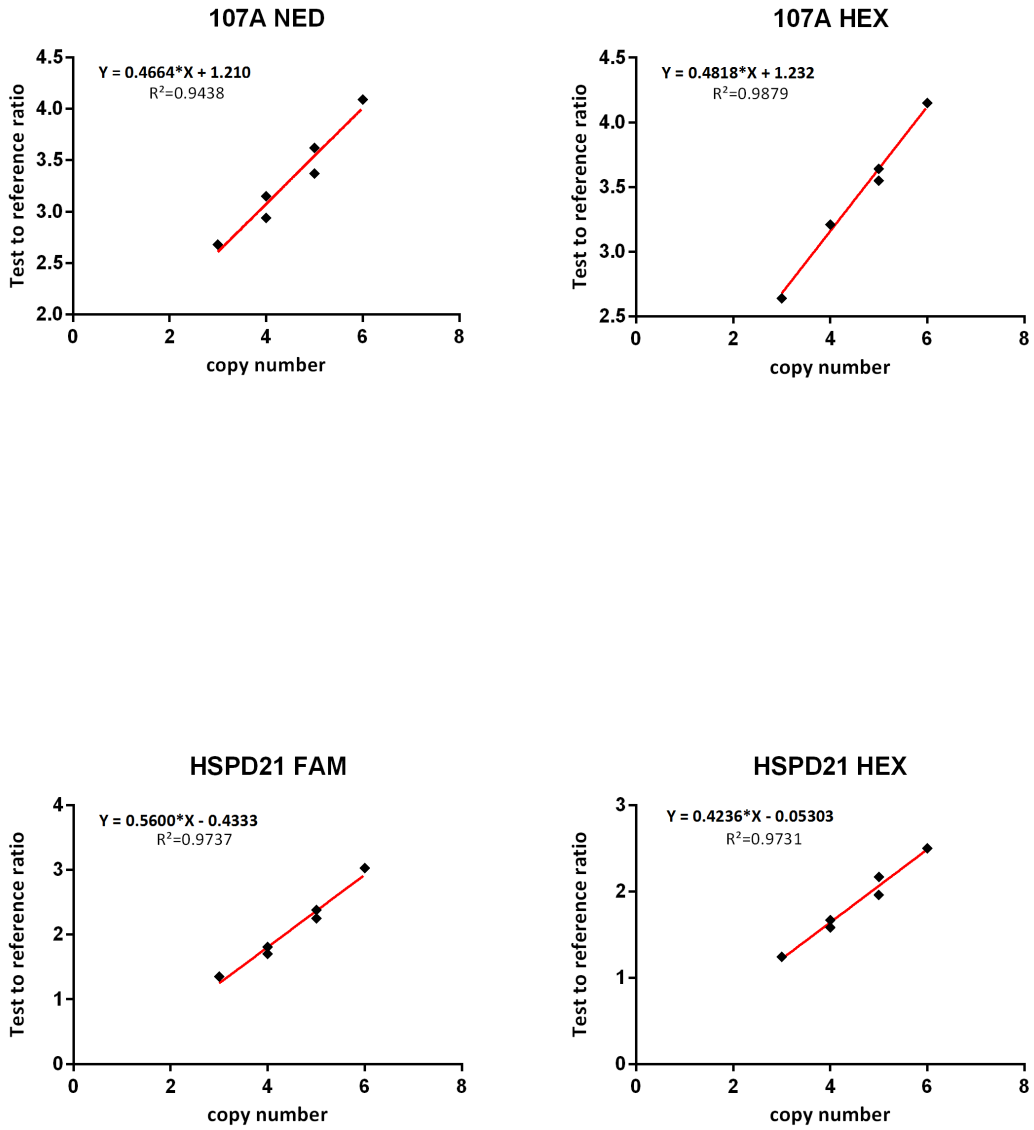


Figure 14: Scatter plots showing the calibration carried out for each experiment with the 6 reference DNA samples of known DEFB copy number for the duplex PRT assay (107A top line, and HSPD21 bottom line). The unrounded test to reference ration of each reference sample was plotted against the corresponding copy number. The linear regression obtained from each experiment was then used to infer the copy numbers of the unknown samples.

2.6.4 Estimation of DEFB CN using Maximum Likelihood analysis:

For DEFB CN estimation, the Maximum Likelihood (ML) approach was adopted from Aldhous et al. 2010. In summary, the likelihood method considered the joint probability of all generated data for a given sample across a range of copy numbers from 1 to 9 with error following the Gaussian distribution for each PRT assay. For each set of data from the triplex test, a copy number maximizing the likelihood of all the observations was calculated and a copy number called.

An estimate of the degree of confidence in each diplotype is calculated by $-2(\ln b - \ln a)$, where (*a)*is the probability of the most likely copy number diplotype, and (*b)*is the combined probability of all other copy number diplotypes (from 1 to 9). This statistic follows a χ^2 distribution, so P values for each copy number diplotype call can be determined using χ^2 tables.

2.6.5 Deduction of DEFB CN using the weighted mean raw PRT values:

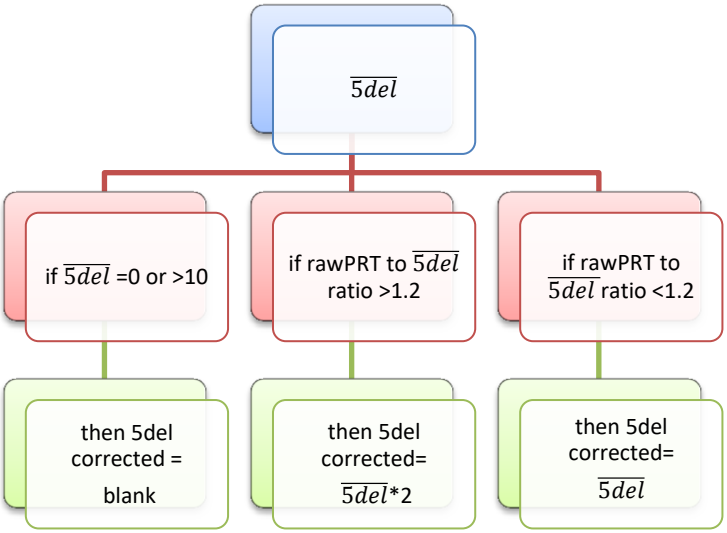
Alongside the DEFB CN estimated using the ML approach, an alternative method to confer DEFB CN was employed by Walker et al., 2009; the weighted mean integer value (hereafter referred to as raw PRT) was also used to infer DEFB CN.

The weighted mean value for the PRT assays (without the 5del assay) was calculated as follows:

(PRT107 average value*0.447) + (HSPD21 average value*0.553).

The numbers 0.447 and 0.553 are considered weightings which are proportional to the reciprocal of the variance of the particular assay (PRT107 and HSPD21 respectively) for a large dataset generated by Dr. Rob Hardwick and Helen Bogle.

For the 5 del assay; if the average value is equal to 1 or larger than 10, then the average value is not taken into account and kept as blank, however if the ratio of the value of the PRT weighted mean to the 5del average value is > 1.2, then the 5del corrected value is taken as 5del average value multiplied by 2. If not, then the 5del average value as taken as the corrected 5del value.



Ultimately, the final weighted mean average is calculated as follows:

If the 5del corrected value is a blank, then the weighted mean PRT is used as the final weighted mean CN. However, if 5del corrected has a value, the weighted mean CN is calculated as follows:

(0.354*5del corrected value) + (0.357*HSPD21 average value) + (0.289*PRT107 average value.

2.7 Digital Droplet PCR (ddPCR)

ddPCR is a method for performing digital PCR that is based on water-oil emulsion droplet technology. A sample is fractionated into 20,000 droplets and PCR amplification of the template molecules occurs in each individual droplet, providing an absolute count of target DNA copies per input sample.

A duplex TaqMan PCR-based assay was carried out for genotyping the CN of DEFB using *RNaseP* as the reference gene. Primers and probes for this assay were adapted from previous work (Fode et al., 2011). The sequences of the primers and probes were as follows:

[4] $\overline{5del}$ denotes the 5del assay average value.

Primer and/or Probe	Sequence (5' → 3')
DEFB103-F	CAT AGG GAG CTC TGC CTT ACC A
DEFB103-R	TGC AGA ACA CAC CCA CTC ACT C
DEFB103 probe	FAM - TGG GTT CCT AAT TAA C – MGB

The TaqMan® RNase P Control Reagents (VIC™ dye) kit (Applied Biosystems) containing a mixture of primers and probe for the *RNaseP* gene was used as the positive control in this experiment. Optimisation of the ddPCR protocol using the gold standard controls was performed to assess best method of DNA preparation and concentration. This included using *Hae III* restriction enzyme digested template, undigested template, templates of different concentrations as well as running a gradient PCR to pinpoint optimum annealing temperature.

The PCR reactions were carried out using the following components in a final volume of 22µL:

Component	Amount (µL)
2X TaqMan® Universal PCR mix	11
RNase P (900nM Primer/250nM) mix	1.1
DEFB103 (900nM Primer/250nM) mix	1.1
Template (5ng/µl)	1
RNase-free Water	7.8

Each plate contains a "no template control" and the six gold-standards used in the PRT assay. 20µl of the reaction mix was placed in a droplet generator cartridge (Biorad), in dedicated sample wells. 70µl of droplet generation oil (Biorad) was placed in the oil dedicated wells. The cartridge was then placed in a QX100 droplet generator (Biorad), a device using microfluidics to combine oil with the sample reaction to generate ~200000 monodispersed, nanolitre-sized droplets for each sample, suitable for ddPCR. 40µl of the emulsion obtained per sample was then transferred to a 96-well plate (ThermoScientifics), sealed with aluminium foil (Thermo Scientifics,) using a plate sealer device (Eppendorf) and amplified to end point using a standard thermal cycler, using the following protocol:

Step	Temperature (°C)	Time	No. of Cycles
Initial Denaturation	95	10 min	1
Denaturation Annealing/extension	94 58	30 sec 60 sec	40
Final Extension	98	10 min	1

The plate was then loaded on a QX100 droplet reader (Biorad) and results were analysed using the QuantaSoft software (Biorad). Reads with more than 10000 droplets generated were considered acceptable. Each gating separating the reads in *DEFB103+/RNaseP-*, *DEFB103-/RNaseP+*, *DEFB103-/RNaseP-*, *DEFB103+/RNaseP+* was manually checked to ensure the presence of a discrete separation between the different droplet populations. The ratio of the total number of single positive droplets for *DEFB103* on the total number of single positive droplets for *RNaseP* called the diploid copy number for each sample in analysis.

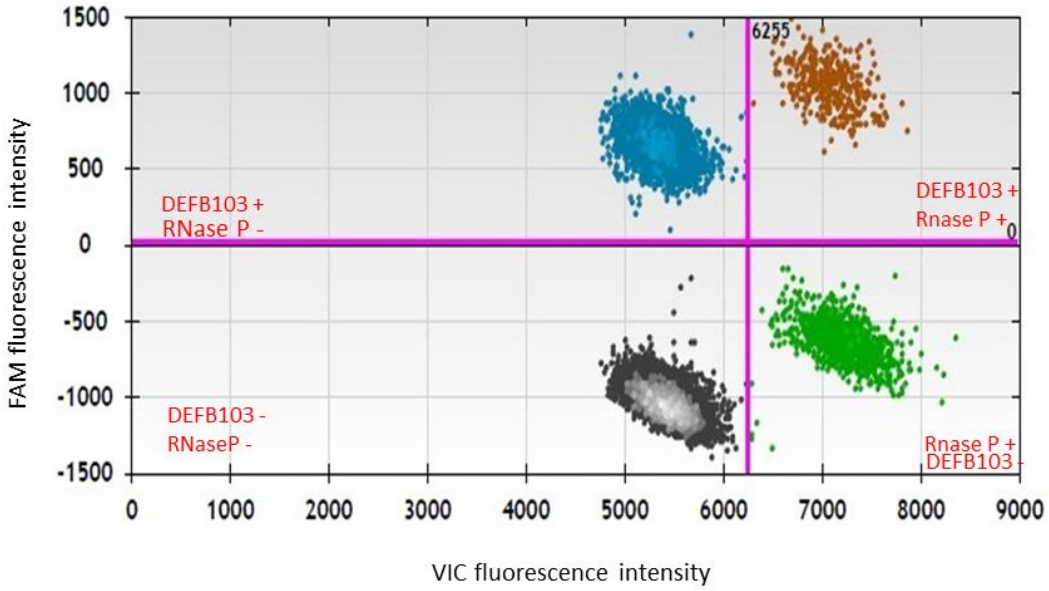


Figure 15: example result from a ddPCR experiment showing 2-D plot of droplet fluorescence. Black cluster shows droplets with neither reference nor target gene. Green cluster shows droplets with positive reference gene. Blue cluster shows droplets with positive target gene, and the brown cluster shows droplets that have both reference and target genes.

2.8 Pneumolysin treated cell lines (PnCells)

Pneumolysin is a pore forming toxin that binds cholesterol in host cell membranes. It is released from *Streptococcus pneumonia* during respiratory infection by autolysis of the bacteria. The cell lines were cultured by our collaborator, Dr. Rob Hirst as described below:

Ciliated primary human airway cultures (NHBE were grown from brush biopsies in basal epithelial growth media (BEGM) at 37°C. Basal cells were seeded onto collagen-coated Transwell inserts (Corning) and cultured using differentiation medium (50:50 BEBM and DMEM) and 100nM trans-retinoic acid. The cells were then incubated with different concentrations of pneumolysin (42.5 Hu and 170Hu)for 20 minutes at 37°C, pelleted, placed in a slow freeze cryopreservation container at -80^{O}C overnight before being moved to long term storage in a -150^{O}C freezer, and then given to us. Cells were used for DNA and RNA extraction and the supernatant was used for protein quantification using enzyme-linked immunosorbent assay (ELISA).

The experiment was carried out two times, each time with a different batch of NHBE cells. Each batch was passaged three times.

2.8.1 RNA and DNA extraction from Pn treated Cells

DNA and RNA were extracted by the Maxwell® 16 instrument using the Maxwell® 16 Blood DNA Purification Kit and MaxWell® 16 LEV simplyRNA Cells Kit (Promega) according to the manufacturer's instructions.

2.8.1.1 Quantification of Extracted RNA

The RNA concentration and purity was determined using ThermoScientific NanoDrop™ 1000 Spectrophotometer.

The ratio of absorbance at 260nm and 280nm is used to assess the purity of the RNA with a ratio of ~2.0 generally accepted as "pure" RNA. If the ratio is appreciably lower, it may indicate the presence of protein, phenol or other contaminants that absorb

strongly at or near 280nm. After quantification, all RNA samples were diluted to 5ng/µl.

2.8.1.2 Reverse transcription–PCR (RT-PCR) of extracted RNA

30ng of RNA was transcribed into cDNA using Invitrogen's SuperScript III First – Strand Synthesis Super Mix kit as per the kit's instructions. Total RNA isolated from Pn treated cell lines were primed for first-strand synthesis by using the random hexamer primer provided with the kit in a total volume of 20µl. Random hexamers are the most nonspecific priming method, and are typically used when the mRNA is difficult to copy in its entirety. With this method, all RNAs in a population are templates for first-strand cDNA synthesis, and PCR primers confer specificity during PCR.

The following components were mixed thoroughly by pipette on ice:

Component	Amount (µL)
Total RNA (5ng/µl)	6
Random Hexamer (50ng/µl)	1
Annealing Buffer	1
DEPC water	0
Total	8

The mixture was incubated at 65°C for 5 minutes, then immediately placed on ice for at least a minute. Briefly centrifuged and added the following on ice:

Component	Amount (µL)
2X First-Strand Reaction Mix	10
SuperScript III®/RNaseOUT™ Enzyme Mix	2

The samples were then incubated at 25°C for 10 minutes followed by 50 minutes at 50°C. The reaction was terminated at 85°C for 5 minutes and cDNA was stored at -20°C for later use.

2.8.2 Expression studies

2.8.2.1 Quantitative PCR (qPCR)

The cDNAs were firstly diluted down to 1:7 ratio and analysed by qPCR using TaqMan®Expression Assays for *DEFB4* (Hs00175474_m1),*UBC* (Hs00824723_m1) and *PPIA* (Hs04194521_s1). PCR was performed using TaqMan® Universal Master Mix II, with UNG (Applied Biosystems) following the manufacturer's manual on LightCycler® 480 (Roche). The protocol was Hold at 50°C for 2 minutes, followed by 40 cycles of 95°C for 15 seconds and 60°C for 1 minute. PCR was performed in quadruplet for each sample and data analysed using the qpcR library in R (Spiess, 2014).

2.8.2.2 Enzyme Linked Immunosorbent Assay (ELISA)

Prior to the ELISA technique, total protein in the supernatant was measured using Bradford Protein Assay (Bio Rad Laboratories), then diluted down by a factor of 4 before being used in the ELISA.

96-well ELISA plates (NuncMaxisorp, ThermoScientific) were coated with 0.5µg/mL goat anti-hBD-2 (Human BD-2 ELISA Development Kit, PeproTech®) at room temperature overnight. The plates were then washed with 0.05% Tween 20 in PBS and blocked with 1% BSA in PBS for 1 hour at room temperature. PnCells supernatants and a serial dilution of hBD-2 standard (Human BD-2 ELISA Development Kit, PeproTech®) from 2000pg/mL to 0pg/mL were added to the ELISA plate in duplicate and incubated for 2 hours at room temperature, followed by 2 hours incubation with 0.5µg/mL biotinylated goat anti-hBD-2 detection antibody. Plates were thoroughly washed and incubated with Avidin-HRP at a 1:2000 concentration for 30 minutes at room temperature. The plates were then washed and 2,2'-Azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) liquid substrate solution (ABTS) (Sigma) was added and monitored for colour development with ELISA plate reader (FLUOstar Omega, BMG LABTECH) at 405 nm with wavelength correction set at 650 nm at 5-minute intervals for half an hour. Results were examined and those with a reliable standard curve were analysed. A reliable standard curve had O.D readings which do not exceed 0.2 units for the blank or 1.3 units for the 2000pg/mL concentration.

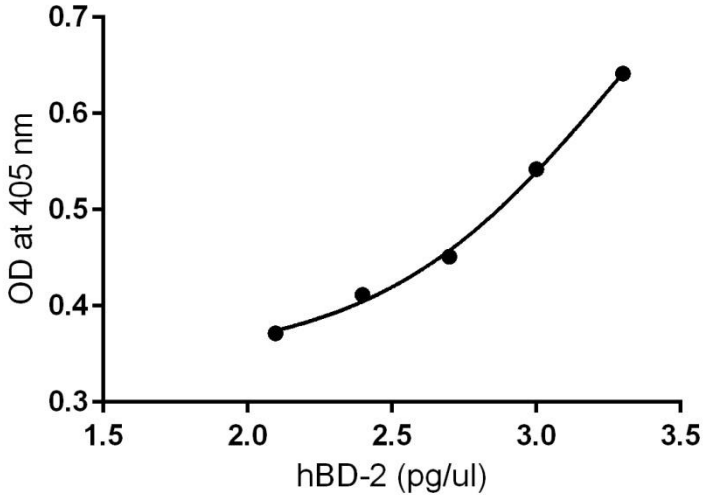


Figure 16: Example ELISA standard curve Absorbance was measured at the relevant OD and protein concentration determined by standard curve.

Sample concentrations were derived from standard curves, and analyses carried out in GraphPad Prism v. 6 software.

2.9 Bioinformatics analysis

2.9.1 Comparison of different CNV calling methods

HapMap and Human Random Control (HRC) samples that have been typed for DEFB CN using PRT were compared to samples typed by: array comparative genome hybridization method (NimbleGen and Agilent platform); digital-detection technology (nCounter); sequence read-depth method (Genome STRip) and PCR based (ddPCR). All calculations and figures generated for comparison purposes were done using 'R' statistical computing program.

2.9.2 NimbleGen aCGH breakpoints

NimbleGen data was kindly provided by our collaborators; Dr. Jacky McArthur and Dr. Donna Albertson from the University of California, San Francisco. NimbleGen arrays with 60-mer oligonucleotide probes all across chromosome 8 were used as indicated

by the NCBI36/hg18 Human Genome Assembly on USCS Genome Browser website (http://genome.ucsc.edu/).

Probes falling within the DEFB regions were checked for gains and losses using the 300bp and 600bp tracks on SignalMap V1.9 software (NimbleGen Systems) for 14 HRC samples and 54 HapMap samples, taking into account gains and losses using the 300bp and 600bp tracks and recorded them on an excel sheet. The regions were then checked for genes and/or repeat elements on the NCBI36/hg18 USCS genome browser. Interestingly, 12 samples showed a loss in the *DEFB107* gene. They were further investigated using a 3-Primer assay as follows:

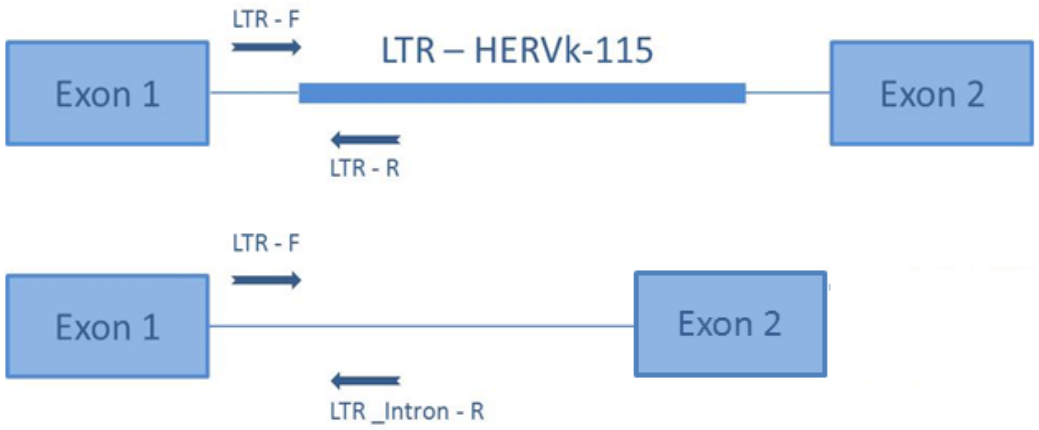


Figure 17: Schematic diagram for the location of the 3 primers used in further investigating the hERV-k115 polymorphic insertion.

The Primer sequences used

Primer	Sequence (5' → 3')	Product length (bp)
LTR – F	AGCCTCATTTAACTTTGGTGC	1189
LTR – R	GGCTATGAAGCAATGGCCTA	
LTR_Intron – R	GAGGTCACTGTGATCAAAGAT	408

The PCR was done using the master mix and thermocycling program described in 'Generic PCR Optimization' (2.4) above with annealing temperature of 62°C. The amplicons were then resolved on 2% agarose gel at 120V for 1 hour.

2.9.3 DEFB CNV region analysis using NimbleGen aCGH

A custom NimbleGen tiling oligonucleotide array (NimbleGen Systems Inc.) was designed with 190,240 probes covering the DEFB region (chr8:6,185,000-84,681,910; hg18). Probe design, array fabrication, and array CGH experiments, including DNA labelling, hybridization, array scanning, data normalization, and $\log_2$ copy-number ratio calculation were performed by NimbleGen Systems Inc. Array CGH was carried out on 68 cell line DNA samples with cell line AF0105 (Hollox et al. 2005) used as reference DNA for all hybridizations.

2.10 Statistical analysis

Table 5 below describes the different platforms used in carrying out the required statistical analyses:

Table 5: Platforms used to carry out statistical analyses.

Software	Used for	Available from
SPSS v. 22	GLM, regression variable plot	Provided by the university
GraphPad Prism v. 6	Categorical contingency tables, ELISA analysis	Provided by the university
R Statistical program v. 3.2.0	Mosaic plots, PCA calculations, histograms in scatter plots, histogram & Kernel density plots, Manhattan plots, Power analysis and qPCR analysis.	http://www.r-project.org/
Plink v. 1.0.7	SNP analysis and haplotype associations	http://pngu.mgh.harvard.edu/~purcell/plink/
Genome Graphs	Plotting probe positions	https://genome.ucsc.edu/cgi-bin/hgGenome

3 Genomic structure and variability of β-Defensin region

3.1 Introduction

β-defensin genes in humans are arranged into 3 main clusters at 8p23.1 (9 genes), 20p13 (5 genes) and 20q11.1 (9 genes), with an additional small cluster of 4 genes on chromosome 6p12 (Ganz, 2003; HGNC, n.d.). At the 8p23.1 genomic location, 7 DEFB genes are found on a repeat unit that is typically present at 2–8 copies in the population, with a modal CN of 4 (Hollox, 2012; Hollox, Barber, et al., 2008). The mutation rate at this locus is extremely fast (~0.7% per gamete) (Abu Bakar et al., 2009), suggestive of the high level of variability in this genomic region. Individuals with 1 copy are very rare (Zhang et al., 2014). In our previous study; (Wain et al., 2014), we determined DEFB diploid CN in a cohort of 1149 adults and in a separate cohort of 689 children using PRT. The distributions of copy number observed in each cohort were in good agreement with previously published distributions from the UK population (Fode et al., 2011; Hollox et al., 2003), although in the children cohort we observed 9 children (>1%) with a copy number of 1, which in other cohorts is rare. This indicates that the existence of a null allele could be deleterious and selected against. At the other end of the DEFB CN spectrum lie rare high copy individuals (9–12 copies) with a cytogenetically observable CN amplification at 8p23.1 that has no clinical phenotypic effect (Barber et al., 1998).

3.2 Study rationale

10 genes and pseudogenes (*DEFB109, DEFB108, DEFB4, HSPD1P, DEFB103, SPAG11, DEFB104, DEFB106, DEFB105 and DEFB107*) are arranged in a cluster of almost 218 kb enclosed in various kinds of low copy repeats (LCR) arising from segmental duplications (SDs) (Ottolini et al., 2014; Hollox, Barber, et al., 2008). Moreover, this DEFB cluster is embedded in one of two complex SD rich regions involving retroviral elements and olfactory repeat (OR) regions, collectively known as “REPD” (for repeat distal) (Giglio et al., 2001) and another smaller OR region called “REPP” (repeat proximal), 5 Mb

proximal on 8p23.1 which shares a high level of identity with REPD (Sugawara et al., 2003). It is due to this repetitive structure which is intractable to analysis, the locus is one of the few regions with a remaining recalcitrant gap in the most recent human genome assembly; hg38, December 2013 (Taudien et al., 2014). In the hg18 assembly, two clusters, DEFB_R1 and DEFB_R2 (chr8:7,170,368 – 7,366,833 and chr8:7,669,242 – 7,855,043, respectively) are arranged in opposite direction on both sides of the gap (Figure 18). The reason the hg18 genome assembly was used is because the NimbleGen data that was provided to us was based on probes that matched the hg18 build.

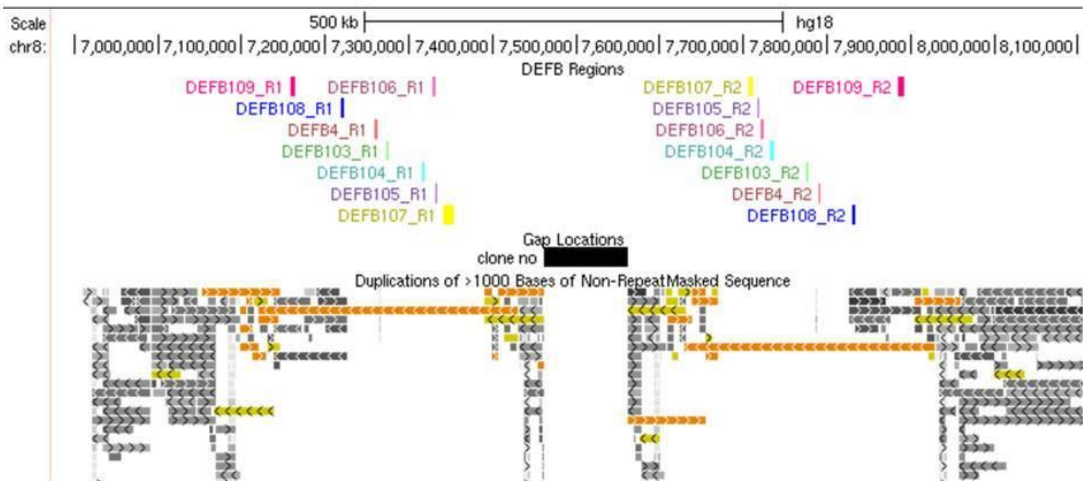


Figure 18: Position of the DEFB regions on chromosome 8p23.1. As dictated by the NCBI hg18 build, two gene clusters are arranged on both sides of a gap and in opposite direction as it is clear by the individual gene names. Segmental duplications are flanking the clusters and/ or are part of it. The segmental duplication track shows regions detected as putative genomic duplications within the golden path. The following display conventions are used to distinguish levels of similarity: Light to dark grey: 90 - 98% similarity, Light to dark yellow: 98 - 99% similarity, Light to dark orange: greater than 99% similarity. Red: duplications of greater than 98% similarity that lack sufficient Segmental Duplication Database evidence (most likely missed overlaps). For a region to be included in the track at least 1 Kb of the total sequence (containing at least 500 bp of non-RepeatMasked sequence) had to align and a sequence identity of at least 90% was required (Bailey et al., 2001, 2002).

Mapping of the β-defensin CNV region has been challenging, however, determining the size of the DEFB CNV region as reviewed by Taudien et al. (2014), was aided by

combining information of the probe positions of the MLPA technique (Groth et al., 2008) and PRT technique (Aldhous et al., 2010) as well as a mixture of microsatellite and multiallelic length polymorphism analyses carried out by Abu Bakar et al., (2009).

At the start, the minimal length of the CNV region was approximated to be a minimum of 240 kb by pulsed-field gel electrophoresis of digested genomic DNA after hybridization with *SPAG11* and *DEFB4* probes (Hollox et al., 2003) however, CNV is confirmed for 115 kb between the most distal MLPA probe in *DEFB4* and the most proximal PRT107A amplicon of PRT (Taudien et al., 2014). At the time, it was not known whether the pseudogenes *DEFB108* and *DEFB109* were included in the CNV region.

In 2014, Taudien at al. established tests for length polymorphisms based on amplification and capillary electrophoresis of 7 indels spread over the entire cluster. CNV was demonstrated for 6 indels between ~ 1 kb distal of *DEFB108* and 10 kb proximal of *DEFB107*, fixing the minimal length of proven CNV to 157 kb including the *DEFB108* pseudogene but excluding *DEFB109* (Taudien et al., 2014). The limitation of this technique is having defined the boundary of the DEFB repeat region by 1 indel in a region that is too complex and polymorphic. Also, indels 1 and 2 showed chimera in 7% and 4% of their sequences most likely due to the co-amplification of paralogues on chromosome 4,8,11 and 12 which gives rise to high variability in the results, especially on the proximal boundary.

The 8p23.1 region is also home to the largest polymorphic inversion known in the human genome (Antonacci et al., 2009) —a ~4.5-Mb-long inversion that was conventionally genotyped using fluorescent in situ hybridization (FISH) (Giglio et al., 2001). Bosch et al. (2009) using Spanish individuals who were found to be homozygous for the 8p23.1 inversion, and by means of dense SNP genotyping of the region, haplotype-based computational analyses and FISH techniques, was able to delineate two different homozygosity tracts composed by a total of 16 SNPs within the inverted region, in the proximity of the REPP and REPD duplicons; The first homozygosity tract is located at chromosome position 8.5Mb, close to the REPD SDs that flanks the 8p23.1 region, and expands ~ 172 kb. It contains 8 SNPs (rs17627505, rs10503393, rs2428,

rs11774860, rs3827811, rs17154769, rs1876836 and rs1039916) which corresponds to the ‘‘CGTCGAGG’’ haplotype. The second tract spans almost 181 kb and lays at 10.8Mb close to the REPP. The conserved haplotype is “TCACGAGA” and constitutes the remaining 8 SNPs (rs1178061, rs1178247, rs3885690, rs2409691, rs13266785, rs10282848, rs10503417, and rs2409719). Bosch and colleagues postulated that these two 8- SNP haplotypes can be used as proxies for the 8p23.1 inversion.

The inversion shows a strong clinal distribution in human populations, with frequencies of ~59% in the Yoruba, ~20%–50% in Europeans, and ~12%–27% in Asians (Salm et al., 2012). Though originally considered a neutral polymorphism and despite the high frequency of inversion heterozygotes, the prevalence of the recurrent inverted, duplicated, and deleted 8p has been estimated at only 1 in 20,000 (Hollox, Barber, et al., 2008). This ectopic recombination has an associated phenotype of developmental delay, mental retardation, facial dysmorphisms, agenesis of the corpus callosum, and other problems including congenital heart disease (Devriendt et al., 1995). This phenomenon has also been described in the parents of children carrying other genomic disorders (Lupski, 1998) such as Hunter syndrome (Bondeson et al., 1995) and Williams-Beuren syndrome (Bayés et al., 2003).

FISH is not amenable to high-throughput analyses and requires viable cells. Moreover, the size of the inversion’s single copy region (~4.5 Mb) approaches the practical resolution of metaphase FISH (Raap, 1998); in 2012, Salm and colleagues presented a novel, robust high-throughput method to genotype 8p23-inv.; Phase Free Inversion Detection Operator (PFIDO). PFIDO is a bioinformatics tool that utilizes SNP data to enable accurate, high-throughput genotyping of the 8p23.1 inversion. Salm and colleagues applied this tool retrospectively to diverse genome-wide SNP genotypes datasets and were able to infer the orientation of the 8p23-inv. The inversion was genotyped as the alleles *N* and *I*; *N* stands for non-inverted allele and refers to the orientation represented in human genome reference (hg19) and *I* to the inverted allele (Salm et al., 2012).

In this chapter, three fundamental questions relating to the DEFB CNV region will be answered; firstly, pinpointing the exact boundaries of the DEFB CNV region. This is

essential because it might give insight into the mutational history of the locus and aid in comprehending if the region is closely tied to the polymorphic recurrent inversion of 8p23.1, which is known to involve the REPP and REPD regions. Secondly if the DEFB region reveals any heterogeneity, this is of significance because it sheds light on whether a disease association is present regardless of the DEFB region state or is it conditional to it. And lastly, if there is a specific SNP/SNPs associated with and or can predict a certain copy number or a copy number cut-off using genome wide SNP genotypes. SNPs are well genotyped and hence a plethora of information is available that can be used in algorithms which enables prediction of DEFB CN on very large datasets without direct CN genotyping.

3.3 Results

3.3.1 NimbleGen array data analysis reveals polymorphic retroviral insertion in *DEFB107*

To address this, a tiling oligonucleotide array with 190,240 probes covering the region between chr8:6,185,000 and chr8:84,681,910 was designed and aCGH performed on a series of 68 DNA samples with known DEFB copy number. Initial inspection confirmed the CNV of the DEFB region, together with the more complex CNV of the REPD region flanking the DEFB region.

NimbleGen data viewer (SignalMap software) has a window averaging step which correspond to 10 (300 bp), 20 (600 bp) and 50 (1500 bp) times the median probe spacing on the design. Probes that fall into a defined base pair window size are averaged and a new position is assigned to this, which is the midpoint of the window. This averaging has two net effects; it reduces the size of the dataset, resulting in faster computation times, and it reduces the noise in the data. Large window sizes results in less noise, but less sensitivity when it comes to finding smaller segments.

The 300bp and 600bp tracks for all probes within both DEFB regions as dictated by the hg18 build was visually checked for 164 HapMap samples using the SignalMap software. Only ‘gains’ and ‘losses’ appearing on both the 300 bp and 600 bp tracks (Figure 19) for each sample were recorded (40 samples). Areas of gain and loss were

specified by Log_2 transforming the ratio values of the probe signal intensities of (Cy3/Cy5). See Appendix 1

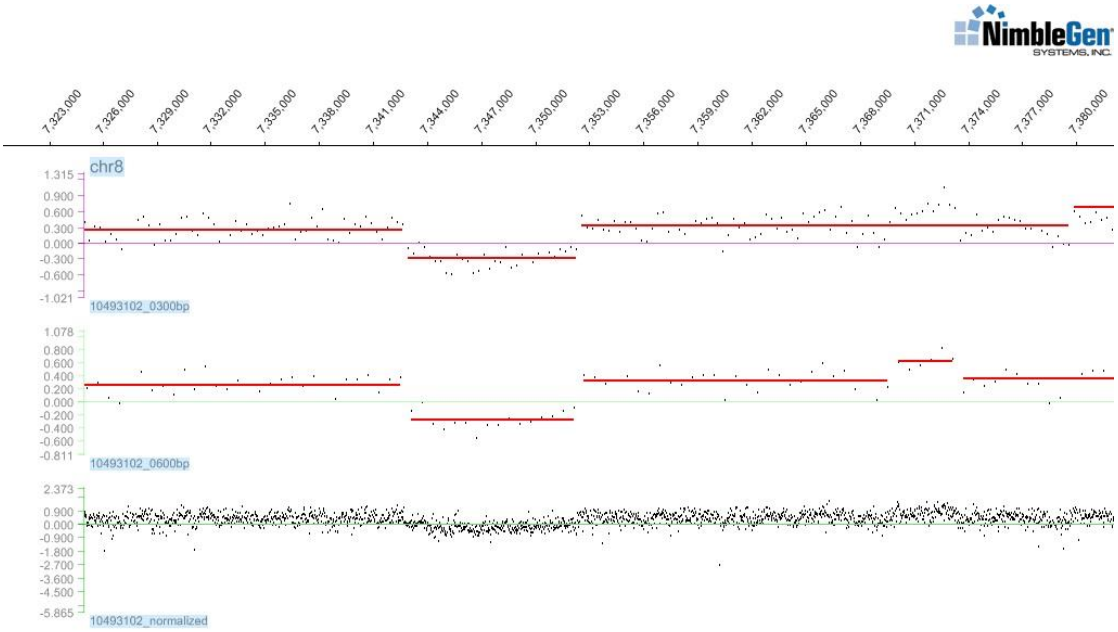


Figure 19: Data displayed in SignalMap software. The 300 bp, 600 bp and normalized signal of sample NA19204 is shown, the red solid lines show the best fit line through the signals indicating areas of gain or loss.

Of the 40 samples, 12 showed a presence/absence polymorphism of an LTR element in the intron of *DEFB107* confirming the results of Turner et al., (2001) which revealed an insertional polymorphism of an endogenous retrovirus, hERV-k115. In an attempt to further explore the hERV-k115 polymorphism, a 3 primer assay using two different primer pairs; 'LTR_R' and 'LTR_Intron-R' (2.9.2) was carried out on two HRC samples; co888 revealed a loss in *DEFB107* gene as illustrated on SignalMap software tracks and co007 that showed no breakpoints on the track, i.e. neither gains nor losses in *DEFB107* gene

Figure 20 below shows the results. Both samples revealed a 408 bp band that reflects an absence of hERVk-115 insertion.

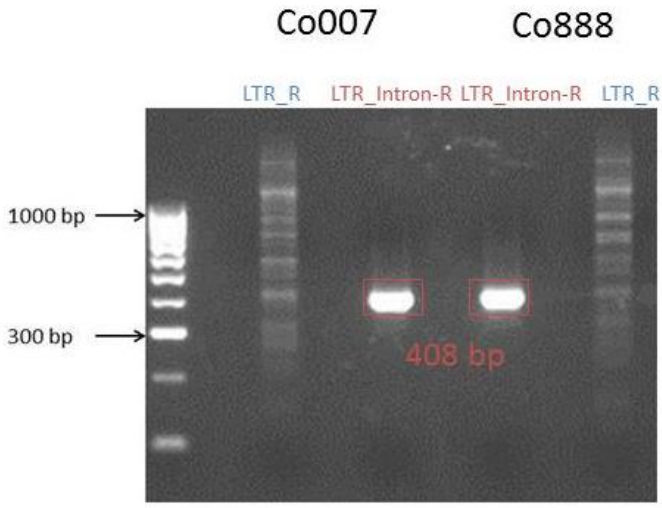


Figure 20: Results of 3-primer assay on samples Co007 and Co888. Both samples show the band revealing an absence of hERVk-115 (408 bp).

The first assay with a reverse primer binding to the LTR did not work as planned. With this limitation, we can only be positive of the absence of the LTR by the presence of the 408 bp band. Due to the heterogeneity of the DEFB region, we can only say that one of the many copies carried by a certain individual has or does not have the hERVk-115 polymorphism.

3.3.2 NimbleGen array data analysis to determine DEFB CNV region size

Approaches that detect CNV regions based on the transition from normal diploid copy number to variable copy number are unlikely to be effective in the DEFB regions because the DEFB CNV block is embedded within repeat-rich regions REPP and REPD. Instead, in order to test the extent of the human DEFB CNV contiguous block, the squared correlation coefficient (r^2) pairwise between the $\log_2$ ratio for each aCGH probe and the DEFB copy number determined by triplex PRT was calculated across all 68 samples. The reasoning behind this method is that intensity values from aCGH probes that are measuring the same CNV as the PRT will, on average, be strongly correlated with copy number measured by PRT, across a large number of samples. Equally, those intensity values from aCGH probes outside the CNV region measured by PRT will not be strongly correlated with copy number measured by PRT. Importantly, this last principle holds whether the aCGH probes map to a diploid non-CNV region or a more complex CNV unrelated to the CNV measured by the PRT.

These r^2 values were plotted against the two assembled DEFB repeats present in the human reference genome and showed a contiguous region of 322kb where the $\log_2$ ratio of the aCGH probes is correlated with the DEFB copy number (Figure 21).

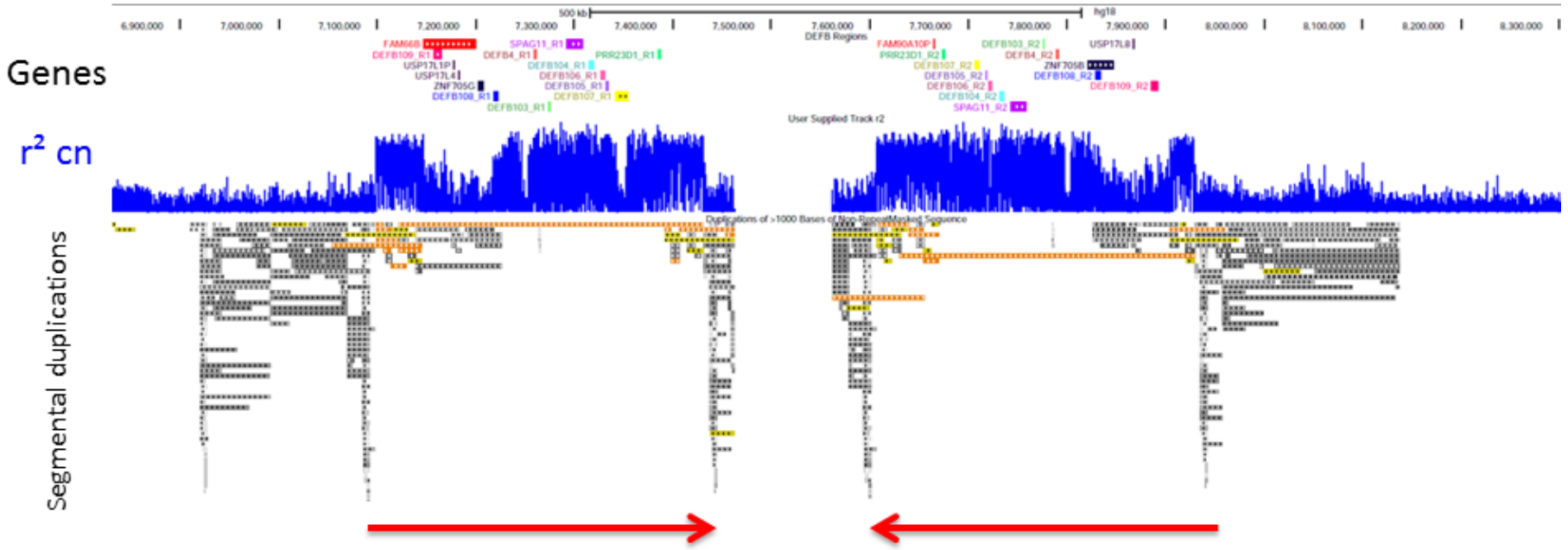


Figure 21: Analysis of CNV of DEFB regions. The correlation of each individual aCGH probe with the copy number of 68 samples estimated by PRT is shown as the track r^2 cn. DEFB genes mapping to the region, segmental duplications as defined by Bailey et al. (2001), and genomic position is also shown. The red arrows indicate the copy number variable repeat (322kb).

This region includes the DEFB genes *DEFB4, DEFB103, DEFB104, DEFB105, DEFB106*, and *DEFB107* as expected, as well the sperm-associated glycoprotein *SPAG11* and the proline rich 23 domain containing 1 gene (*PRR23D1*). *SPAG11* is related to the DEFB genes, and is both antimicrobial and necessary for the initiation of sperm maturation (von Horsten et al., 2004). *PRR23D1* is transcribed and predicted to encode a protein, as yet of unknown function. Other human *PRR23* family members (*PRR23A, PRR23B*, and *PRR23C*) are testis-specific genes, according to the RNA sequencing (RNA-Seq) data provided by Illumina BodyMap 2 and analysed by E. Hollox, strongly suggesting that this family has a role in the male reproductive system. Within the 322kb contiguous region is a small section which shows a lower level of correlation, due to it being comprised of a low copy repeat that also maps to chromosome 4p16.1, 11q13.4, and 12p13.31. This small section contains *DEFB109*, *FAM90A10* (Bosch et al., 2008), and *FAM66B* gene families, as well as a *ZNF705* gene, members of the KRAB - associated zinc-finger family of transcription factors (Huntley et al., 2006), and members of the *USP17L* family of deubiquitinating enzymes (Burrows et al., 2005).

3.3.3 Inversion status analysis

In total, 1009 HGDP and HapMap samples had matching data for DEFB CN and inversion status. The inversion status of 911 samples was deduced by PFIDO and the remaining 98 by FISH (Salm et al., 2012). The distribution of each DEFB CN according to the inversion genotype is shown in Figure 22.

Since there was no prior hypothesis on the nature of any association, a categorical contingency table was created for PFIDO and FISH separately, with each DEFB CN / Inversion genotype a separate cell in the table.

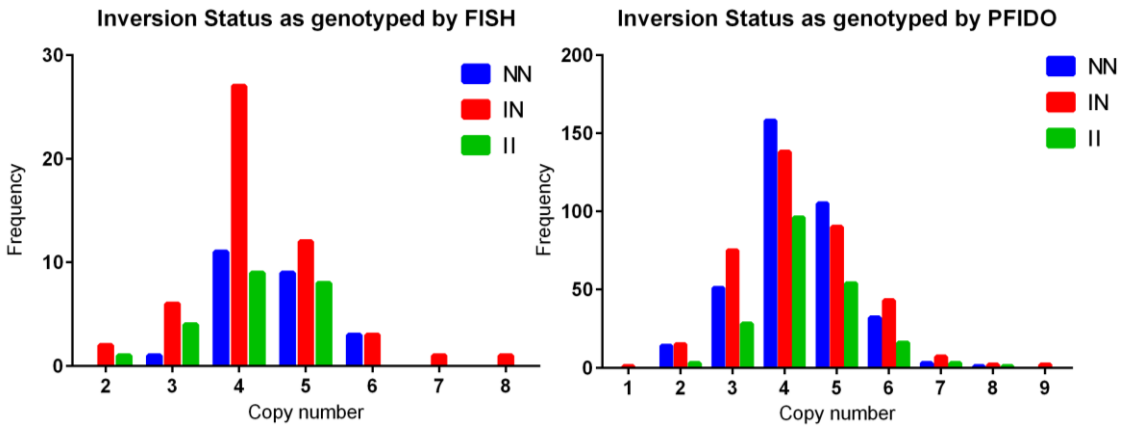


Figure 22: the frequency of each DEFB copy number according to Inversion status as genotyped by FISH and PFIDO. '*N*' stands for non-inverted allele and refers to the orientation represented in human genome reference (hg19) and '*I*' to the inverted allele.

A categorical contingency table analysis was then carried out. Because the table has more than two columns and two rows, only the chi-square test for trend is performed and Pearson's residual deviation from the expected cell count was checked. The p-value for samples genotyped by FISH was 0.638 and that for PFIDO was 0.142; both of which were not statistically significant and confirmed there was no relation between DEFB CN and inversion status.

3.3.4 DEFB CN region and SNP Linkage Disequilibrium

In this subsection, a subset of 891 SHCS samples (2.1.3.2) were used for the analysis, the reason being the availability of SNP information for chromosome 8 and matching DEFB calls using PRT for the cohort. Chromosome 8 genotypes were uploaded onto Plink software, and a list of 802 SNPs used by Salm et al. (2012), publically available online were extracted from the dataset. A linear regression association analysis was carried out using the DEFB CN as the phenotype and the SNP genotypes as a predictor, assuming an additive model.

The 3 SNPs with the highest p-values were:

Table 6: Results of the linear regression association analysis carried out using the DEFB CN as the phenotype and the SNP genotypes as a predictor for the three SNPs with highest p-value.

SNP	B- value	R^2	p-value
rs2979234	0.388	0.019	4.63×10^{-5}
rs17149723	0.215	0.017	1.22×10^{-4}
rs11774836	-0.188	0.017	2.03×10^{-4}

The Manhattan plot (Figure 23) for the above SNPs shows that the results have not achieved genome-wide significance. The low r^2 values show that each SNP individually has very little predictive power for DEFB CN. The corresponding Q-Q plot (Figure 24) of observed p-values against expected p-values show that the majority of points fall within the 95% CI, nonetheless, some points fall below the grey area which means that there are fewer 'significant observations' than expected by chance, however the difference is minimal. Hence we deduced that the above 3 SNPs are weakly associated with DEFB CN in this cohort.

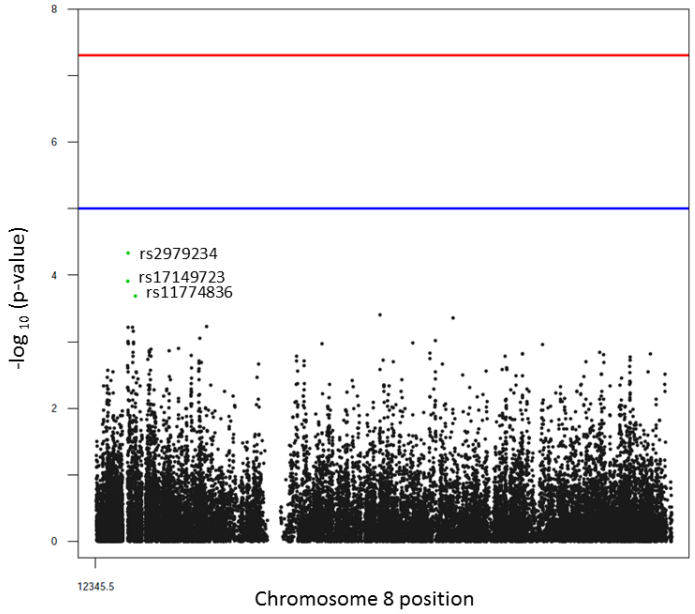


Figure 23: Manhattan plot for the genome-wide association study (GWAS) of SNPs across chromosome 8 in the SHCS cohort with DEFB CN. The large gap represents the centromere and the REPD and REPP are the tiny gaps on either side of the SNPs. The x-axis represents chromosome 8 position. The y-axis shows the p-value for association test at each locus on the log scale, the three SNPs with highest p-values are coloured in green. Suggestive threshold p $<10^{-5}$ (blue line) and Genome-wide p $<5 \times 10^{-8}$ (red line) are also shown.

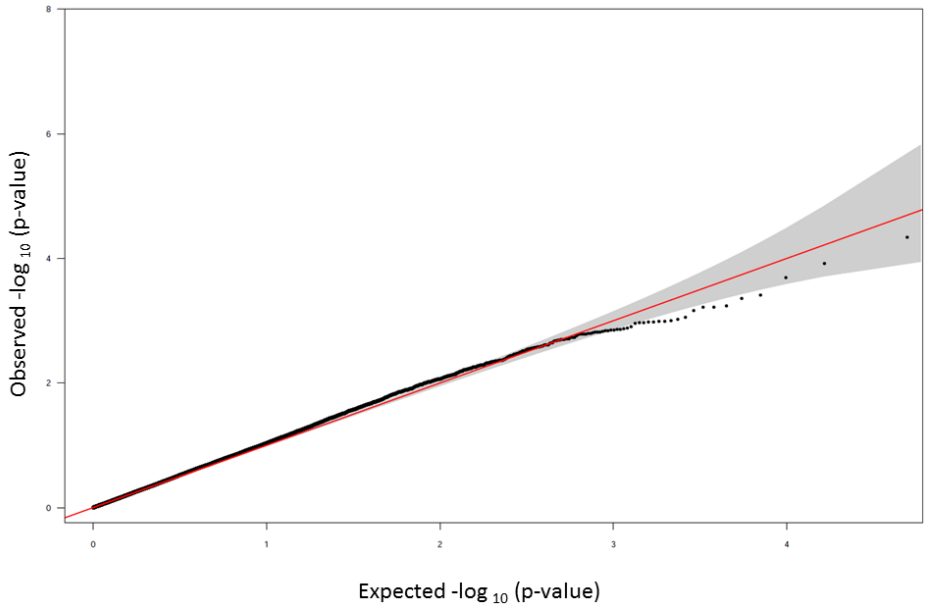


Figure 24: Q-Q-plot of observed p-values on the log scale against expected p-values on the log scale for DEFB CN and SNP association. The red line is drawn where y-axis value = the x-axis value, and the grey area represents the 95% CI.

Since our population is a sample of unrelated individuals, a haplotype-based association with a quantitative trait (DEFB CN) was carried out and the results summarised in Table 7 below

Table 7: Results of haplotype-based association with a quantitative trait (DEFB CN)

Haplotype	B-value	R^2	p-value	SNPs forming the haplotype
TAG	0.483	0.023	$8.3x10^{-6}$	rs2979234|rs17149723|rs11774836
GGA	-0.215	0.019	$5.7x10^{-5}$	rs2979234|rs17149723|rs11774836
GAG	0.115	3.6×10^{-3}	0.082	rs2979234|rs17149723|rs11774836
TAA	0.095	1.5×10^{-4}	0.726	rs2979234|rs17149723|rs11774836
GGG	0.019	1.4×10^{-4}	0.727	rs2979234|rs17149723|rs11774836
GAA	-0.021	1.2×10^{-5}	0.919	rs2979234|rs17149723|rs11774836

The TAG and GGA haplotype show significant weak correlation with DEFB CN. TAG and GGA haplotypes are opposite each other, which is what we expect to see. According to the b-values, presence of the TAG haplotype is associated with an increase in CN and presence of the GGA haplotype is associated with a decrease in CN.

The SNP with the strongest evidence of association; SNP rs2979234 was then uploaded onto the LocusZoom software and a plot was generated to show the SNP rs2979234 position and any associations it may have with neighbouring genes and SNPs in a flanking region of +/-400 kbs.

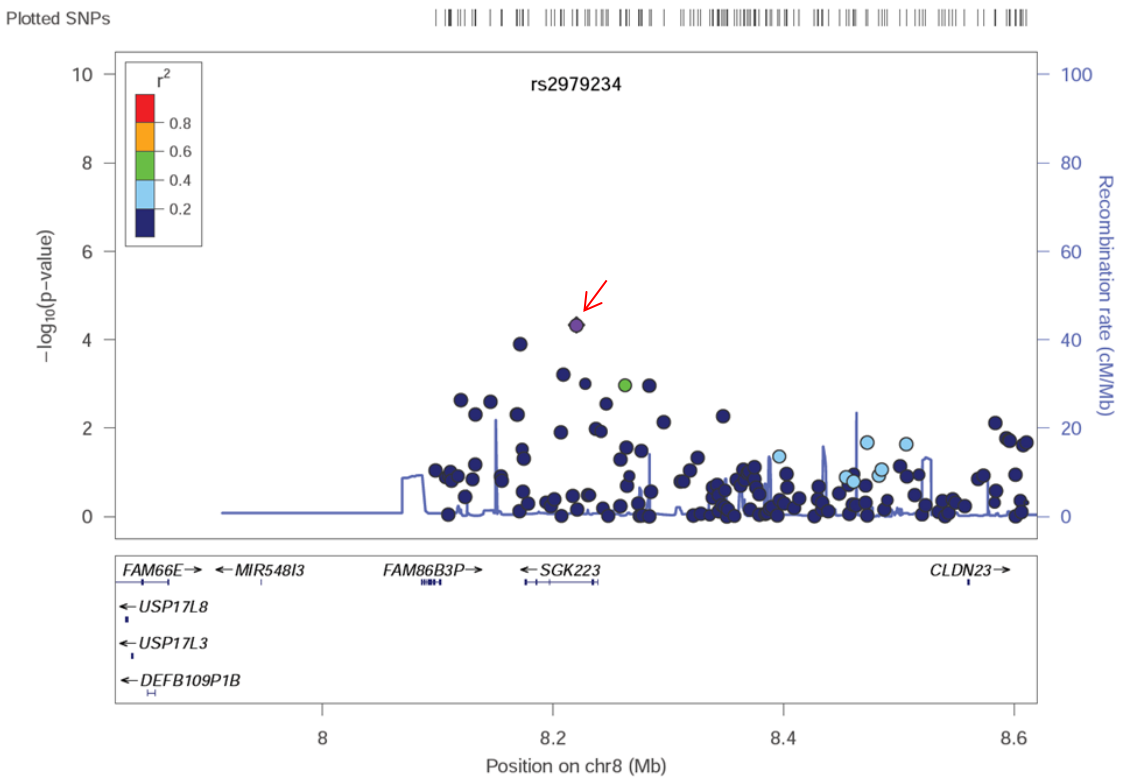


Figure 25: Pairwise results of LD SNPs at 8p23.1. The x-axis represents chromosomal position on chromosome 8 with the location of genes in a flanking region of +/-400 kbs. The left y-axis shows the p-value for association tests at each locus (dot) on the log scale. The right y-axis provides recombination rates in centimorgans per megabase in the chromosomal region identifying recombination hotspots in the region (blue line). The diamond shaped dot with a red arrow represents the SNP in question. Other SNPs in the region are represented by circles. The colours indicate linkage disequilibrium per the r^2 map on top left.

Linkage disequilibrium associated with the SNP rs2979234 appears to fall just outside the *SGK223* gene, with an r^2 value between 0.4-0.6. All the rest show very weak LD with r^2 of 0.4 and below. In other words, there is no association with flanking SNPs, therefore any disease associated with CNV will be with the CNV itself not SNPs, and SNP GWAS will not detect a signal here.

3.4 Discussion

The NimbleGen aCGH data analysis of the 40 samples showed no heterogeneity within the DEFB region. The only exception is a presence/absence polymorphism of an LTR element in the intron of *DEFB107* confirming the results of Turner et al., (2001) revealing an insertional polymorphism of an endogenous retrovirus, HERV-k115 to be exact, however contradicting its prevalence of 16%. Our data showed that 12 samples out of the 40 (almost 30%) showed a loss in *DEFB107* in one of the regions. A 3-primer assay failed to assertively pinpoint the presence of the HERVk-115 polymorphism; it only confirmed the absence of the insert. Let's take for example a 5 copy individual, if 3 out of the 5 copies had the insertional polymorphism, then the 408bp (no insertion) band will appear due to the 2 other copies. If on the other hand, a 2 copy individual does not have the HERVk-115 insert in either copy, then it will also show the 408bp band. Due to this complexity, no further studies were carried out to further investigate the hERVk-115 insertional polymorphism.

Moreover, expression studies could not be carried out as *DEFB107* is mainly expressed in the epididymis and samples were difficult to find. Our analysis also showed some heterogeneity involving the *FAM66/DEFB109P* genes. The function of the *FAM66* gene is not known and all copies of *DEFB109* in this region are pseudogenes according to Taudien et al., (2014).

The NimbleGen aCGH data was also used to determine the length of the CNV that we measure by the PRT method. The CNV spans the DEFB genes as a block corresponding to the main segmental duplication identified in the region with two small regions showing a more complex duplication structure, but where copy number reflects that of the DEFB region. I deduced that the size of the contiguous DEFB CNV region to be 322Kb with two copies assembled in hg18, agreeing with the findings of Hollox et al., (2003). Gene content of CNV was confirmed by Forni et al., (2015) using exome sequence mapping.

When it comes to the inversion studies, no correlation was found between DEFB CN and inversion status. The negative results could perhaps be best explained by the differences in mutation rates. Slowest, representing the oldest variation are SNPs and

their haplotypes. Followed by the inversion which is recurrent and leaves a faint yet detectable signal of association with the SNP haplotype, and finally there is the CNV, which mutates too quickly to leave signal with a flanking SNP allele and seemingly too quickly to leave a signal with the inversion. It also sheds light on the fact that diseases, like Psoriasis, are associated with DEFB CNV regardless of their orientation and inversion status.

4 Comparison of β-defensin copy number typing methods

4.1 Introduction

It is of great importance to find accurate techniques to infer exact copy number in genes that exhibit copy number polymorphisms in the human genome. This not only provides an understanding of the biological basis of these important variations, but also sheds light on whether specific copy number variations truly influence gene dosage, phenotypic variation and susceptibility to disease (Redon et al., 2006). The DEFB gene cluster on chromosome 8p happens to fall in this category.

The DEFB gene cluster on chromosome 8p23.1 may influence susceptibility to certain diseases (Jones et al., 2014; Abe et al., 2013; Jaradat et al., 2013; Hardwick et al., 2012; Taudien et al., 2012; Zhou et al., 2012; Janssens et al., 2010; Hollox et al., 2008), and to date, a gold standard method for DEFB gene copy number quantification has not been established (Nuytten et al., 2009; Groth et al., 2008; Chen et al., 2006; Linzmeier & Ganz, 2005; Hollox et al., 2003). Hollox et al. (2003) reported concordant gene copy numbers of *DEFB4*, *DEFB103*, and *DEFB104* by using MAPH. Groth et al. in 2008, also using MLPA confirmed this concordance, whereas qPCR analysis carried out by Chen et al.(2006) detected discordance in intra-individual DEFB CN. Their approach involved amplification of three target loci (*DEFB4* or *DEFB103* or *DEFB104*) and the single-copy reference locus (human serum albumin, *ALB*) in a single PCR reaction for each sample. The results of DEFB CNV typing of these 3 genes showed intra-individual differences. In *DEFB4*, the 3-, 4- and 5-copy number variants were the main genotypes (20.45%, 52.27% and 20.45%, respectively); the other copy variants were less frequent. For *DEFB103*, the 3-copy number variant was the most frequent genotype (65.91%), with 2-copy and 4-copy variants occurring with frequencies of 20.45% and 13.64%, respectively. In contrast, in *DEFB104*, the 2-copy variant was the dominant genotype (70.45%). There were 29.55% individuals showing 3 copy numbers of *DEFB104*. According to Chen and colleagues, there was no overlap between the groups of different copy numbers for all 3 defensin genes. It is due to these inconsistencies that conflicting disease associations within case-control studies emerge. To be more precise

as explained by Hollox (2010); In an association study, difference in copy number distribution between cases and controls is interpreted as evidence of an influence of that particular polymorphism on disease. However, there may be other causes for that difference: population stratification or variance in the physico-chemical properties of the DNA between cases and controls.

4.2 Study rationale

It is known that case-control association studies are particularly vulnerable to inaccuracies in the raw data, particularly systematic bias between case and control DNA which results in a false positive evidence for association (Hollox, 2010; Clayton et al., 2005). Therefore, accurate typing of copy number variation is a critical step in understanding the relevance of CNVs to human disease. This was demonstrated in the case of Crohn's disease; Fellermann et al. in (2006) using qPCR, reported an association between low DEFB CN and Crohn's disease of the colon from relatively small sample size cohorts (71 colonic Crohn's patients and 169 controls). In contrast, Bentley et al. (2010) reported an association between Crohn's disease and higher DEFB CN, using qPCR measurement in 466 cases and 329 controls. Shortly afterwards in the same year, Aldhous and co-workers used PRT to assess DEFB CN in more than 1500 UK DNA samples including more than 1000 cases of Crohn's disease. A subset of these (625 samples) was typed using both PRT-based methods and standard qPCR methods. Comparing the PRT-based results with Bentley et al. (2010) and Fellermann et al. (2006), found no evidence to support the reported association of Crohn's disease with either low or high DEFB CN (Aldhous et al., 2010).

It is essential to understand that the underlying biological assumption is that DEFB CN varies discretely as complete integers (2, 3, 4 etc.), hence the raw data produced by the various techniques is expected to reflect that underlying biological reality by the clustering of unrounded data about the integer values. The more gathered and less spread out the cluster is around an integer number, the more precise the typing method is.

In this chapter, data from the various methods used to quantify DEFB CNV has been compared for concordance within different cohorts. Data generated using PRT and ddPCR were done in the lab, whereas data from aCGH platforms, nCounter and Genome STRiP were provided to us through our collaborators or were publically available online.

4.3 PRT intra-method consistency

The method adopted by our lab for DEFB CN typing is PRT (2.6). PRT is particularly suited for analysis of complex regions of the genome such as SDs which is what surrounds the DEFB region. One particular strength of PRT is that the first pass assignment of integer copy number is quite high at 93% for the DEFB locus (Armour et al., 2007). This is mainly useful for association studies as repeat testing of 10% of samples is within acceptable limits with respect to cost and labour. PRT has been shown to be superior to qPCR in terms of determining integer CN and accuracy in CN measurement >4 (Fode et al., 2011; Aldhous et al., 2010). Also, Cukier et al. (2009) implies in their research that qPCR generates a significant increase in false-positive CNV results due to sample degradation under standard laboratory storage conditions and emphasize the need to assess sample integrity immediately prior to real-time qPCR experiments.

Armour and co-workers (2007) showed that PRT accuracy is comparable to other well validated methods such as MLPA, which is relatively expensive, time-consuming and requires larger amounts of sample DNA (Zhang et al., 2014). PRT workflow procedure is relatively rapid, genotyping a 96 well-plate with PRT would normally take 8 hours as it does not require the lengthy overnight hybridization steps that are components of both MLPA and MAPH (Armour et al., 2007). With respect to cost, it is a relatively an inexpensive means to determine CN in large scale association studies as the electrophoresis step contributes most to the cost of the test per sample, combining the PCR products in a single capillary which is done in PRT allows multiple measures to enhance the accuracy of the test without proportionate increases in cost (Walker, Janyakhantikul & Armour, 2009). In addition to all this, PRT requires low quantities of

genomic DNA (5–10 ng) (Zhang et al., 2014). According to Zhang et al. (2014), MLPA is superior to qPCR and PRT for DEFB CN determination If accuracy has the highest priority. In this subsection, I would like to assess the quality of the copy number calls from PRT by comparing the calls of the differently labelled primers used across the total plates typed.

In essence, two PRT assays (PRT107A and HSPD21) were designed to specifically co-amplify DNA sequences that occur within the DEFB CNV repeat unit and reference sequence that is not copy number variable (having a diploid copy number of 2). The amplicons slightly differ in size and hence produce separate peaks when run on a capillary electrophoresis as shown in Figure 26 below. The third assay (rs5889219) examines a triallelic indel polymorphism (123/126/128 bp) that can have different sizes in individual DEFB repeats.

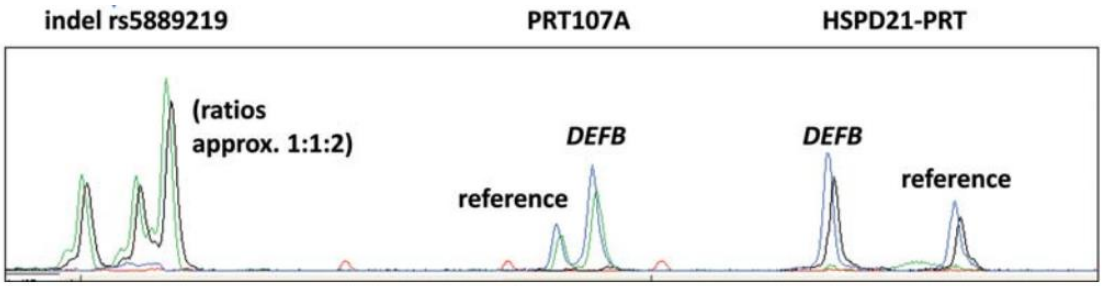


Figure 26: Trace for a sample with a DEFB CN of 4. reflected in the ratios between the 'test' (labelled DEFB) and 'reference' peaks in the two PRT systems HSPD21 and PRT107A, and in the approximate 1:1:2 ratios of alleles (consistent with a total of four copies) for the multi-allelic indel rs5889219. Each analysis is performed in two parallel but independent replicate amplifications using different fluorescent labels which are combined before capillary electrophoresis. Reproduced from (Aldhous et al., 2010)

A total of 99 96-well plates across different cohorts were typed for DEFB CN using PRT throughout the course of this PhD. Each PCR was carried out twice using different fluorescently labelled primers (NED- or HEX-labelled PRT107AF, FAM- or HEX-labelled HSPD21R, HEX- or FAM-labelled 5DEL). In total; four master mix reactions were prepared for a given DNA plate, with the last column on each plate reserved for six

gold-standard control samples (c0088 (cn=4), c0207 (cn=5), c0849 (cn=6), c0913 (cn=3), c0940 (cn=4), c0969 (cn=5)).

After running the capillary electrophoresis, each sample will have 6 ratios that need to be combined together and normalised in order to infer a DEFB integer copy number. For the gold standards, the test to reference peak value for each assay was calculated and plotted against their known copy numbers resulting with a linear regression line that is later used to normalise raw peak area ratios of samples with unknown DEFB CN.

Table 8 below shows the values of the mean peak ratios of test to reference for each PRT assay and the standard deviation for each standard across the 99 plates genotyped for DEFB CN. The values represent the raw un-normalised ratio of reference to test peak areas. This does not reflect the absolute DEFB CN on its own.

Table 8: Table of the mean values of test to reference peak ratios and (standard deviation) for the standards across all the plates typed, categorised by the labelled primer used.

	NED-107A assay	HEX-107A assay	FAM-HSPD21 assay	HEX-HSPD21 assay
Co913 DEFB CN=3	2.81 (0.26)	2.83 (0.24)	1.36 (0.09)	1.33 (0.16)
Co940 DEFB CN=4	3.23 (0.23)	3.26 (0.25)	1.81 (0.11)	1.75 (0.21)
Co088 DEFB CN=4	3.25 (0.23)	3.21 (0.25)	1.74 (0.11)	1.67 (0.18)
Co207 DEFB CN=5	3.75 (0.36)	3.77 (0.36)	2.24 (0.13)	2.14 (0.24)
co969 DEFB CN=5	3.70 (0.24)	3.70 (0.24)	2.22 (0.12)	2.19 (0.19)
Co849 DEFB CN=6	4.26 (0.34)	4.28 (0.34)	2.68 (0.18)	2.64 (0.25)

In this section, all the raw values of the gold standard controls for each plate were combined and plotted together according to the labelled primer. Figure 27 and 28 below shows the distribution of the raw values of the six gold standards as typed by PRT for the NED-labelled PRT107A Assay, HEX-labelled PRT107A Assay, FAM-labelled HSPD21A Assay and HEX-labelled HSPD21 Assay, and the distribution of the 4 DEFB CN raw values across the assays. It can be inferred from the graphs that each sample for

each assay gives a similar value across the spectrum hence increases the consistency of results and confirm reliability of PRT in inferring DEFB CN.

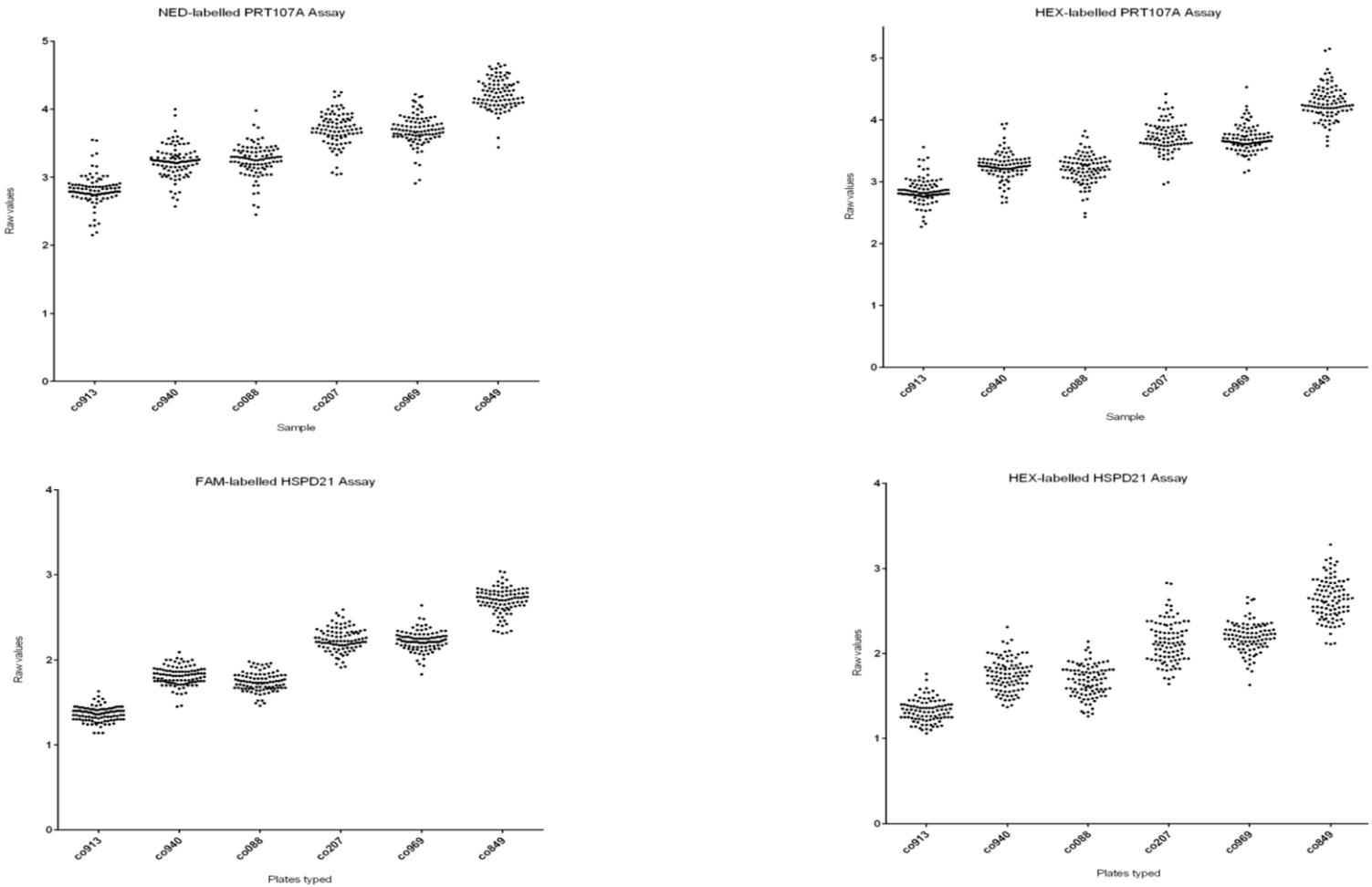


Figure 27: Distribution of the raw values of the six gold standards as typed by PRT for NED-labelled PRT107A assay, HEX-labelled PRT107A assay, FAM-labelled HSPD21A assay and HEX-labelled HSPD21 assay, in ascending order of DEFB CN (c0913 (cn=3), c0940 (cn=4), c0088 (cn=4), c0207 (cn=5), c0969 (cn=5)) and c0849 (cn=6)).

The two samples with a copy number of 4 (co940 and co088) were then plotted on the same graph for visual evaluation of variation across the typed plates for all four primers across the two assays.

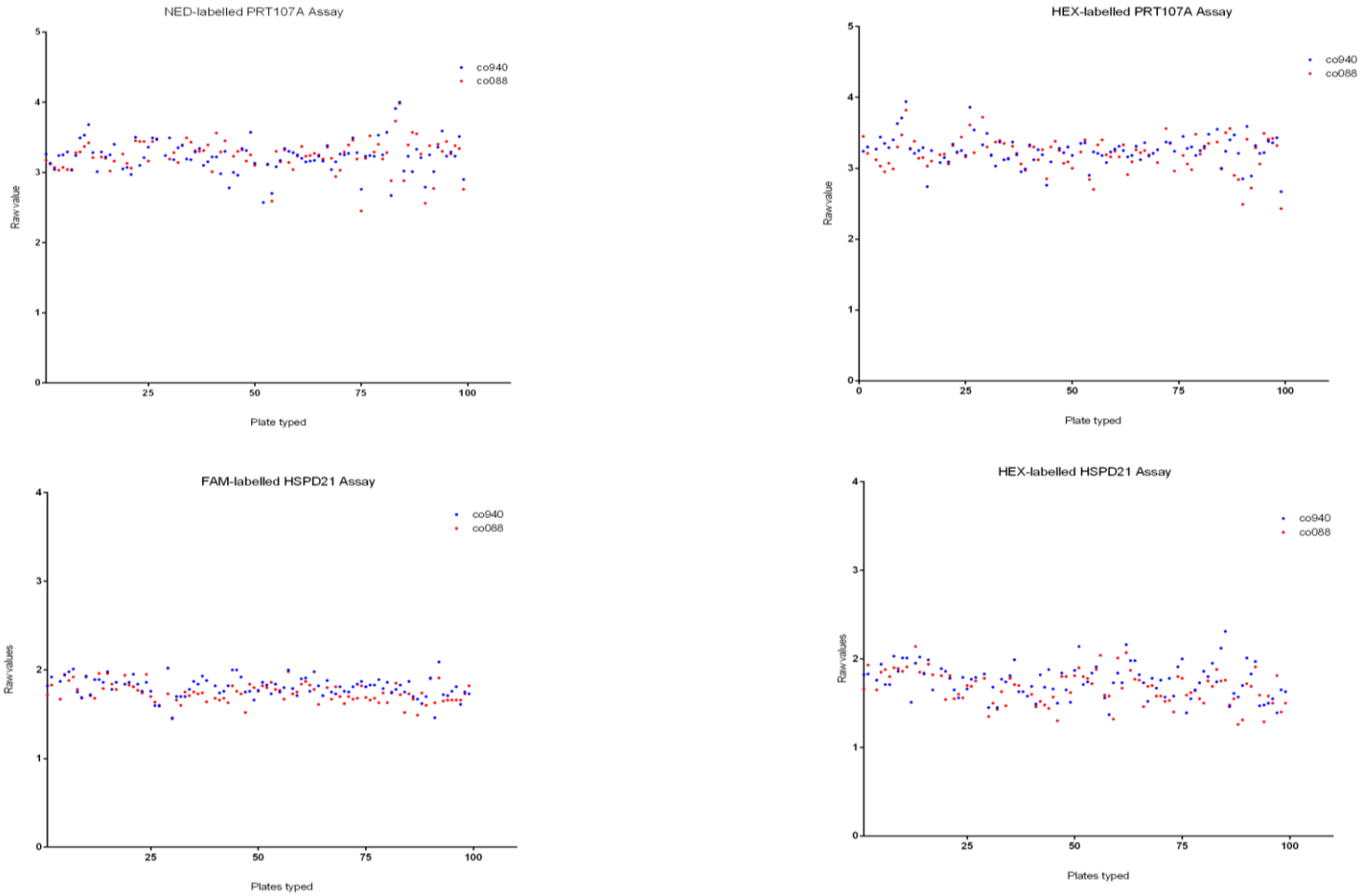


Figure 28: the raw values of NED and HEX-labelled PRT107A assay and FAM and HEX-labelled HSPD21 assay for the samples co940 and co088 (DEFB CN of 4) plotted on the same graph across all typed plates.

The clustering of raw values across the different controls with regards to the two differently labelled primers for each assay is consistent for each DEFB CN called (Table 9). Co913 which has a DEFB CN of 3 shows clustering around 2.8 for the PRT107A assays while it clusters around 1.3 for the HSPD21 assays. Co940 and co088 both have a DEFB CN of 4 and cluster around 3.2 for PRT107A assays and 1.8 for the HSPD21 assays. As for Co207 and co969 which have DEFB CN of 5, raw values cluster around 3.7 for the PRT107A assays and 2.2 for the HSPD21 assays and finally, co849 with a DEFB CN of 6 has raw values clustering around 4.2 for the PRT107A assay and 2.7 for the HSPD21 assays.

Further investigations were done by looking at the clustering of NED-labelled PRT107A and FAM-labelled HSPD21 ratios in Figure 29 and HEX-labelled PRT107A and HEX-labelled HSPD21 ratios in Figure 30 below. As expected, both typing from PRT assays in this system produced clear clusters of results corresponding to copy number diplotypes of 3, 4, 5 and 6.

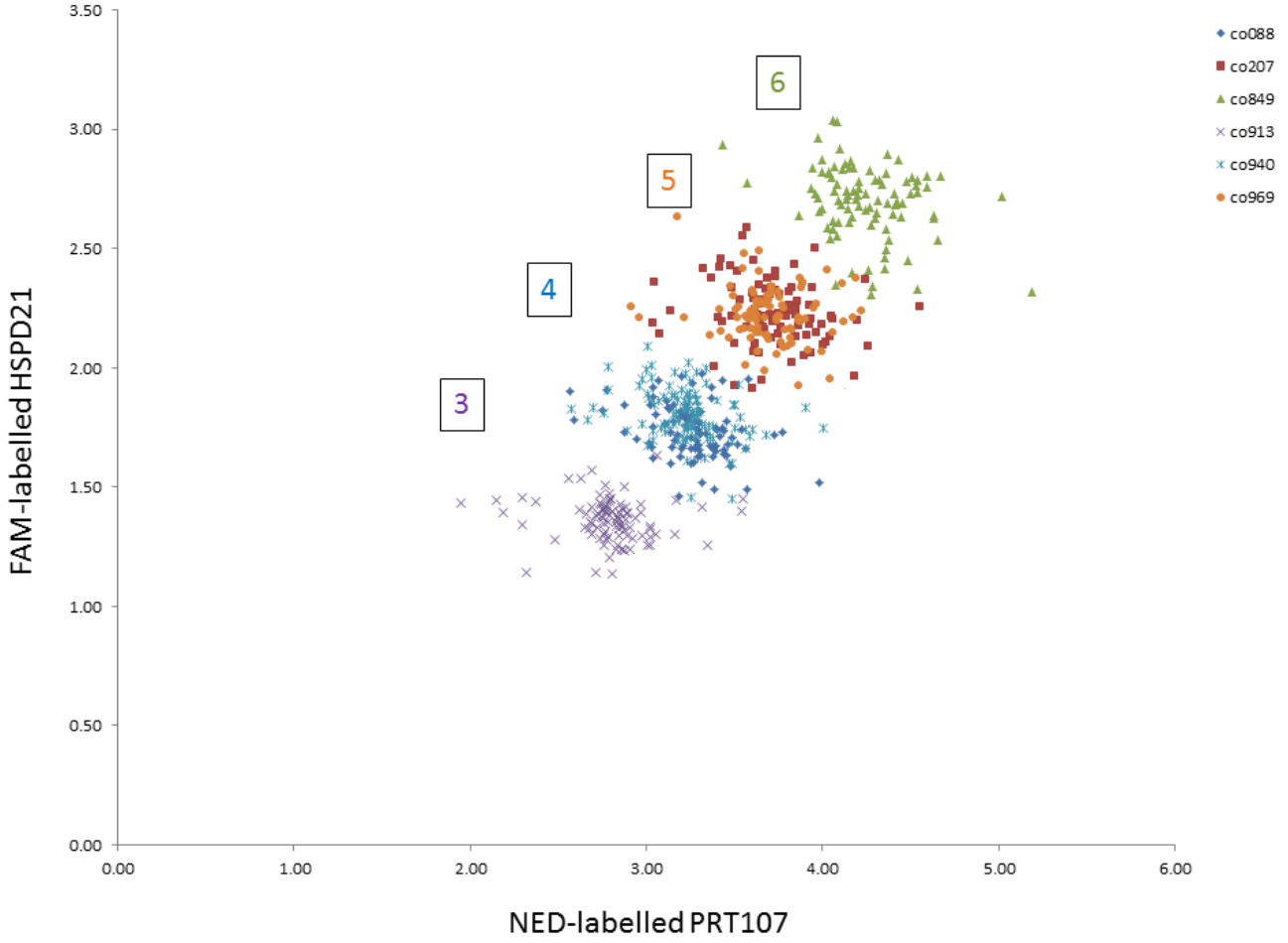


Figure 29: Scatterplot of PRT results between NED-labelled PRT107A and FAM-labelled HSPD21 for standard controls across all plated typed showing clear clustering around the integer copy number called. The DEFB CN was called using the maximum likelihood approach.

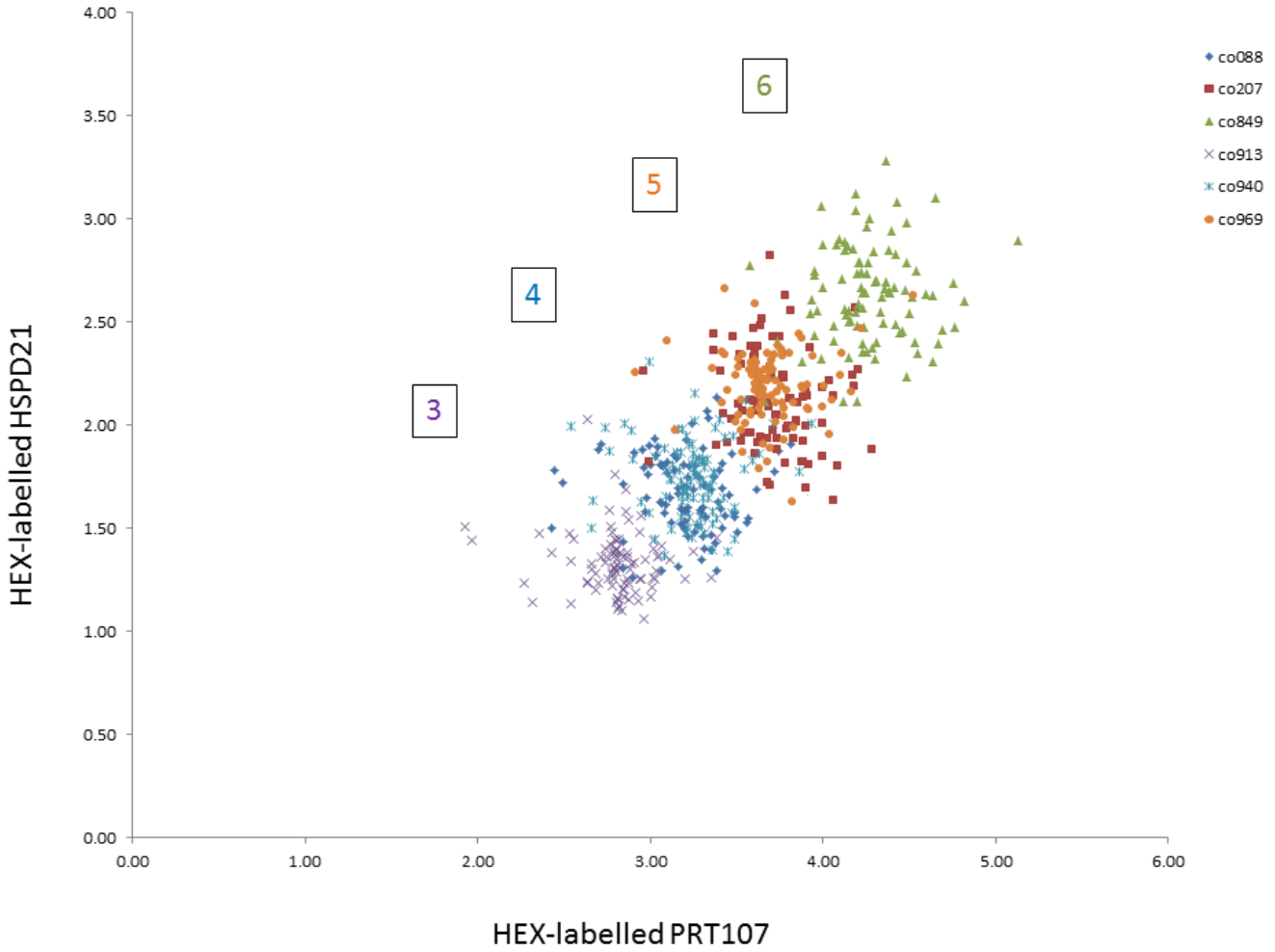


Figure 30: Scatterplot of PRT results between HEX-labelled PRT107A and HEX-labelled HSPD21 for standard controls across all plated typed showing clear clustering around the integer copy number called. The DEFB CN was called using the maximum likelihood approach.

4.4 Comparison of different DEFB CNV typing methods

HapMap and HRC samples that have been typed for DEFB CN using PRT were compared to the same samples typed by NimbleGen aCGH, Agilent aCGH, nCounter, ddPCR and genome STRiP (please see Appendix 2). All calculations and diagrams generated for comparison purposes were done using R statistical computing program. The reason behind choosing PCA to be plotted against the raw PRT CN in all comparisons (except for ddPCR) is because PCA accounts for as much of the variability in the data as possible.

4.4.1 aCGH vs. PRT

4.4.1.1 NimbleGen aCGH vs. PRT

NimbleGen aCGH data was provided by Dr Jackie McArthur and Dr Donna Albertson from the University of California, San Francisco. NimbleGen arrays with 60-mer oligonucleotide probes all across chromosome 8 were used as dictated by the NCBI36/hg18 human genome assembly on UCSC Genome Browser website (http://genome.ucsc.edu/). All test samples were labelled with Cy3 and were referenced to sample AF0105, which was labelled with Cy5.

We had matching data of DEFB CN for 14 HRC samples and 54 HapMap samples. Probes falling between positions chr8:7,125,834 – 7,354,232 (first DEFB region) and chr8:7,706,553 – 7,892,432 (second DEFB region) as dictated by the NCBI36/hg18 human genome assembly for each sample were selected – a total of 15,266 probes. The corrected ratio of Cy5:Cy3 corresponding to each probe was used to calculate the first principal component (PCA).

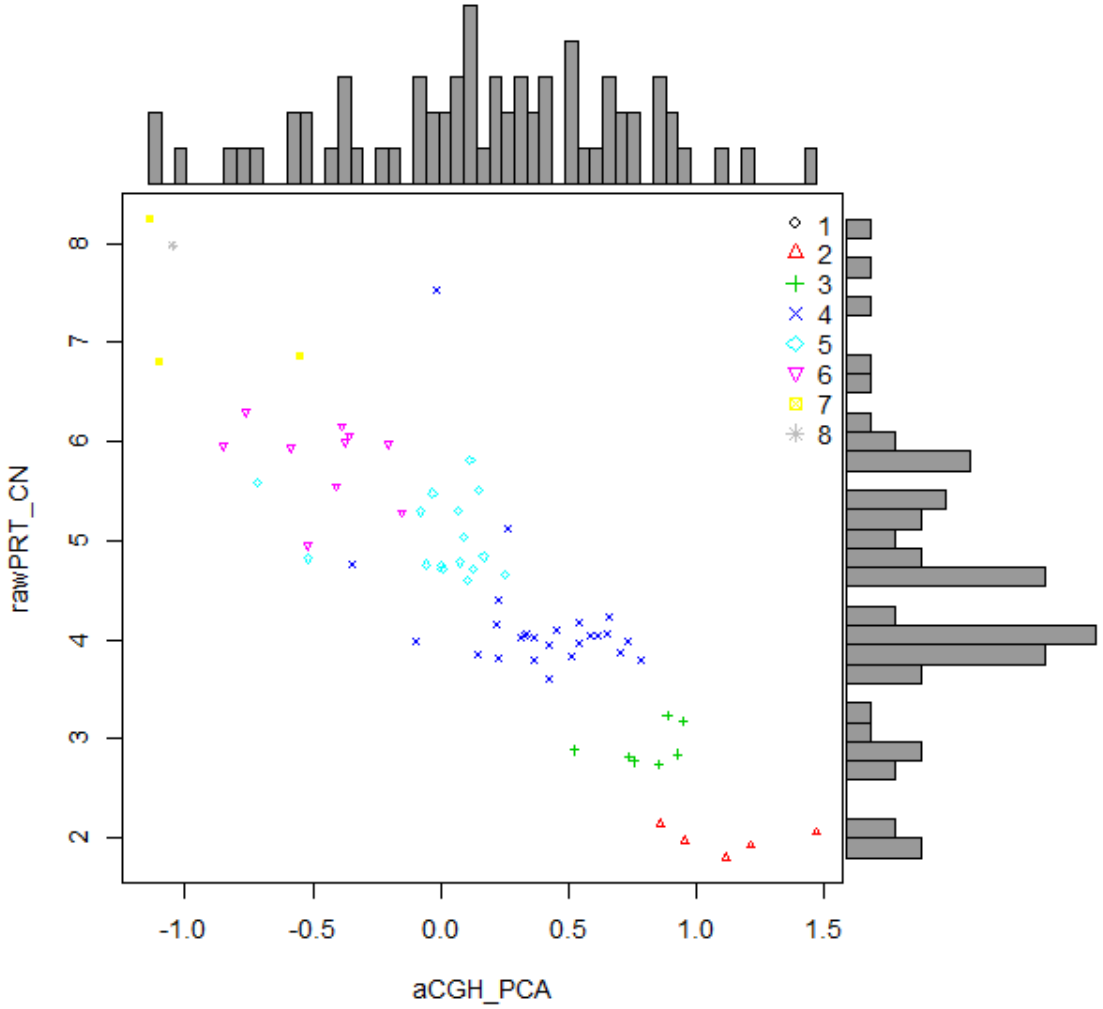


Figure 31: The DEFB CN of 68 samples typed by PRT was compared to same set of samples typed by NimbleGen aCGH data. For NimbleGen aCGH data, the PC1 for 15000 probes across the 68 samples was calculated and plotted (x-axis) against raw normalized PRT CN results (y-axis). Each point on the scatterplot represents a different sample, classified by integer CN as shown in the legend in the upper right corner.

Looking at the scatterplot of the NimbleGen aCGH vs PRT (Figure 31), well distinct clusters cannot be seen easily; however, looking at the histograms, one can see defined clusters for PRT, especially with lower copy number. The noisy NimbleGen aCGH data could be a result of the mismatch of sequence between both DEFB regions due to SNPs. Checking the 'segmental duplication' track on the UCSC genome browser (hg18 assembly), a 99% match is observed i.e. 1 in 100bp mismatch, giving an average

0.6 bp mismatches in 60 (length of probe), if the Poisson distribution is applied, a 43% chance of probes having at least one mismatch per copy is predicted.

4.4.1.2 Agilent aCGH vs. PRT

Agilent aCGH data was publically available from (Conrad, Pinto, et al., 2010). The Agilent 105K CNV genotyping array was designed by the WTCCC in collaboration with Conrad and co-authors (2010). After pilot experiments, each locus was targeted with at least 10 probes. For the DEFB region 60-mer oligonucleotide probes falling between positions chr8:7,121,228 – 7,431,058 as dictated by the NCBI36/hg18 Human Genome Assembly on UCSC Genome Browser website (http://genome.ucsc.edu/) were used. All test samples were referenced to a pool of 10 genomic cell-line DNAs from the ECACC. Matching data for 164 HapMap samples was available. The corrected ratio corresponding to 62 probes were used to calculate the PCA for Agilent aCGH probes.

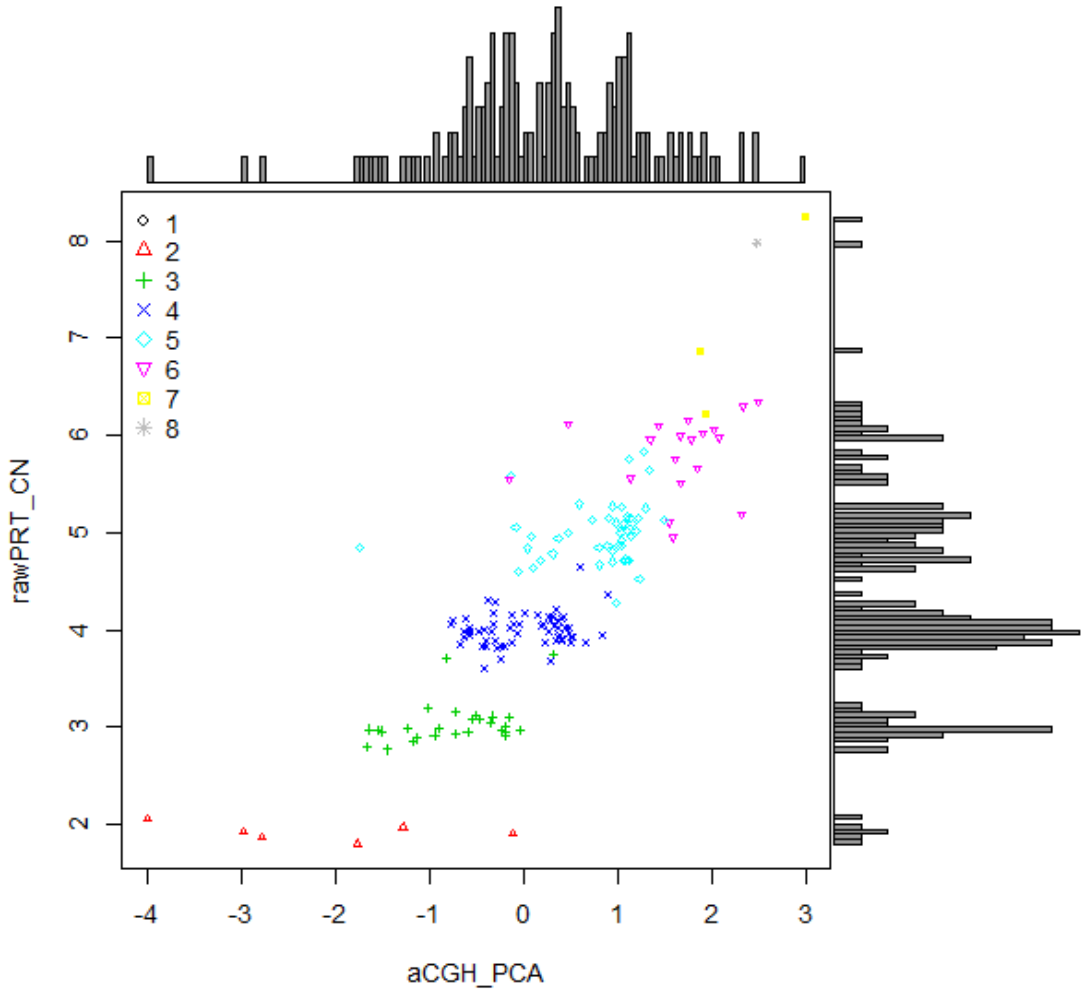


Figure 32: The DEFB CN of 164 samples typed by PRT was compared to same set of samples typed by Agilent aCGH data. For Agilent aCGH data, the PC1 for 62 probes across the 164 samples was calculated and plotted (x-axis) against raw normalized PRT CN results (y-axis). Each point on the scatterplot represents a different sample, classified by integer CN as shown in the legend in the upper left corner.

Samples having a CN of 2 as typed by Agilent aCGH do not cluster at all. This might be due to the limitations of the software in not recognizing the signal as an extra copy but merely background noise. As for the rest of the data, the clusters spread widely in comparison to PRT, even with lower copy number samples, but with clearer cut-offs as the most responsive subset of probes calling the copy number were chosen from the NimbleGen aCGH experimental strategy. Conrad et al (2010) designed the probes according to NimbleGen aCGH standard protocol, with the exception that probes with up to 100 close matches in the genome were included, allowing greater coverage of segmentally duplicated regions. The final design provided 1 probe per 56bp median density across the genome (Conrad, Pinto, et al., 2010).

4.4.2 nCounter vs. PRT

nCounter data was provided by Dr. Andrew Sharp from Mount Sinai School of Medicine, New York. nCounter used 6 probes covering the first DEFB repeat unit as dictated by the NCBI36/hg18 Human Genome Assembly on UCSC Genome Browser website (http://genome.ucsc.edu/). The specific probe positions are summarised in the below table.

Gene	Strand	Position	Maps to
DEFB103A	+	chr8:7274359-7274449	Intron
DEFB104A	+	chr8:7315440-7315529	Exon/Intron
DEFB105A	+	chr8:7333919-7334013	Intron
DEFB106A	+	chr8:7330225-7330317	Intron
DEFB107A	+	chr8:7341702-7341801	Intron
SPAG11	+	chr8:7307325-7307407	Intron

Matching data for 164 HapMap samples was available. The values for the 6 probes used in the nCounter system were used to calculate the PCA.

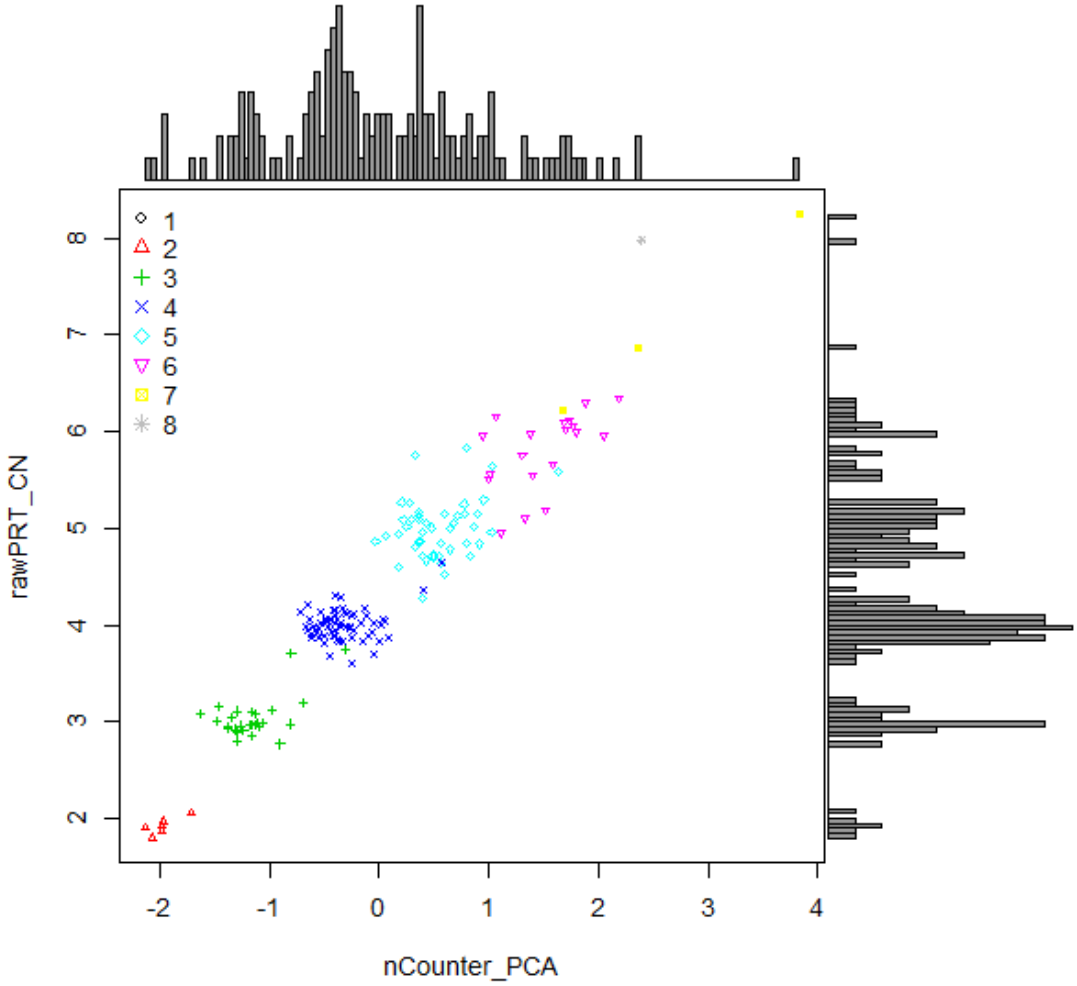


Figure 33: The DEFB CN of 164 samples typed by PRT was compared to same set of samples typed by nCounter data. For nCounter data, the PC1 for 6 probes across the 164 samples was calculated and plotted (x-axis) against raw normalised PRT CN results (y-axis). Each point on the scatterplot represents a different sample, classified by integer CN as shown in the legend in the upper left corner.

As for the nCounter data, again, the data clusters quite nicely for PRT suggesting that the probes reflect the underlying absolute integer CN, and unlike Agilent aCGH, the nCounter data has a greater spread only with higher copy numbers. The diagram suggests that nCounter is a good system for typing DEFB CN.

4.4.3 ddPCR vs. PRT

Upon optimisation, as previously described in (2.7), the gold standard controls were typed for DEFB CN using ddPCR. The DEFB CN calls were accurate and reproducible per user when ran in duplicate as it is demonstrated in Figure 34 and the threshold of quadruplets was tweaked manually (1.2.7)

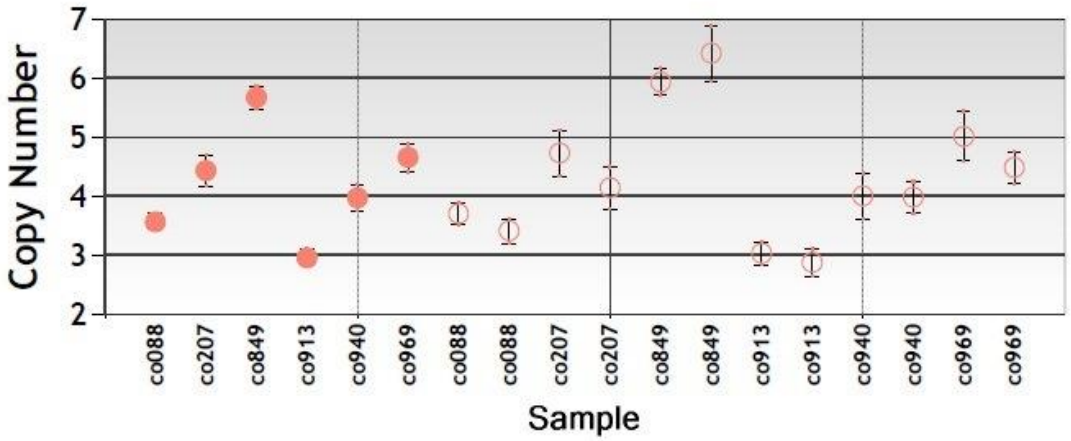


Figure 34: the DEFB CN of the 6 gold controls as typed by the ddPCR optimised protocol. The solid coloured dots show the combined average DEFB CN for each sample. Each sample was run twice and the individual values are shown next to each other.

ddPCR was used to type DEFB CN for 242 samples taken from HRC – 1 and 2 plates and a Portuguese cohort. The results were plotted against CN typed by PRT for the same data set represented in a scatterplot/histogram diagram. For clarification, the ddPCR assay was carried out just once, not in triplicate as recommended, hence the results might not provide true comparability.

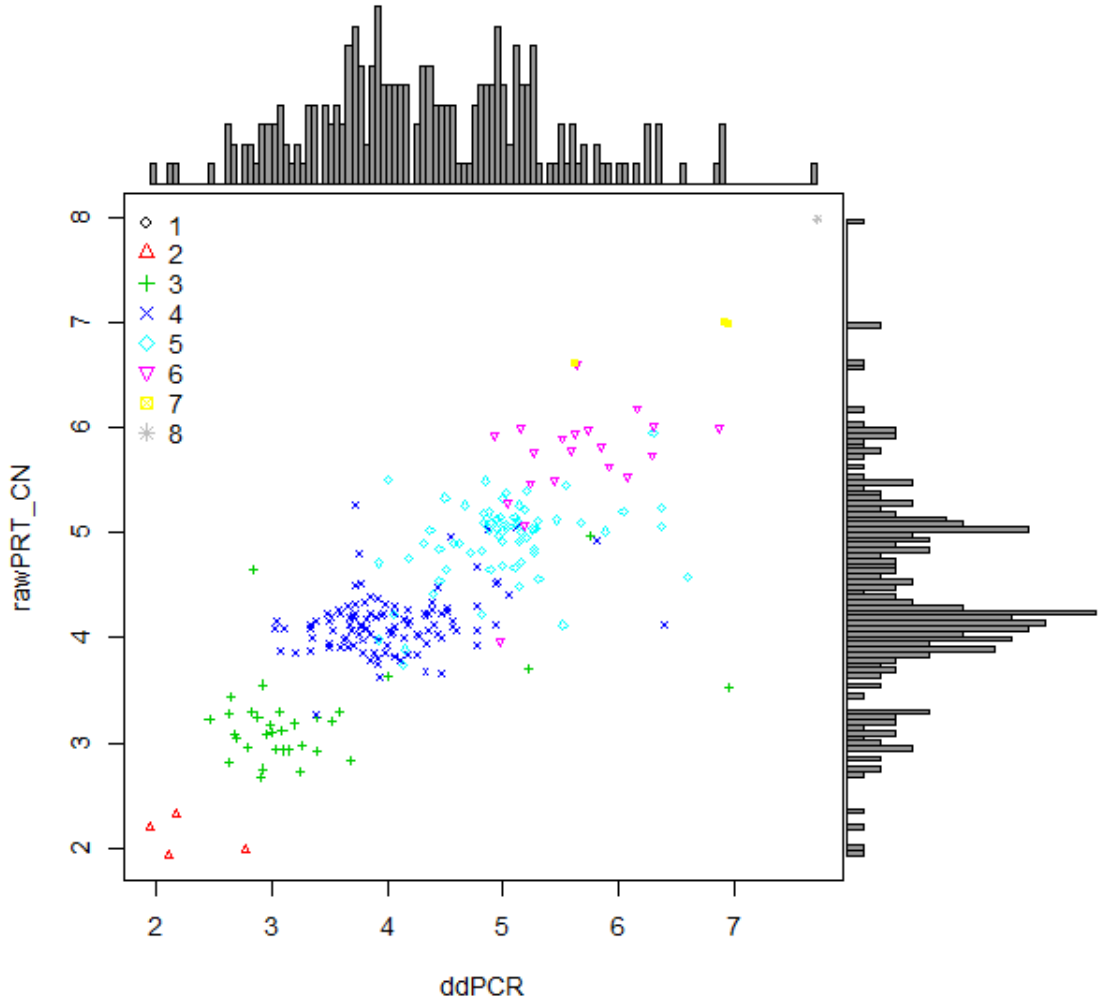


Figure 35: The DEFB CN of 242 samples typed by PRT was compared to same set of samples typed by ddPCR. ddPCR CN on (x-axis) plotted against raw normalised PRT CN results (y-axis). Each point on the scatterplot represents a different sample, classified by integer CN as shown in the legend in the upper left corner.

The samples as typed by ddPCR cluster quite nicely along the integer copy number of DEFB as compared to PRT. The results would have been more accurate if the samples were done in triplicate, unlike this study here where they have been done once only.

4.4.4 Genome STRiP vs. PRT

Genome STRiP s a suite of tools for discovering and genotyping structural variations using next generation sequencing data. The methods are designed to detect shared

variation using data from multiple individuals. The data was publically available from (Handsaker et al., 2015). Matching data for 131 HapMap Samples was available. DEFB copy number calls were compared with copy number estimates made previously by PRT (Figure 36).

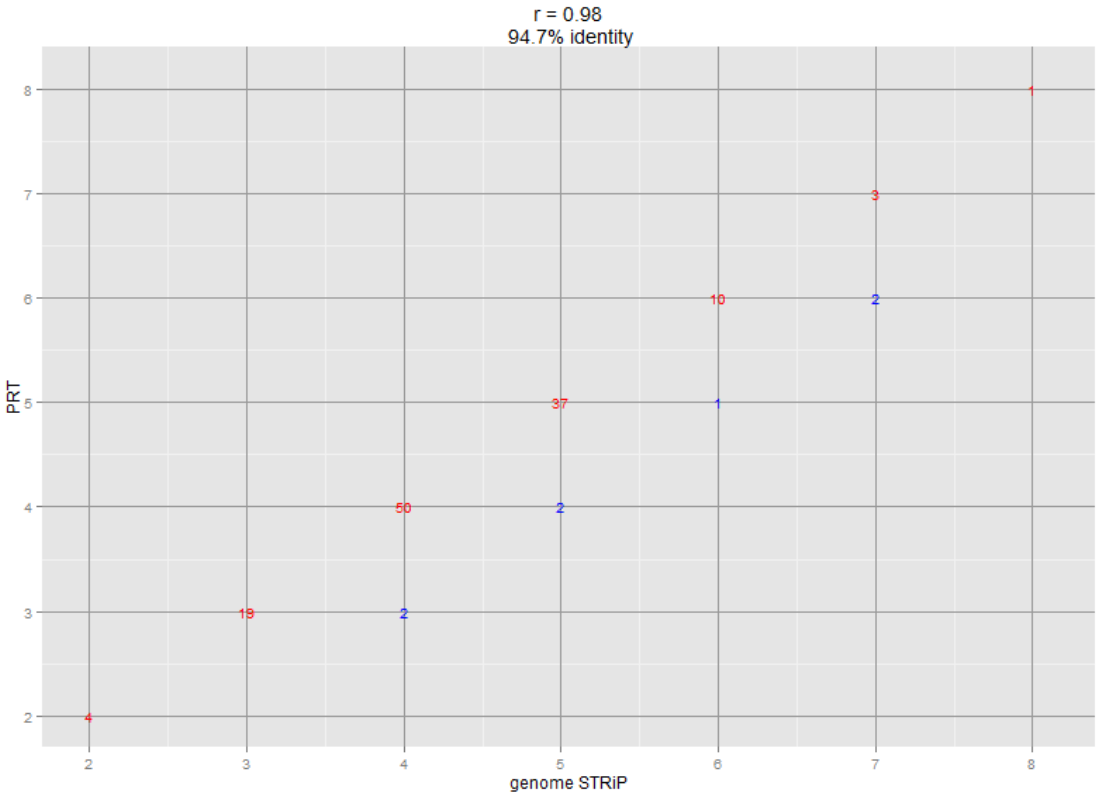


Figure 36: comparison of genome STRiP integer calls and DEFB CN as estimated using ML approach, PRT. The figures in red indicate the numbers of samples concordant for that particular copy number. The numbers in blue indicate the numbers of discordant samples.

It is clear that copy number calls made using genome STRiP agree well with PRT. 124 samples show consensus copy number, giving 5% of samples showing discrepancies.

Genome STRiP combines information from read depth "RD" analysis which detects SVs by analysing the read depth-of-coverage, read pair "RP" analysis which assesses the orientation and spacing of the mapped reads of paired-end sequences, split-read "SR" analysis which evaluates gapped sequence alignments for SV detection and haplotype features of population-scale sequence data for genotyping CNV integer numbers (Mills et al., 2011). Hence, the discrepancies could be due to genome STRiP method

limitations; as SR and RP, according to Cantsilieris et al., (2013) approaches are unreliable in SD regions of the genome (such as the DEFB region), and the RD method is poor at characterizing CNV breakpoints.

4.5 Discussion

Agilent aCGH and NimbleGen aCGH first developed CGH as a method for copy number comparison between differentially labelled target and reference DNAs by measuring the fluorescence ratio along the length of each chromosomal region, indicating relative losses or gains in a target sample using fluorescence in situ hybridization. CGH arrays use arrays of bacterial artificial chromosomes, cDNA, or long synthetic oligonucleotides to probe specific regions of interest for copy number assessment (Li & Olivier, 2013). These probes can either be selected for an even genome-wide distribution or for unique, repetitive regions. NimbleGen aCGH and Agilent aCGH platforms contain substantial amount of probes in SD regions. In general, SDs show higher levels of false positive and false negative call rates in comparison to unique regions of the genome (Pinto et al., 2011). The success for resolving copy number polymorphism (CNP) into discrete copy number classes depends on the genetic property of the CNP, the density and performance of the probe and the parameters used for normalization. It is worth mentioning that probe number does not mean improved coverage or resolution (Cantsilieris et al., 2013). When it comes to DEFB CN calling, NimbleGen aCGH has 15,266 probes across the DEFB region and Agilent aCGH has 62, and as shown in Figure 31 and 32 above, none provide as good a clustering around an integer number call as PRT does, with Agilent aCGH being more precise due to the careful selection of probe positions across the DEFB region unlike NimbleGen aCGH.

A major drawback of using NimbleGen aCGH and Agilent aCGH in calling DEFB CN is that estimation of the copy number call is relative and depends on how well the reference sample is characterised. For example, a loss in the reference sample can be interpreted as a gain in the test sample even though the test sample may have a diploid CN of two (Cantsilieris et al., 2013).

As far as nCounter is concerned, the ability to call absolute DEFN CN is very specific especially among the lower copy number calls. We can conclude that nCounter and PRT are equally good in calling copy number, however when looking at 1 gene at a time, nCounter is not feasible as it is capable of profiling up to 800 regions of the human genome in a single reaction making its running cost high.

The DEFB CN calls as typed by ddPCR were accurate and reproducible when run in duplicate as it is demonstrated in Figure 34 above, after the threshold of quadruplets was tweaked manually (1.2.7). This step is time consuming since each sample has to be looked at individually to set the threshold at the right position according to the user, hence typing large cohorts becomes prohibitively time-consuming. When a copy number is high for example, ddPCR would require an extra step in sample preparation to be done (such as restriction enzymes) to efficiently separate linked copies of the target gene.

As for genome STRiP method, the results when plotted against PRT had a regression coefficient of 0.98 and almost 95% identity. This proved to be a good calling method for DEFB CN however; it requires sequenced genomes to generate the data which is not always possible due to ethical and financial considerations.

In summary, triplex PRT proved to be the best method for typing DEFB CN. PRT uses a small amount of DNA of 5ng/µl, and uses 1 pair of primers to amplify test and reference loci which increases consistency and reproducibility. During this PhD, 7173 samples were genotyped for DEFB CNV using PRT, 443 failed to type from the first run. I.e. PRT has a high pass rate of 94% in returning an integer value, with a significant p-value from the first run. PRT is relatively rapid and suitable for typing thousands of samples in large case-control association studies.

5 β-Defensin copy number and response to pneumolysin toxin in lung epithelial cells

5.1 Introduction

Bronchial epithelial lining fluid (ELF) contains various antimicrobial substances to protect against pathogenic insult. The antimicrobial components of the ELF are lysozyme, lactoferrin, secretory phospholipase-A2, and DEFB (Bals, 2000). hBD-1 is expressed constitutively in the epithelia of the urogenital tract, trachea, and respiratory tract (Medzhitov & Janeway, 2000). hBD-2, which is salt sensitive (Bals et al., 1998) and hBD-3 are expressed mainly in the respiratory tract, and their expression increases in response to infections and inflammatory mediators (Ganz, 2003). In addition, these two hBDs show strong antimicrobial activity against pathogens of respiratory infections, including *P. aeruginosa*, and thus they seem to function in airway mucosal defence (Ganz et al., 2009; Yanagi et al., 2005; Ganz, 2003; Linzmeier et al., 1999). The most common chronic respiratory diseases are chronic obstructive pulmonary disease (COPD) and Asthma. Lung function measures are heritable traits that predict morbidity and mortality in the general population (Young et al., 2007). The ratio of forced expired volume in 1 second (FEV_1) to forced vital capacity (FVC) is used in diagnostic criteria for COPD, whilst the FEV_1 (expressed as % predicted FEV_1) contributes to measures of COPD severity (Wain et al., 2014).

According to the WHO (2015), 64 million people currently have COPD and in 2012, 6% of all deaths globally were attributed to COPD. It is projected that by 2030 COPD will become the third leading cause of death worldwide. Although the major risk factor for COPD is smoking, there is a genetic component (Silverman et al., 1998). Asthma on the other hand, is one of the major non-communicable diseases, with 235 million people currently suffering from asthma. It is a common disease among children and most asthma-related deaths occur in low- and lower-middle income countries (WHO, 2013b).

5.2 Study rationale

Three DEFBs are expressed in significant levels in airway epithelia: hBD-1 (encoded by *DEFB1*) which is expressed in the lung (Goldman et al., 1997), hBD-3 (encoded by *DEFB103*) which is expressed in the trachea (Jia et al., 2001), and hBD-2 (encoded by *DEFB4)*, also expressed in the lung (Bals et al., 1998). hBD-2 expression is up-regulated by mucoid *P.aeruginosa*, and has very effective bactericidal activity against it (Harder et al., 2000). These facts encouraged the first study to test the hypothesis – raised by Hollox et al., in (2003); that variable DEFB CN affects immune system function. The study was carried out on a cohort of 355 patients with cystic fibrosis (CF) (Hollox et al., 2005). The hypothesis was that increased DEFB CN would be associated with improved lung function in patients with CF. The DEFB CN genotyping was done by the MAPH technique. There were no significant correlations between DEFB CN and each respiratory clinical parameter measured; the mean and current FEV_1, and mean and current FVC, suggesting that DEFB CN did not affect lung function in patients with CF.

More recently, GWAS of SNPs identified 26 regions of the genome showing associations with FEV_1 and/or FEV_1/FVC (Artigas et al., 2011; Hancock et al., 2010; Repapi et al., 2010; Wilk et al., 2009). Together, these 26 variants are responsible for only 3.2% of the additive polygenic variance in FEV_1/FVC (Soler Artigas et al., 2011). Of the 26 regions, 8 have been reported to be associated with COPD (Wilk et al., 2012; Castaldi et al., 2011). GWAS for asthma have shown association with at least 10 genomic regions, comprising those encoding proteins involved in the immune response (Wan et al., 2012; Moffatt et al., 2010). These variants explain only around 4% of asthma heritability (Cookson & Moffatt, 2011). Hence, more variants explaining the missing heritability need to be found.

Genome-wide linkage analyses in the Boston Early-Onset COPD study have provided significant evidence for linkage of airway obstruction to chromosome 8 (Silverman et al., 2002) and, in particular, for FEV_1 to the genomic region 8p23 (Palmer et al., 2003). Also, studies carried out by Liao et al., (2012), Andresen et al.(2011) and Levy et al., (2005) reported some evidence for association with *DEFB1* in COPD and asthma.

In light of these findings, Wain et al. in (2014) carried out a study to check association of DEFB CN with lung function as a quantitative trait as well as for association with COPD and asthma using case-control subsets. Details of the study and its replication can be found in Wain et al. (2014). No evidence for association of DEFB CN with lung function in all individuals in either cohort was found, however a moderate signal of association of DEFB CN with COPD was observed within the adult Gedling population (OR=1.4, 95% CI: 1.02–1.92, p=0.039); nonetheless, there was no evidence for this association in the replication cohort (OR=0.89, 95% CI: 0.72–1.07, p=0.217). No observable association was found between DEFB CN and doctor-diagnosed asthma either after removing outliers which gave a preliminary association (OR=1.18, 95% CI: 0.96–1.44, p=0.12) (Wain et al., 2014).

A recent review by Antoni and colleagues (2015) showed that patients with chronic respiratory diseases (COPD, chronic bronchitis and/or asthma) are at a higher risk of Pneumococcal disease, including community-acquired pneumonia (CAP) and invasive pneumococcal disease (IPD) than individuals without these comorbidities. A fold increase of between 1.3 and 13.5 for CAP and 1.3 and 16.8 for IPD has been observed. These findings, in addition to Lee et al. (2004)'s results that hBD-2 is active against a number of respiratory microbes including *Haemophilus influenzae*, *Moraxella catarrhalis*, and *Streptococcus pneumoniae*, confirm that DEFB plays a major role in the healthy and diseased lung (Hiratsuka et al., 2003).

In 2013, Kim and colleagues investigated the molecular mechanism by which hBD-2 expression is induced in response to *S. pneumoniae* in human airway cells. They demonstrated that induction of hBD-2 expression is mediated by pneumolysin (Pn), which is a major virulence protein well-conserved among all clinical *S. pneumoniae* isolates. Kim and colleagues carried out their experiments on human alveolar epithelial A549 cells and demonstrated that that Pn clearly induced hBD-2 expression in dose-dependent and time-dependent manners. The problem with Kim's experimental design is the use of A549 cell lines for the experiment because according to a study carried out by Hellmann et al., (2001), which systematically screened for alterations in the expression of 600 genes in normal human bronchial epithelial (NHBE) cells as well as in several lung carcinoma lines, including A549 cell lines, found that the differential

expression of 17 genes was observed in all four carcinoma lines compared to NHBE cells. It also established that the direction of all 17 gene expression differences, either upregulation or downregulation relative to NHBE cells, was the same for all four carcinoma lines, underscoring their common molecular features. Each lung tumour line also exhibited a number of unique differences compared to both normal cells and the other tumour cell lines. These differences may be due to differences in the cellular origin and/or pathology of the cell lines studied. Hence, the results obtained from carcinoma cell lines do not necessarily have the same effect on normal human bronchial epithelial cells.

In summary, little is known about whether DEFN CN alters how normal lung cells react in response to treatment with Pn. To further understand the role DEFB CN plays in the lungs and respiratory system, we decided to develop a laboratory model to easily enable us to investigate whether DEFB expression levels differ with CN in response to treatment with Pn.

5.3 Experimental design

The experiment was based on two facts; that DEFB expression in humans is related to CN in healthy individuals (Jaradat et al., 2013, 2015; Jansen et al., 2009), and that hBD-2 is strongly induced by Pn in human airway cells as has been demonstrated by Kim et al., (2013). However, Kim's findings were not previously supported by Scharf et al., (2012); as they showed that *S. pneumoniae* induced hBD-2 and hBD-3 release in human bronchial epithelial cells regardless of bacterial viability and the exotoxin Pn.

Dr. Rob Hirst kindly provided us with the required materials for the experiment; ciliated primary human airway cultures in the form of normal human bronchial epithelial (NHBE) cell lines and Pn. NHBE are isolated from epithelial lining of airways above bifurcation of the lungs and consist of the surface epithelial cells and mucus glands. The surface epithelial cells are made up of three principal cell types: basal, goblet, and ciliated cells, of which the latter two form a suprabasal columnar structure and are necessary for mucociliary clearance (ScienCell Research Laboratories, 2016) and have been used as controls in previous publications (Furukawa et al., 2015).

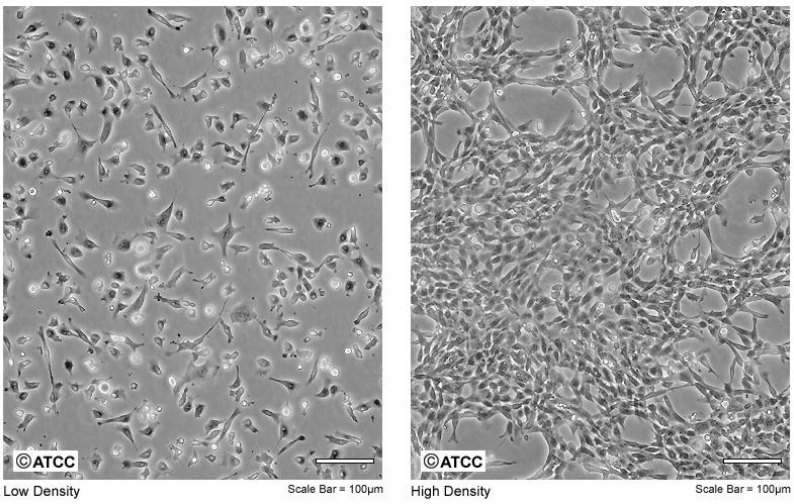


Figure 37: Human Bronchial Epithelial Cells growth at low density (left) and high density (right). Reproduced from (http://www.atcc.org/products/all/CRL-2503.aspx#characteristics). Accessed April 19th 2016.

The experiment layout is summarised in Figure 38 below. The experiment was carried out twice, each time with a different batch of NHBE cells. Each batch was passaged three times.

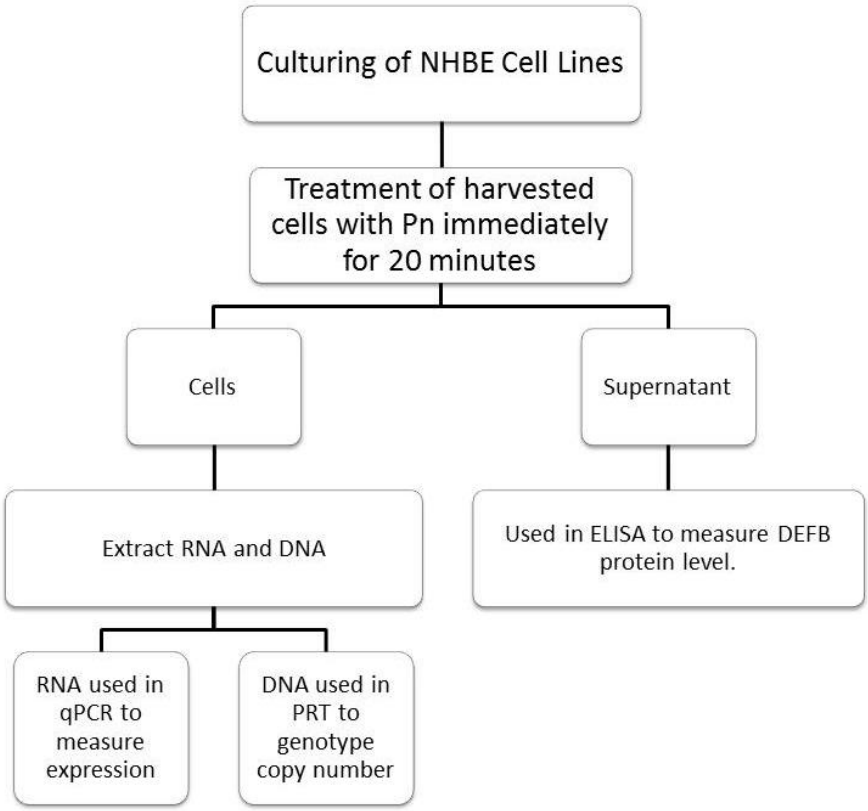


Figure 38: layout of design for the Pn-treated cells experiment.

Dr. Rob Hirst recommended the Pn treatment level to be 42.5 and 170 Hu as these levels were non-toxic to the cells.

5.4 Results

5.4.1 First NHBE cell line

5.4.1.1 DNA samples – genotyping of DEFB4 CN using PRT

After TRIzol extraction, DNA concentration was checked and appropriate amounts of DNA (to be a total of 5ng/µl) were used in PRT. The first NHBE cell line had a DEFB CN of 4.

5.4.1.2 RNA samples – mRNA quantification using qPCR

The RNA was reverse-transcribed to cDNA which was used in qPCR. Before carrying out the qPCR, appropriate controls were to be chosen. An ideal reference for this experiment must have an amplicon size similar to our target gene (*DEFB4*) and must be highly expressed in the lung tissue and diseased lung cell lines (as documented on Expression atlas: http://www.ebi.ac.uk/gxa/home). Table 9 below shows all qPCR control genes offered by Applied Biosystems TaqMan Expression Assays. The differential and baseline expression of each gene as documented by the expression atlas on EMBL-EBI (n.d.) was looked at. The genes that satisfied the above mentioned criteria were Peptidylprolyl Isomerase A (*PPIA)* and Ubiquitin C (*UBC)*. These genes were also deemed as accurate qPCR controls for clinicopathological analysis of lung specimens as published by Nguewa et al., (2008).

Table 9: A comparative table of all qPCR controls offered by Applied Biosystem TaqMan Expression Assays showing the gene, amplicon size, expression level in the lungs and different lung-related cell lines as found on www.ebi.ac.uk/gxa/home

Gene	Amplicon Size (bp)	Lung tissue[5]	Lung tissue[6]	Cell Lines				Remarks
				A549[7]	AG445	IMR-90[8]	NHLF[9]	
18S	61	-	-	-	-	-	-	-
ACTB	171	High	High	Low	Low	High	High	↑ in lung adenocarcinomas
HPRT1	72	Low	Low	High	Low	Low	High	↑ in adenocarcinomas + non-small cell
B2M	81	High	High	Low	Low	Low	High	
GUSB	96	High	High	Low	Low	Low	Low	
HMBS	62	Low	Low	Low	Low	Low	Low	
IPO8	71	Low	High	High	High	Low	High	
PGK1	73	High	High	Low	Low	Low	Low	↑ in adenocarcinomas + non-small cell + adrenocortical ↓poliovirus
POLR2A	61	Low	Low	High	Low	High	Low	
PPIA	97	High	High	High	High	High	High	↓poliovirus
RPLP0	105	Low	Low	Low	Low	High	High	
TBP	65	Low	Low	Low	Low	Low	Low	
TFRC	66	High	Low	Low	Low	Low	Low	↑ in non-small cell ↓tobacco smoke condensate treatment
UBC	71	High	High	High	High	High	High	

The qPCR fluorescence raw values for the samples in this experiment were analysed in the R package 'qpcR (Spiess, 2014) using the Permutation approach'; this involves binding observations together according to ties. The ties bind observations that can be independent measurements from the same sample. In qPCR terms, this would be a

[5]RNA-Seq of human individual tissues and mixture of 16 tissues from Illumina body map
[6]RNA-seq of coding RNA from tissue samples of 95 human individuals representing 27 different tissues in order to determine tissue-specificity of all protein-coding genes
[7]A549 cells are adenocarcinomic human alveolar basal epithelial cells
[8]Human Caucasian fetal lung fibroblast
[9]Normal Human Lung Fibroblasts

real-time PCR for two different genes (gene of interest and reference gene) on the same sample. If ties are omitted, the observations are shuffled independently.

The samples were analysed twice, with each reference gene. The package calculates the ratio between *DEFB4* expression and *PPIA* and *UBC* expression levels consecutively in samples treated with 42.5 Hu and 170 Hu of Pn to controls (0 Hu Pn). The median, CI levels and a p-value are calculated, the p-value gives a measure against the null hypothesis that the expression levels in the initial group (*DEFB4*) occurred by chance.

Table 10: Summary of the results of qPCR analysis for NHBE cells batch 1

Cells from batch	Pn amount (Hu)	Reference gene	Median (CI)	Expression change	p-value
1	42.5	*PPIA*	1.05 (0.45 – 1.65)	No	0.24
1	170		1.01 (0.53 – 2.20)	No	0.21
1	42.5	*UBC*	0.94 (0.50 – 1.20)	No	0.16
1	170		1.07 (0.57 – 1.57)	No	0.17

***: significant p-value (p<0.05)**

The below box and whisker plot provides a visual presentation of the qPCR results.

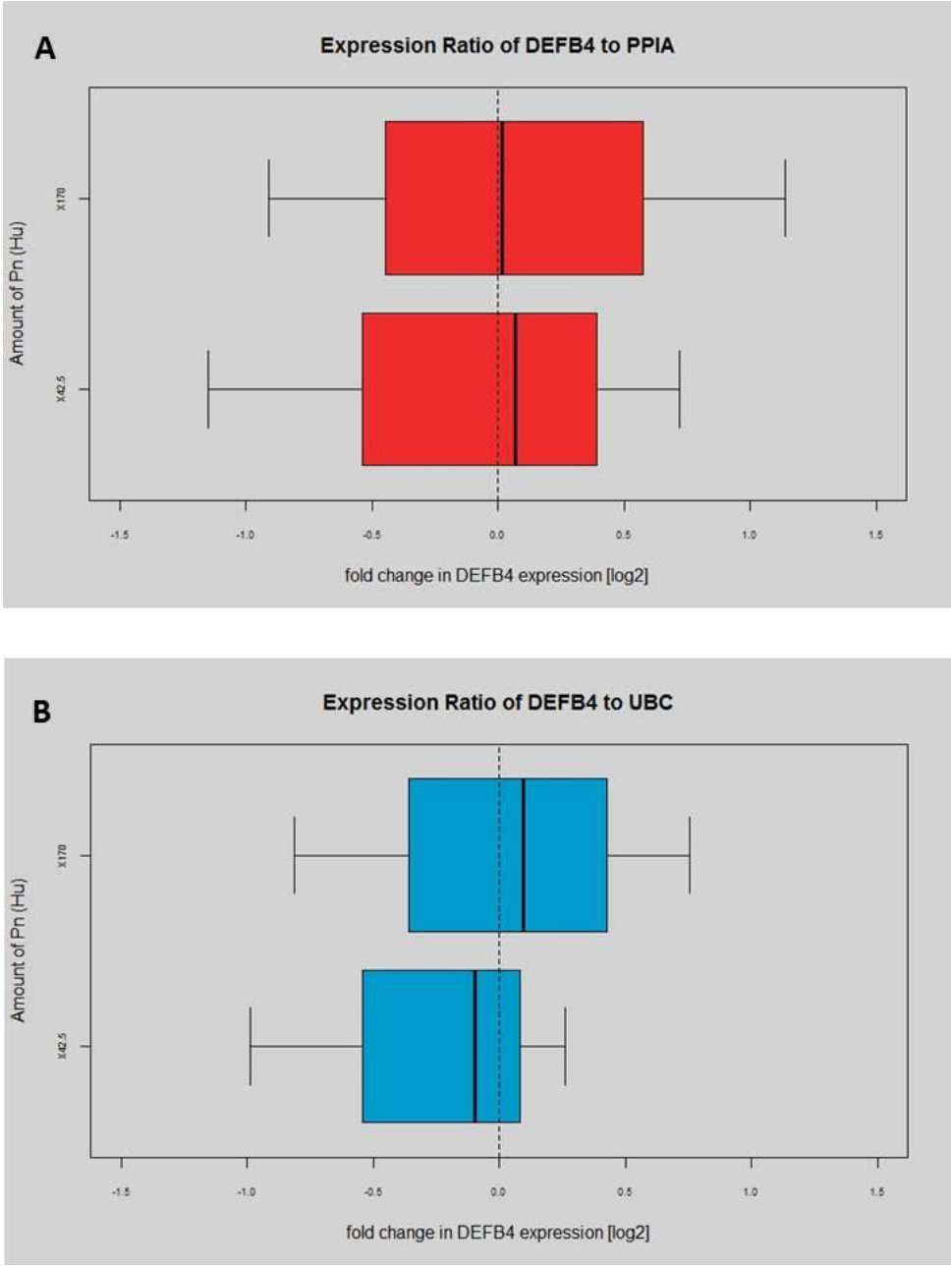


Figure 39: Box and whisker plot for qPCR results. A) Results of expression ratio of *DEFB4* to *PPIA*. B) Results of expression ratio of *DEFB4* to *UBC*. The x-axis shows the $\log_2$ fold change in *DEFB4* expression and the y-axis shows the amount of Pn in Hu used to treat the cells

Increased mRNA expression was defined as N-fold ≥2.0, "normal" expression was an N-fold ranging from 0.5 to 2.0, and decreased mRNA expression was N-fold ≤0.5.

When compared to *PPIA*, the *DEFB4* expression level of cells did not show evidence of any change in expression (Figure 39A). The same was shown when expression levels of *DEFB4* were compared to *UBC* as a reference gene (Figure 39B).

5.4.1.3 Serum for protein quantification using ELISA

The total protein (TP) concentration for the supernatants was measured using Bradford Assay. hBD-2 concentrations were normalised to total protein corrected for ELISA dilution and descriptive statistics for each group were calculated.

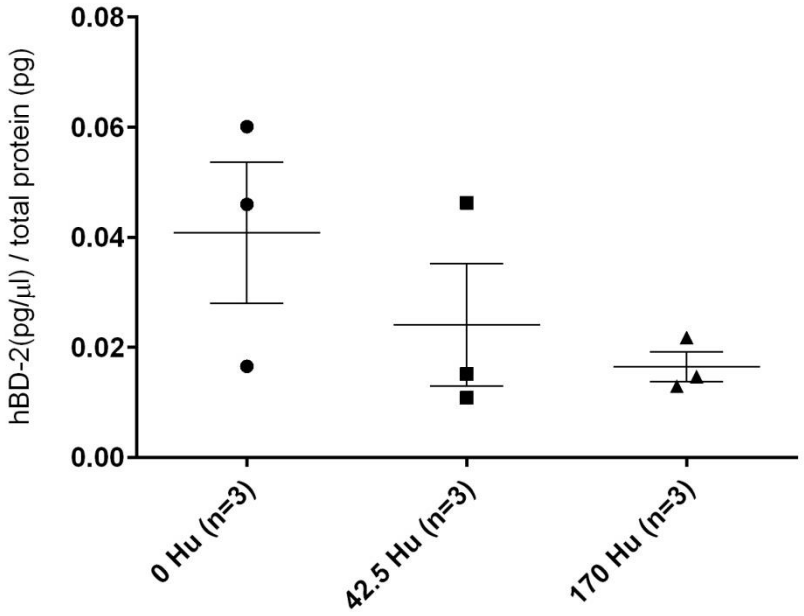


Figure 40: concentration of hBD-2 relative to total protein. The x-axis represents the haemolytic unit (Hu) of Pn treatment, 0 Hu being the control and y-axis is the concentration of hBD-2 normalised to total protein (pg/µl). The data is from 3 different passages.

The mean concentration of hBD-2 of the control group was the highest at 0.041 pg/µl (SEM = 0.013) and as the amount of Pn used for treatment increased; the concentration of hBD-2 relative to total protein decreased. The mean was 0.024 (SEM = 0.011) and 0.016 (SEM = 0.003) for cells treated with 42.5 and 170 Hu respectively.

In a different analysis, hBD-2 concentrations were expressed as hBD-2 release fold when treated with 42.5 and 170 Hu of Pn, compared to protein release concentration in control cells (not treated with Pn). Results are summarised in the table below:

Passage	Amount of Pn treatment (Hu)	TP conc. (pg/µl)	TP conc. corrected for ELISA dilution (pg/µl)	hBD-2 conc (pg/µl)	normalisation of hBD-2 to TP conc (pg/µl)	Normalisation of hBD-2 to control (0 Hu of Pn)	hBD-2 release fold change
1	0	1.59×10^{5}	3.98×10^{4}	1.83×10^{3}	4.60×10^{-2}	0	0
1	42.5	2.06×10^{5}	5.14×10^{4}	2.38×10^{3}	4.63×10^{-2}	2.99×10^{-4}	6.50×10^{-3}
1	170	1.57×10^{5}	3.93×10^{4}	8.56×10^{2}	2.18×10^{-2}	-2.42×10^{-2}	-5.26×10^{-1}
2	0	2.52×10^{5}	6.31×10^{4}	1.05×10^{3}	1.66×10^{-2}	0	0
2	42.5	3.24×10^{5}	8.10×10^{4}	1.23×10^{3}	1.52×10^{-2}	-1.41×10^{-3}	-8.53×10^{-2}
2	170	2.57×10^{5}	6.42×10^{4}	9.43×10^{2}	1.47×10^{-2}	-1.88×10^{-3}	-1.13×10^{-1}
3	0	1.80×10^{5}	4.49×10^{4}	2.70×10^{3}	6.01×10^{-2}	0	0
3	42.5	1.66×10^{5}	4.16×10^{4}	4.51×10^{2}	1.09×10^{-2}	-4.93×10^{-2}	-8.19×10^{-1}
3	170	3.49×10^{5}	8.72×10^{4}	1.13×10^{3}	1.29×10^{-2}	-4.72×10^{-2}	-7.85×10^{-1}

The cells of each passage treated with the same amount of Pn were grouped together and an average and standard deviation was calculated. Results are summarised in the table below and used to generate

Figure 41

Amount of Pn treatment (Hu)	Average hBD-2 release fold change to 0Hu of Pn	Standard deviation	Z-test value – compared to control	p-value
0	0	0		
42.5	-0.30	0.26	2	0.0456
170	-0.47	0.20	4	0.0001

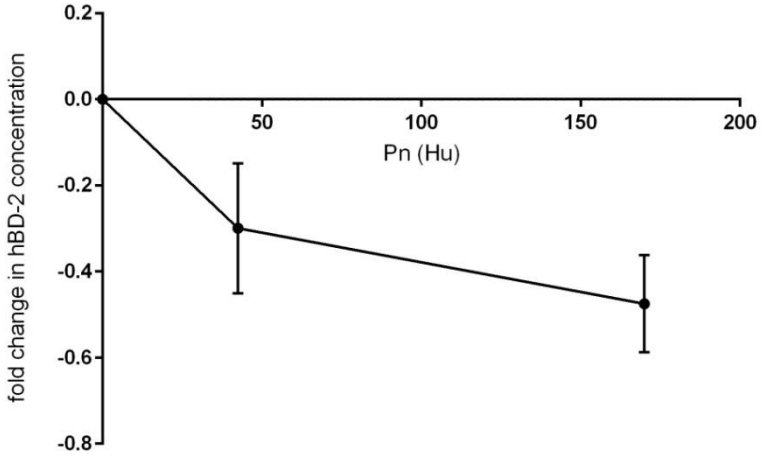


Figure 41: hBD-2 release fold change in cells treated with 42.5 and 170 Hu of Pn consecutively compared to control cells. The data is from 3 different passages.

As it is clear from

Figure 41 above, the amount of hBD-2 released in treated cells compared to control cells decreased by 0.3 and 0.5 fold respectively as the treatment strength increased, disagreeing with the results of Kim et al., (2013).

The experiment was repeated in a different NHBE cell line and results reported below

5.4.2 Replication experiment – second NHBE cell line

5.4.2.1 DNA samples – genotyping of DEFB4 CN using PRT for replication experiment

After TRIzol extraction, DNA concentration was checked and appropriate amounts of DNA (to be a total of 5ng/µl) were used in PRT. The second NHBE cell line had a DEFB CN of 5.

5.4.2.2 RNA samples – mRNA quantification using qPCR for replication experiment

As with the initial experiment, the samples were analysed twice, with each reference gene. The package calculates the ratio between *DEFB4* expression and *PPIA* and *UBC* expression levels consecutively in samples treated with 42.5 Hu and 170 Hu of Pn to controls (0 Hu Pn). The median, CI levels and a p-value are calculated, the p-value gives a measure against the null hypothesis that the expression levels in the initial group (*DEFB4*) occurred by chance.

Table 11: Summary of the results of qPCR analysis for NHBE cells batch 2 – replication experiment.

Cells from batch	Pn amount (Hu)	Reference gene	Median (CI)	Expression change	p-value
2	42.5	*PPIA*	1.09 (0.50 – 2.25)	No	0.23
2	170		0.96 (0.43 – 1.66)	No	0.25
2	42.5	*UBC*	1.03 (0.50 – 2.22)	No	0.22
2	170		0.87 (0.46 – 1.51)	No	0.22

The below box and whisker plot provides a visual presentation of the qPCR results.

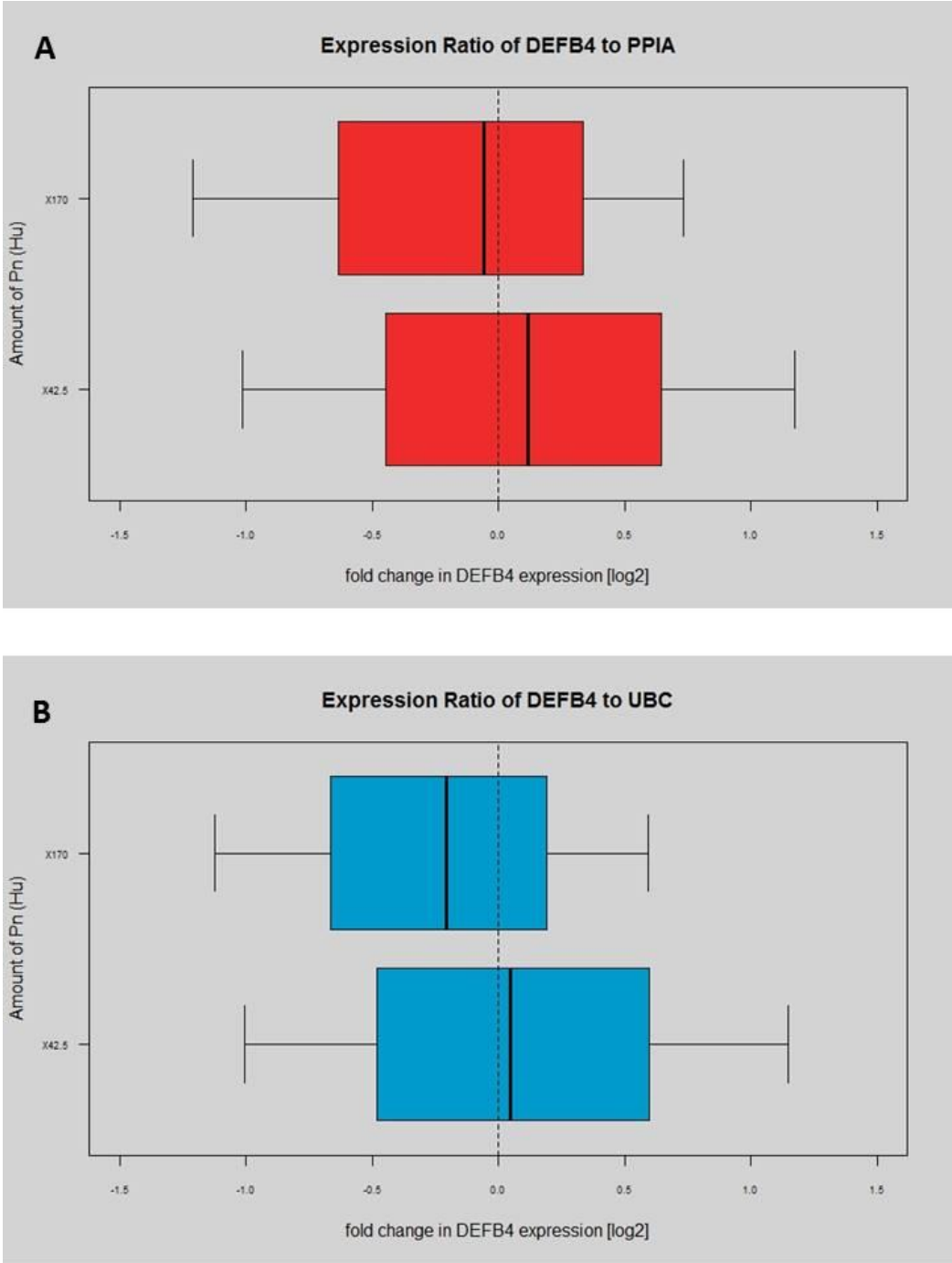


Figure 42: Box and whisker plot for qPCR replication experiment. A) Results of expression ratio of *DEFB4* to *PPIA*. B) Results of expression ratio of *DEFB4* to *UBC*. The x-axis shows the $\log_2$ fold change in *DEFB4* expression and the y-axis shows the amount of Pn in Hu used to treat the cells

When compared to both *PPIA* and *UBC*, the *DEFB4* mRNA expression level of cells did not change.

The results of the replication experiment reproduced the results of the original experiment, confirming that hBD-2 mRNA levels in response to treatment with Pn did not change.

5.4.2.3 Serum for protein quantification using ELISA for replication experiment

The total protein (TP) concentration for the supernatants was measured using Bradford Assay as in the original experiment. hBD-2 concentrations were normalised to total protein corrected for ELISA dilution and descriptive statistics for each group were calculated.

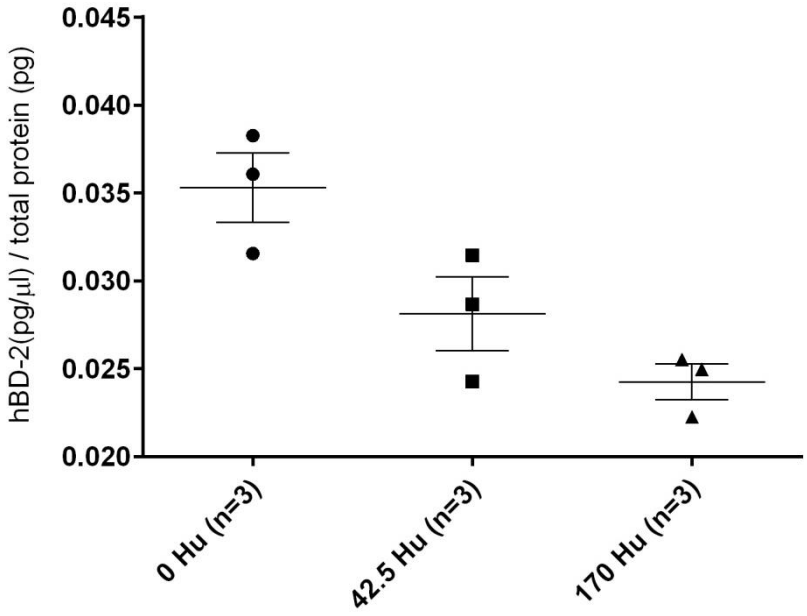


Figure 43: concentration of hBD-2 relative to total protein. The x-axis represents the haemolytic unit (Hu) of Pn treatment, 0 Hu being the control and y-axis is the concentration of hBD-2 normalised to total protein (pg/µl). The data is from 3 different passages.

The mean concentration of hBD-2 of the control group was the highest at 0.036 pg/µl (SEM = 0.002) and as the amount of Pn used for treatment increased; the concentration of hBD-2 relative to total protein decreased. The mean was 0.028 and

0.024 with SEM of (0.002) and (0.001) for cells treated with 42.5 and 170 Hu respectively.

In a different analysis, hBD-2 concentrations were expressed as hBD-2 release fold when treated with 42.5 and 170 Hu of Pn, compared to protein release concentration in control cells (not treated with Pn). Results are summarised in the table below:

Passage	Amount of Pn treatment (Hu)	TP conc. (pg/ul)	TP conc. corrected for ELISA dilution (pg/ul)	hBD-2 conc (pg/ul)	normalisation of hBD-2 to TP conc (pg/ul)	Normalisation of hBD-2 to control (0 Hu of Pn)	hBD-2 release fold change
1	0	1.83×10^{5}	4.59×10^{4}	1.75×10^{3}	3.83×10^{-2}	0	0
1	42.5	2.80×10^{5}	7.00×10^{4}	1.70×10^{3}	2.43×10^{-2}	-1.40×10^{-2}	-3.65×10^{-1}
1	170	2.74×10^{5}	6.84×10^{4}	1.71×10^{3}	2.50 x10 -2	-1.33×10^{-2}	-3.47×10^{-1}
2	0	2.38×10^{5}	5.94×10^{4}	1.87×10^{3}	3.16×10^{-2}	0	0
2	42.5	2.96×10^{5}	7.40×10^{4}	2.33×10^{3}	3.15×10^{-2}	-8.91×10^{-5}	-2.82×10^{-3}
2	170	3.40×10^{5}	8.51×10^{4}	1.90×10^{3}	2.23×10^{-2}	-9.28×10^{-3}	-2.94×10^{-1}
3	0	1.91×10^{5}	4.78×10^{4}	1.72×10^{3}	3.61×10^{-2}	0	0
3	42.5	2.40×10^{5}	5.99×10^{4}	1.72×10^{3}	2.87×10^{-2}	-7.40×10^{-3}	-2.05×10^{-1}
3	170	2.58×10^{5}	6.46×10^{4}	1.65×10^{3}	2.55×10^{-2}	-1.05×10^{-2}	-2.92×10^{-1}

The cells of each passage treated with the same amount of Pn were grouped together and an average and standard deviation was calculated. Results are summarised in the table below and used to generate Figure 44.

Amount of Pn treatment (Hu)	Average hBD-2 release fold change to 0 Hu of Pn	Standard deviation	Z-test value – compared to control	p-value
0	0	0		
42.5	-0.16	0.12	2.3	0.021
170	-0.25	0.06	7	0.0001

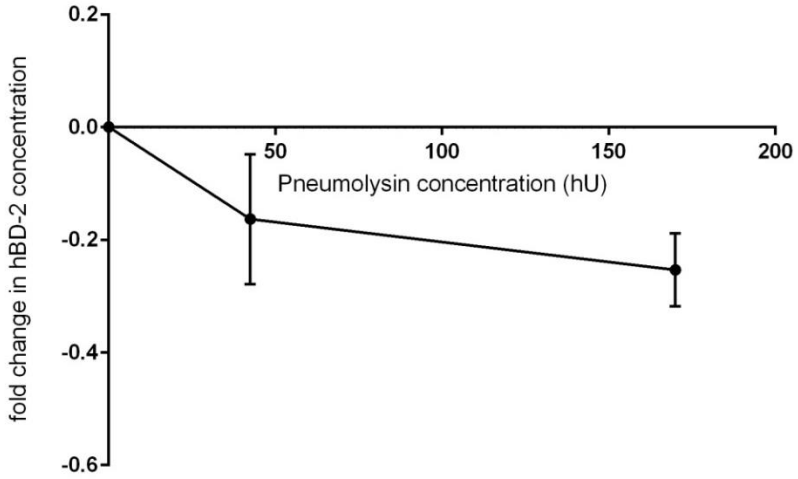


Figure 44: hBD-2 release fold change in replication experiment cells treated with 42.5 and 170 Hu of Pn consecutively compared to control cells. The data is from 3 different passages.

The amount of hBD-2 released in treated cells compared to control cells decreased by 0.2 and 0.3 fold respectively as the treatment strength increased. These results reproduced the results of the original experiment.

5.4.3 Expression comparison between cell lines

The first NHBE cell line had a DEFB CN of 4 whereas the second NHBE cell line used for the replication experiment had a DEFB CN of 5. Comparison analysis to check if DEFB expression and hBD-2 protein levels differ between the two cell lines were carried out and results summarized below

5.4.3.1 mRNA expression comparison

In this subsection, the qPCR crossing point (Cp) results were used. Cp is defined as the cycle at which fluorescence from amplification exceeds the background fluorescence and is standardized by the MIQE guidelines as the quantification cycle (Bustin et al., 2009). A lower Cp correlates with higher target expression in a sample.

The average Cp values for both cell lines in both control (0 Hu of Pn) and experimental conditions (42.5 and 170 Hu of Pn) were used to carry out an unpaired t-test under Welch's correction to compare amount of *DEFB4* mRNA in cells that have a DEFB CN of 4 and cells that have a DEFB CN of 5.

There was a significant difference in the Cp value for cells with a DEFB CN of 4 ($\overline{x}$= 33.72, SEM=0.64) and cells with DEFB CN of 5 ($\overline{x}$=28.87, SEM=0.57) under Welch-corrected t (4) =5.638, and two-tailed p = 0.005. Cell line 1 (DEFB CN =4) had a higher Cp value than cell line 2 (DEFB CN =5) meaning that the amount of *DEFB4* mRNA in cell line 2 was higher in the three experimental conditions (0 Hu of Pn (control), 42.5 and 170 Hu of Pn respectively). These findings suggest that DEFB CN has an effect on the amount of mRNA found in NHBE cell lines.

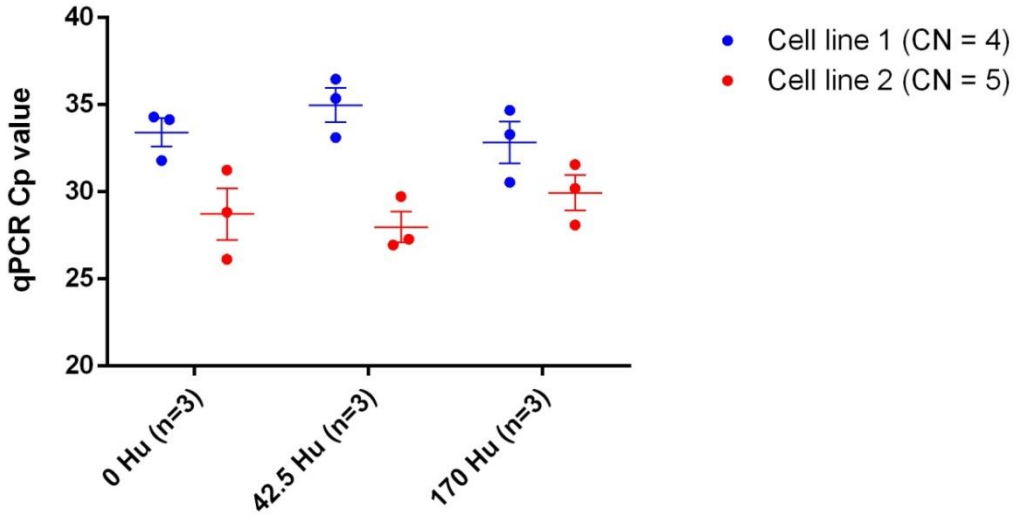


Figure 45: qPCR Cp values for amount of *DEFB4* mRNA in cell line 1 (DEFB CN = 4) and cell line 2 (DEFB CN = 5) in response to treatment with Pn. The data is from 3 different passages.

In order to check if expression fold significantly differs in cells with DEFB CN of 4 and 5 due to treatment with Pn, data from sections 5.4.1.2 and 5.4.2.2 were used. An unpaired t-test was conducted to measure the expression fold change in *DEFB4* relative to the reference *PPIA* and *UBC* genes in cell line 1 to cell line 2.

For using *PPIA* as a reference data, there was no significant difference between *DEFB4* expression change in cells with DEFB CN =4 ($\overline{x}$=0.04, SEM=0.03) and cells with DEFB CN = 5 ($\overline{x}$=0.03, SEM=0.09) upon treatment with Pn. under Welch-corrected; t (1.18) =0.12, and two-tailed p = 0.92

The results showed a similar trend for when *UBC* data was used as a reference. There was no significant difference between *DEFB4* expression change in cells with DEFB CN =4 ($\overline{x}$=7.3 x 10^{-4}, SEM=0.1) and cells with DEFB CN = 5 ($\overline{x}$=-0.08, SEM=0.13) upon treatment with Pn. under Welch-corrected; t (1.86) =0.51, and two-tailed p = 0.66

As far as hBD-2 concentration is concerned, an unpaired t-test was conducted to measure the amount of protein (pg/µl) in cells with DEFB CN of 4 and 5 in control and experimental conditions. There was no significant difference between hBD-2 concentration in cells with DEFB CN =4 ($\overline{x}$=0.027, SEM=0.007) and cells with DEFB CN = 5 ($\overline{x}$=0.029, SEM=0.003), under Welch-corrected t (2.77) =0.26, and two-tailed p = 0.81.

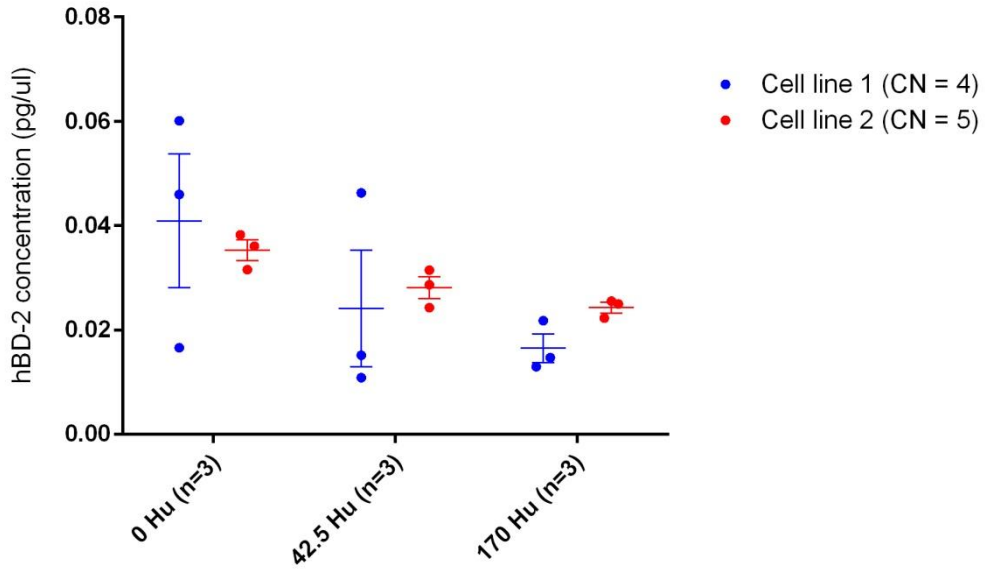


Figure 46: hBD-2 concentration in cell line 1 (DEFB CN = 4) and cell line 2 (DEFB CN = 5) in response to treatment with Pn. The data is from 3 different passages.

5.5 Discussion

The cell line from the original experiment had a DEFB CN = 4 and cells from the replication experiment had a DEFB CN =5. The qPCR results showed no change in *DEFB4* expression fold change relative to reference genes *PPIA* and *UBC* in response to treatment with different amounts of Pn in both cell lines.

When a comparison was conducted to check if a difference of *DEFB4* expression exists between cells carrying different DEFB CN (4 and 5 in this case), cells having a DEFB CN =4 had a significantly lower *DEFB4* mRNA level than cells having a DEFB CN = 5. These results confirm that DEFB CN influence basal expression levels of hBD-2 mRNA in NHBE cells

As far as hBD-2 concentration levels are concerned, treatment with Pn reduced the amount of hBD-2 in the supernatant. These results contradict previous findings of Kim et al., (2013) and Scharf et al., (2012). The contradicting results could be explained by the use of different cell lines in each of the experiments; NHBE cell lines in ours, Human alveolar epithelial A549 cells in Kim's experiment and Primary human small airway epithelial cells in Scharf's experiment. It has been previously confirmed by Hellmann et al., (2001) that A549 cell lines, being carcinoma cell lines had expression differences across 17 genes expression differences, either upregulation or downregulation relative to NHBE cells. Hence, the results obtained from carcinoma cell lines do not necessarily have the same effect on normal human bronchial epithelial cells and cannot mimic *in vivo* studies.

The second factor that could have influenced the results is the different amount of Pn used in each experiment. According to Dr. Rob Hirst's recommendation, we used Pn levels of 42.5 and 170 Hu which are the levels that triggered an initial change in his studies (personal communication, 2015) and that were not high enough to kill off all cells (Hirst et al., 2004). The third factor that might have influenced my results is the amount of time the cells were treated for. Our cells were left in contact with Pn for 20 minutes directly after being harvested, as opposed to 4-6 hours in Kim's experiment and 3-4 hours in Scharf's experiment.

As far as the ELISA results are concerned, hBD-2 concentration levels were not significantly different in cells that have a DEFB CN of 4 and 5 (Figure 46), but hBD-2 concentration levels decreased as levels of treatment with Pn increased in each cell line (

Figure 41 and Figure 44). The explanation behind why mRNA levels showed no change upon treatment with Pn could suggest that hBD-2 is released at sublytic (pro-inflammatory) concentrations of Pn as an early indicator that the epithelial cells are stressed and return to basal levels. Therefore the reduction in hBD-2 levels may be an acute response to treatment with Pn.

Before a solid conclusion about the reaction of DEFB in relation to Pn treatment is deduced based on a model system, the experiment needs to be repeated using NHBE cell lines with different DEFB CN.

6 Association of β-defensin copy number variation in HIV cohorts

6.1 Introduction

Human Immunodeficiency Virus (HIV) infection is prevalent globally with around 36.9 million people estimated to be living with it (WHO, 2015b). Disease prevalence is not consistent throughout the world as shown in Figure 47 below; Sub-Saharan Africa is the most affected region, with 25.8 million people living with HIV in 2014. Also sub-Saharan Africa accounts for almost 70% of the global total of new HIV infections (WHO, 2015b).

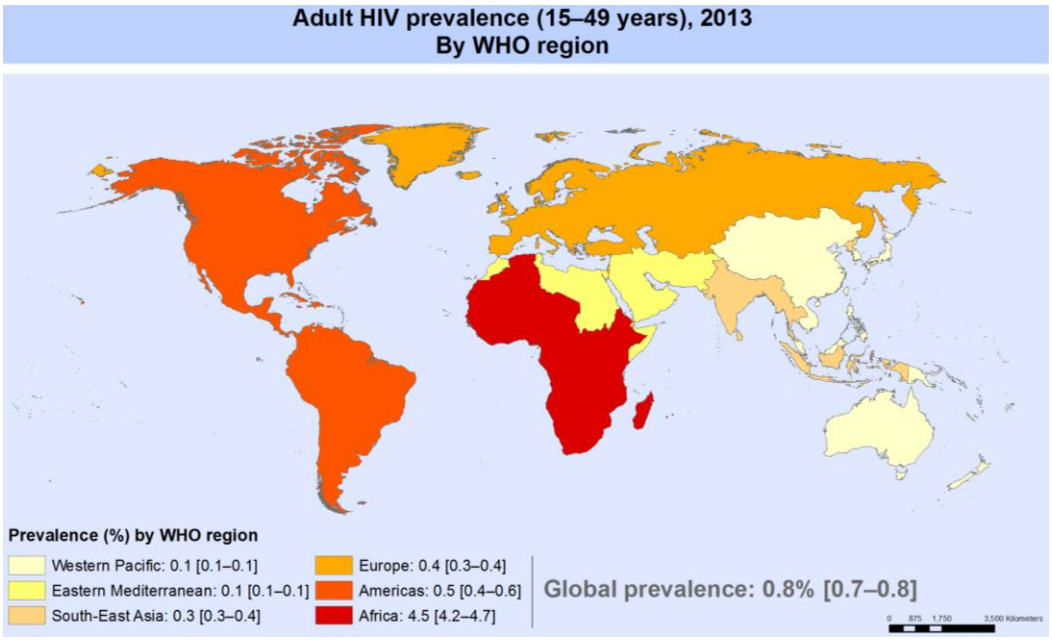


Figure 47: Adult HIV prevalence (15-49 years) for the 2013 distributed according to WHO regions. Reproduced from WHO Map production, HIS department

http://gamapserver.who.int/mapLibrary/Files/Maps/HIV_adult_prevalence_2013.png

It has been discovered that HIV infections causing Acquired Immunodeficiency Syndrome (AIDS) originates from two kinds of HIV; HIV-1, also called the main (M) group and HIV-2. These belong to a group of retroviruses called Lentiviruses (Zimmer, 2012). HIV-1 and HIV-2 share many similarities including intracellular mechanisms of replication, modes of transmission and clinical consequences, however, HIV-1 and HIV-

2 are considered to have been derived from separate simian to human transmissions; HIV-1 from SIVcpz and HIV-2 from SIVsmm (Figure 48) (Sharp & Hahn, 2011).

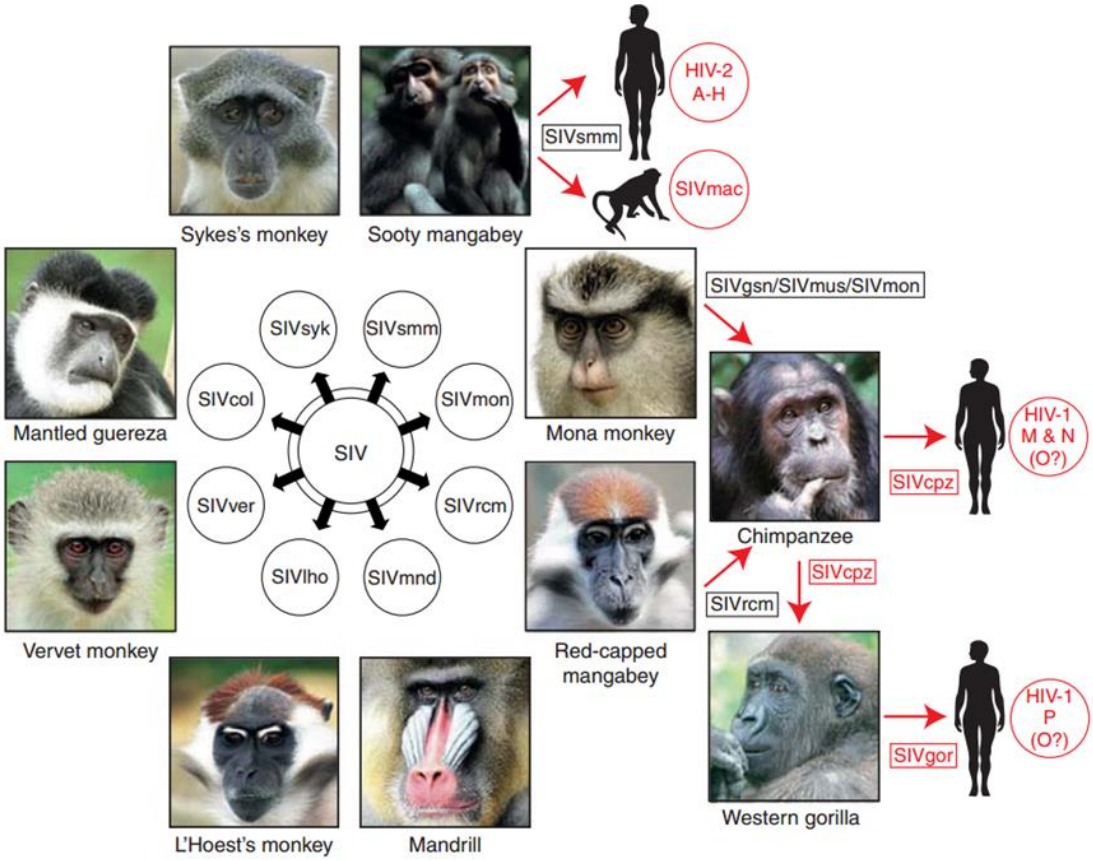


Figure 48: Origins of human AIDS viruses. Old World monkeys are naturally infected with more than 40 different lentiviruses, termed simian immunodeficiency viruses (SIVs) with a suffix to denote their primate species of origin (e.g., SIVsmm from sooty mangabeys). Known examples of cross-species transmissions, as well as the resulting viruses, are highlighted in red. Reproduced from (Sharp & Hahn, 2011).

The major clinical difference between the two infections is that progression to AIDS occurs more slowly in HIV-2 infection compared with HIV-1. Geographically, HIV-1 occurs worldwide whereas HIV-2 is mainly restricted to West Africa and communities in Europe with socioeconomic links to West Africa. In terms of infection, HIV-1 has greater infectivity and is more readily transmitted (Sharp & Hahn, 2011). In this study, HIV-1 is of interest to us.

6.2 Study rationale

HIV is spherically shaped with the inner contents surrounded by an outer phospholipid bilayer (Gelderblom, 1997). The outer surface of the virus is embedded with multiple types of glycoproteins (gp120 and gp41) and act as receptors for binding to host cell membrane receptors. The retrovirus HIV-1 infects host cells through its viral envelope glycoprotein gp120 which links to the CD4 membrane protein of immune system T-cells, monocytes/macrophages, eosinophils, dendritic cells in epithelial tissue, and microglial cells in the central nervous system (Fanales-Belasio et al., 2010). To complete host cell binding, the gp120 complex alters its shape to uncover a domain specific for a chemokine receptor such as *CCR5* or *CCXR4*; usually CCR5-tropic viruses almost always predominate (Weinberg et al., 2006) which it must also bind to (Pierson & Doms, 2003; Farzan et al., 1997; Wu et al., 1996). This double receptor attachment firmly secures the virus on the target cell and the gp41 fuses with the host cell membrane and the viral capsid subsequently is released into the host's cytoplasm (Fanales-Belasio et al., 2010).

All classes of defensins reportedly can suppress HIV replication (Wu et al., 2005; Guo et al., 2004; Chang et al., 2003; Mackewicz et al., 2003). Nakashima et al., described defensins as antiviral in 1993 and has shown that α-defensins inhibit HIV replication, Zhang et al., in 2002 suggested that they might play a role in controlling HIV in some subjects. Research has also found that the presence of low HIV-1 viral load can stimulate expression of hBD-2 and hBD-3 but not that of hBD-1, which is constitutively expressed in human oral epithelial cells (Quiñones-Mateu et al., 2003). According to Weinberg et al., (2006); hBD-2 and hBD-3 produced in oral epithelial cells play an essential role in the prevention of HIV infection as it has been found that oral transmission of HIV is rare (Rogers et al., 1990). Also, hBD-2 and -3 inhibit HIV-1 replication without being cytotoxic to immunocompetent cells. There are conflicting reports on the downregulation of expression of HIV co-receptors by DEFB; more specifically, hBD-1 and hBD-2 did not modulate cell surface HIV co-receptor expression by primary CD4+ T cells, whereas Quiñones-Mateu et al., 2003 showed that hBD-2 and hBD-3 mediated downregulation of surface *CXCR4* but not *CCR5* expression by peripheral blood mononuclear cells (PBMCs) at high salt conditions and in the absence

of serum. Interestingly, hBD-2 is constitutively expressed in healthy adult oral mucosa but the level seems to be diminished in HIV-infected individuals (Sun et al., 2005).

The clinical outcome of HIV-1 infection is highly variable and determined by complex interactions between virus, host and environment. Numerous human genetic factors including copy number variations have been reported to regulate HIV-1 disease (McLaren & Carrington, 2015). To date, studies have shown that copy numbers of certain host genes such as chemokine genes *CCL3L1* and *CCL4L*, ligands for *CCR5*; have been estimated to explain around 13% of variation in viral load (VL) at set point (Fellay et al., 2009). HIV VL set point has been used as a predictor of disease progression (Ho, 1996; Henrard et al., 1995), and more recently as a parameter of HIV vaccine efficacy (Buchbinder et al., 2008). The VL set point is defined as the viral load of a person infected with HIV, which stabilizes after a period of acute HIV infection (Mellors et al., 1997). Higher gene expression of *CCL3L1* and *CCL4L* (Ahuja et al., 2008; Gonzalez et al., 2005; Townson et al., 2002) and immune receptor family of the killer cell immunoglobulin-like receptors (KIR) genes have a defensive effect against HIV infection and progression to AIDS (Pelak et al., 2011). In addition, CNV of DEFB genes may also play an important part in proneness to HIV infection (Hardwick et al., 2012; Milanese et al., 2009). In Brazilian HIV-positive children, median copy number of *DEFB104* was lower compared with HIV-exposed uninfected children and healthy controls, suggesting that *DEFB104* may be involved in protection from vertical transmission of HIV (Milanese et al., 2009). On the other hand, Hardwick et al., 2012 conducted a cohort study in 1002 Ethiopian and Tanzanian patients investigating how DEFB copy number affects HIV VL immediately prior to administration of highly active antiretroviral therapy (HAART) and immune reconstitution following initiation of HAART and concluded that higher DEFB copy number variation is associated with increased HIV load prior to HAART ($p=0.005$) and poor immune reconstitution following initiation of HAART ($p=0.003$). In Mehlotra (2012), it has been shown that higher DEFB CN is an additional genetic factor associated with slower progression to AIDS in the mainly Caucasian Multicentre AIDS Cohort Study (MACS).

In this chapter, an association study of HIV viral load with DEFB CN will be carried out in African and Swiss cohorts, a case and control study will be performed to check if

DEFB CN differs between healthy individuals and HIV infected patients and finally, progression of HIV infected individuals to AIDS will be investigated against DEFB CN to check for an association.

6.3 Estimation of DEFB copy number in HIV cohorts

6.3.1 International AIDS Vaccine Initiative (IAVI) cohort

IAVI is a global not-for-profit, public-private partnership working to accelerate the development of vaccines to prevent HIV infection and AIDS. This perspective cohort was part of the Protocol C project launched in 2006 to further understanding of how HIV progresses and is transmitted. Volunteers were from Kenya, Uganda, Rwanda, Zambia and South Africa (2.1.3.1)

387 samples from the IAVI cohort were successfully typed for DEFB CN using PRT and CN was deduced using both the ML approach and raw PRT calculations. The diploid CN of DEFB varied from 1 to 9 with a modal CN of 4 as demonstrated in Figure 49 below.

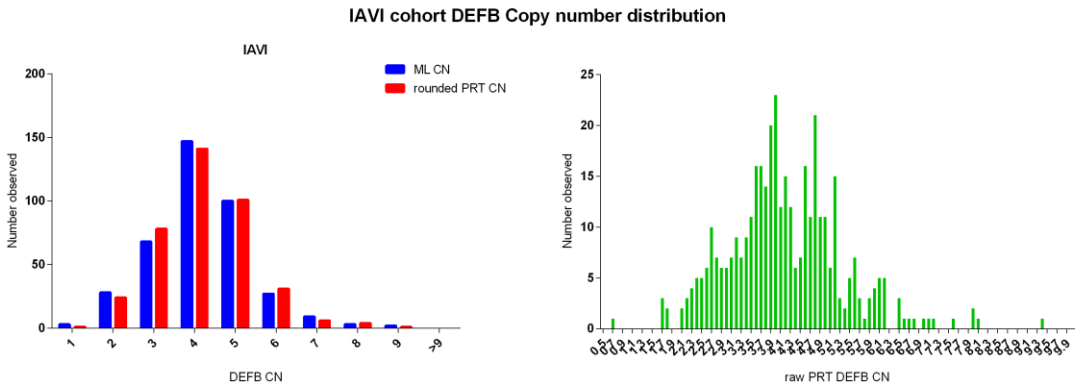


Figure 49: Distribution of IAVI cohort DEFB CN as estimated by ML approach, rounded PRT and raw PRT.

6.3.2 Swiss HIV Cohort Study (SHCS) cohort

The SHCS is representative for the Swiss HIV-epidemic collected by the Centre for HIV-AIDS Vaccine Immunology (CHAVI) founded by the National Institute of Allergy and Infectious Diseases (2.1.3.2).

3155 samples from the SHCS cohort were successfully typed for DEFB CN using PRT and CN was deduced using both the ML approach and raw PRT calculations. The diploid CN of DEFB varied from 1 to 10 with a modal CN of 4 as demonstrated in Figure 50 below

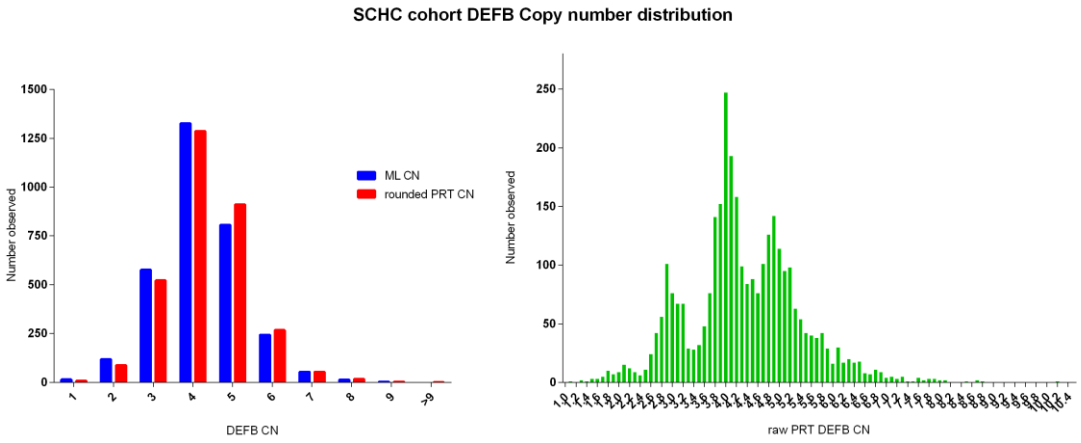


Figure 50: Distribution of SCHC cohort DEFB CN as estimated by ML approach, rounded PRT and raw PRT.

6.4 Association studies of DEFB CN in HIV cohorts

6.4.1 Viral Load at set point

HIV viral load set point has been used as a predictor of disease progression (Ho, 1996; Henrard et al., 1995), and more recently as a parameter of HIV vaccine efficacy (Buchbinder et al., 2008). The VL set point is defined as the viral load of a person infected with HIV, which stabilizes after a period of acute HIV infection (Mellors et al., 1997).

A total of 302 samples for the IAVI cohort and 1525 samples for the SHCS cohort had the required clinical information to test the association of DEFB CN with viral load (VL) at set point (Figure 51).

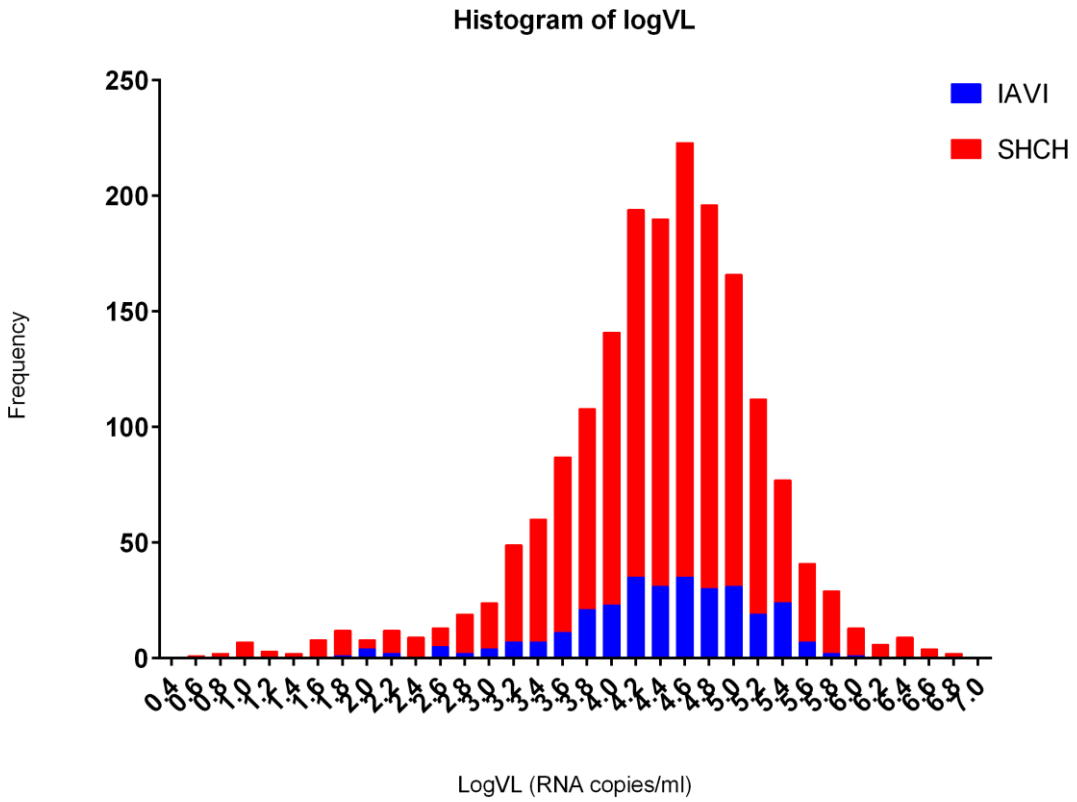


Figure 51: Histogram of the Log values of viral load at set point for the IAVI and SHCH cohorts.

The dependent variable; log transformed values of HIV viral load at set point, was modelled as a normal distribution with an identity link function. The model used type III sum of squares ANOVA and goodness-of-fit was analysed using Wald statistics. Predictor variables and factors were; sex, age, the first, second, and third principal components (PC1, PC2 and PC3) of genetic relatedness from GWAS data (Figure 52), and DEFB CN as estimated by ML approach, raw PRT CN and rounded CN.

The descriptive statistics for the analysed outcomes, factors and covariates are summarised below.

		LogVL (RNA copies/ml)	Age (years)	Sex	PC1	PC2	PC3	Raw PRT	Rounded PRT	ML CN
IAVI n=302	Min.	1.80	17	Male: 178 Female: 124	-0.19	-0.16	-0.20	1.60	2	1
	Max.	5.95	58		0.053	0.1	0.12	8.86	9	9
	Mean	4.40	32		$6.8x10^{-4}$	$-2.3x10^{-5}$	$1.3x10^{-3}$	4.18	4.25	4.23
SHCH n=1525	Min.	0.60	23	Male: 1236 Female: 289	-0.083	-0.60	NA	1.06	1	1
	Max.	6.86	92		0.054	0.11	NA	8.42	8	8
	Mean	4.36	50		$5.8x10^{-5}$	$3.9x10^{-4}$	NA	4.26	4.29	4.19

The principal components used here are a measure of genetic relatedness and reflect population ancestry from GWAS data. The below scree plot illustrates why the first 3 PCs were chosen for inclusion in association studies for the IAVI cohort

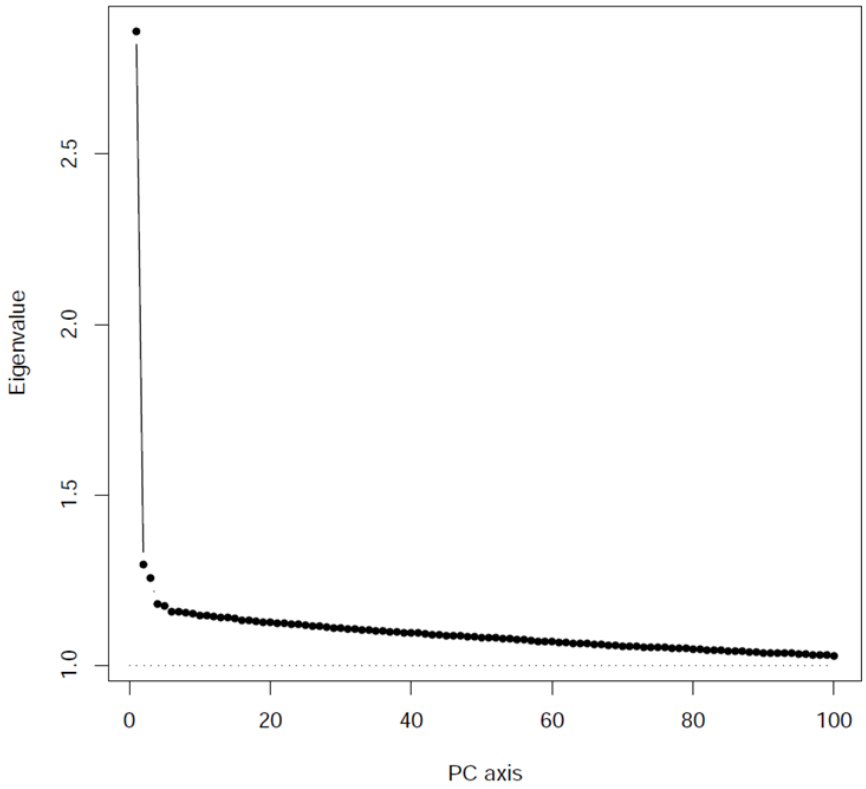


Figure 52: A scree plot for the IAVI cohort of genetic relatedness from GWAS data for the PCA showing the first 200 PCs which illustrates that 99% of the variation occurs in the first 3 PCs provided by Patrick Shea on December 9th 2013 via email.

The results for association are summarized in Table 12 and Table 13 below:

Table 12: Results of the association between log(VL) in IAVI cohort patients and the predictors; sex, age, PC1, PC2, PC3 and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Set to zero because this category is the reference for the specific parameter (*) denotes a significant p-value < 0.05

Outcome: Log(VL) for IAVI				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Sex: Male	0.371	0.196	0.546	$3.4x10^{-5}$*
Sex: Female	0[a]			
Age	0.004	-0.006	0.015	0.392
PC1	3.076	1.599	4.554	$4.5x10^{-5}$*
PC2	-1.584	-3.077	-0.090	0.038*
PC3	1.094	-0.459	2.647	0.167
DEFB raw-PRT CN	0.007	-0.064	0.077	0.853
Sex: Male	0.371	0.196	0.547	$3.3x10^{-5}$*
Sex: Female	0[a]			
Age	0.004	-0.006	0.015	0.392
PC1	3.076	1.598	4.554	$4.5x10^{-5}$*
PC2	-1.579	-3.072	-0.087	0.038*
PC3	1.095	-0.458	2.648	0.167
DEFB rounded CN	0.003	-0.066	0.072	0.924
Sex: Male	0.371	0.196	0.547	$3.2x10^{-5}$*
Sex: Female	0[a]			
Age	0.004	-0.006	0.015	.395
PC1	3.066	1.588	4.543	$4.8x10^{-5}$*
PC2	-1.565	-3.057	-0.073	0.040*
PC3	1.113	-0.441	2.666	0.160
DEFB ML CN	-0.015	-0.080	0.050	0.652

In IAVI cohort, log(VL) at set point was found to be higher in males compared to females. A significant p-value for the PC1 and PC2 value shows that viral load at set point is influenced by ancestry. DEFB CN was not associated with viral point at set point.

Table 13: Results of the association between log(VL) in SHCH cohort patients and the predictors; sex, age, PC1, PC2 and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Set to zero because this category is the reference for the specific parameter (*) denotes a significant p-value < 0.05

Outcome: Log(VL) for SHCH				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Sex: Male	0.439	0.330	0.548	$2.8x10^{-15}$*
Sex: Female	0[a]			
Age	0.002	-0.002	0.006	0.270
PC1	0.498	-1.650	2.646	0.650
PC2	-0.357	-2.473	1.760	0.741
DEFB raw-PRT CN	-0.020	-0.060	0.021	0.335
Sex: Male	0.439	0.330	0.548	$2.8x10^{-15}$*
Sex: Female	0[a]			
Age	0.002	-0.002	0.006	0.269
PC1	0.500	-1.649	2.649	0.648
PC2	-0.350	-2.466	1.767	0.746
DEFB rounded CN	-0.015	-0.055	0.026	0.480
Sex: Male	0.438	0.329	0.547	$3x10^{-15}$*
Sex: Female	0[a]			
Age	0.002	-0.002	0.006	0.267
PC1	0.477	-1.671	2.625	0.663
PC2	-0.341	-2.457	1.775	0.752
DEFB ML CN	-0.026	-0.066	0.014	0.203

In the SHCS cohort, males were found to have a higher viral load at set point compared to females. Neither genetic relatedness nor DEFB CN was found to an effect as was observed in IAVI cohort.

6.4.2 Case-Control comparison

This study aimed to investigate whether DEFB CN differs in HIV infected individuals compared to uninfected individuals. It was carried out on the 3155 participants of SHCS cohort as cases and matched to 2034 HIV-negative individuals that were genotyped as controls by Dr. Rob Hardwick (a former lab member) for other purposes. The control cohort includes samples from Human Random Control Panels, Gedling

Cohort and Leicester Children Cohort. The analysis was carried out using DEFB CN as estimated by the raw PRT calculations.

The descriptive statistics of the DEFB CN as calculated by the raw PRT for controls and cases is summarized in Table 14 below

Table 14: Descriptive statistics of DEFB CN as calculated by raw PRT for controls and cases.

	Controls	Cases
N=	2034	3155
Minimum	0.85	1.06
25% percentile	3.58	3.74
Median	4.11	4.17
75% percentile	4.87	4.91
Maximum	10.95	10.15
Mean	4.23	4.28
Standard Deviation	1.11	1.04
Lower 95% CI	4.18	4.24
Upper 95% CI	4.28	4.32

The cumulative frequency of DEFB CN in cases and controls is shown in Figure 53 below

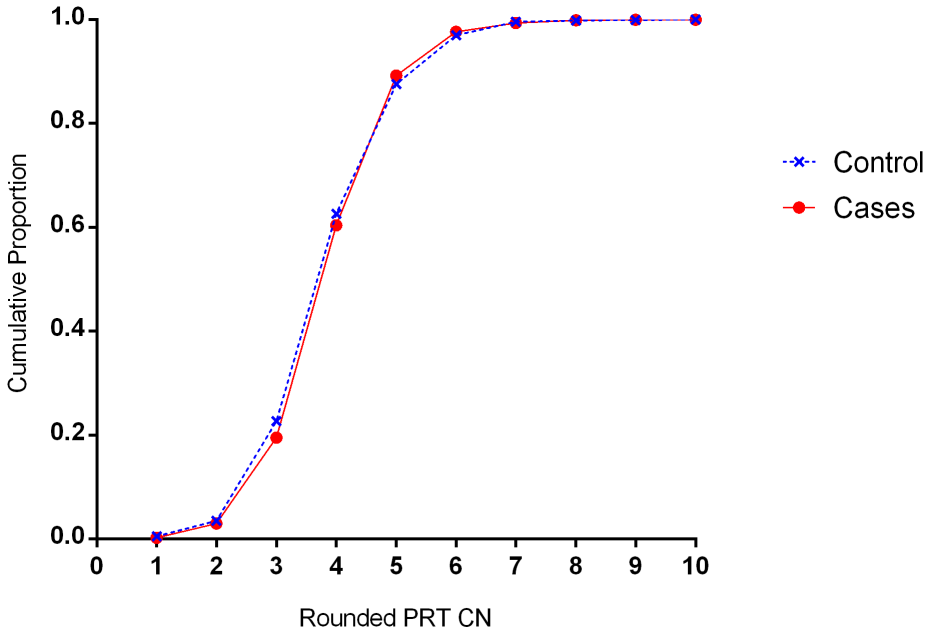


Figure 53: Cumulative frequency distribution of rounded raw PRT DEFB CN in HIV cases and controls.

An association study was carried out to investigate if DEFB CN differed between cases and controls ('cases' were modelled as response, treating 'controls' as the reference category. This study was carried out using the logit link function, binomial distribution of the generalised linear model with DEFB CN as estimated by ML approach, raw PRT CN and rounded CN as covariates.

Table 15: Results of the association between participant type (case vs. control) and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN. ([a]) Cases were modelled as response, treating 'controls' as the reference category. (*) denotes a significant p-value < 0.05

Outcome: Cases vs. Controls [(a)]				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
DEFB raw-PRT CN	0.046	-0.006	0.098	0.084
DEFB rounded CN	0.033	-0.019	0.085	0.214
DEFB ML CN	-0.009	-0.061	0.042	0.725

Unfortunately, the DEFB CN does not differ between people infected with HIV and healthy individuals.

6.4.3 HIV Progression Studies

In this section, whether DEFB CN was a contributor in the progression of HIV infection to AIDS was investigated. Clinical progression data for 229 HIV infected individuals from the SHCS cohort was available. The mean DEFB CN as calculated by PRT was found to be 4.37. The data has been divided into two groups; those having a DEFB CN below 4.37 and those having a DEFB CN of 4.37 and above.

	N	**Progressed**	**Did not progress**
< 4.37	120	75	45
≥ 4.37	109	69	40

The Kaplan-Meier estimator for survival function was used. Individuals who did not progress to AIDS or have dropped out of the study before the end time were marked as censored. The survival data for the two groups is as follows:

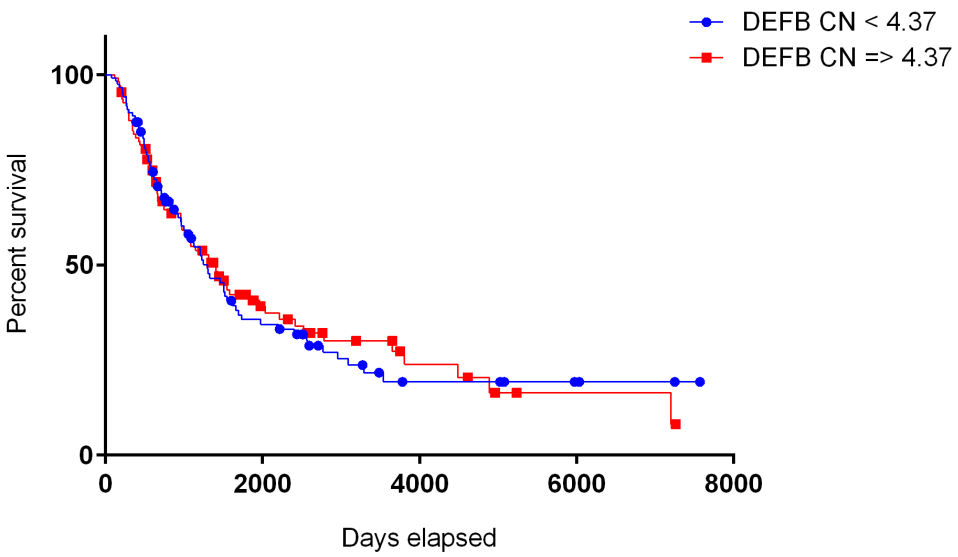


Figure 54: a survival curve for the progression of HIV infected individuals to AIDS. The participants were divided into groups; individuals with DEFB CN below the mean (4.37) and individuals with DEFB CN of 4.37 and above.

The median survival times for individuals with DEFB CN <4.37 and individuals with DEFB CN ≥4.37 is 1299 and 1398 respectively. This means that at day 1299, 60 individuals were still alive for the group which has a DEFB CN of < 4.37, whereas at day 1398, 54 individuals were still alive of those having a DEFB CN of ≥ 4.37.

The log-rank test (Mantel-Cox test) reported a p-value of 0.85 which shows that DEFB CN has no significant difference in progression from HIV infection to AIDS.

To further confirm the result, a Cox regression was carried out and the results are summarized below:

Table 16: Cox regression result summary showing the variables year of birth, sex, log transformed viral load values and DEFB CN as calculated using the raw-PRT approach. (*) denotes a significant p-value < 0.05.

Outcome: Cox Regression					
Variable	**B**	**SE**	**Wald**	**p-value**	**Exp(B)**
Year of Birth	-0.004	0.009	0.239	0.625	0.996
Sex	0.264	0.205	1.660	0.198	1.302
Log(VL)	0.650	0.098	44.301	$2.8x10^{-11}$ *	1.915
DEFB Raw-PRT CN	0.152	0.084	3.253	0.071	1.164

Table 16 above shows that the only parameter from the ones tested that affect the progression of a participant to AIDS is viral load. The higher the viral load is, the quicker the HIV infected person progresses to AIDS. This confirms the previous results of the Kaplan-Meier function that DEFB CN has no effect on progression pace from HIV infection to AIDS.

In a different attempt, using DEFB CN as calculated by ML, the data was divided into two groups; those having a DEFB CN equal to or below 4 and those having a DEFB CN of above 4. This was done in an attempt to replicate the analysis criteria Hardwick et al., (unpublished) carried out during their investigations.

	N	Progressed	Did not progress
DEFB CN ≤ 4	142	90	52
DEFB CN > 4	87	54	33

The Kaplan-Meier estimator for survival function was used. Individuals who did not progress to AIDS or have dropped out of the study before the end time were marked as censored. The survival data for the two groups is as follows:

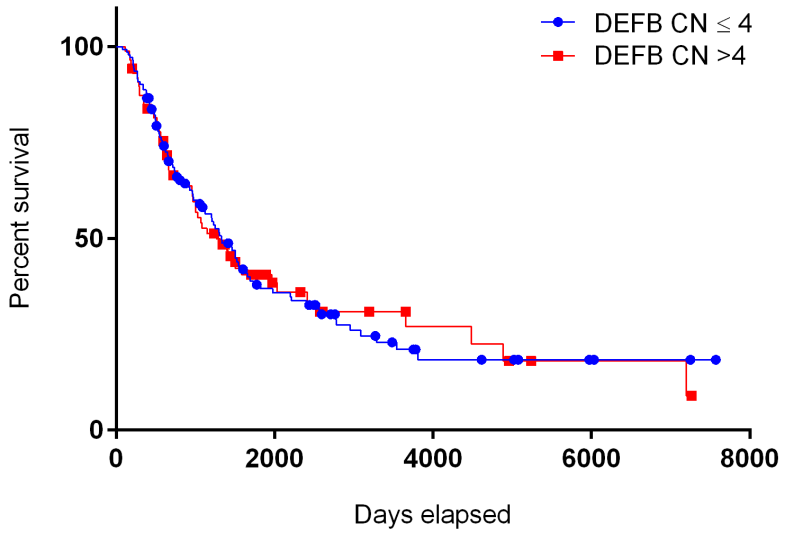


Figure 55: a survival curve for the progression of HIV infected individuals to AIDS. The participants were divided into groups; individuals with DEFB CN as calculated by ML above the median (4) and individuals with DEFB CN of 4 and below.

The median survival times for individuals with DEFB CN ≤ 4 and individuals with DEFB CN > 4 is 1330 and 1267 respectively. This means that at day 1330, 45 individuals were still alive for the group which has a DEFB CN of ≤ 4, whereas at day 1267, 27 individuals were still alive of those having a DEFB CN of > 4.

The log-rank test (Mantel-Cox test) reported a p-value of 0.98 which shows that DEFB CN has no significant difference in progression from HIV infection to AIDS.

To further confirm the result, a cox regression was carried out and the results are summarized below:

Outcome: Cox Regression					
Variable	**B**	**SE**	**Wald**	**p-value**	**Exp(B)**
Year of Birth	-0.005	0.009	0.293	0.588	0.995
Sex	0.287	0.204	1.99	0.159	1.33
Log(VL)	0.638	0.097	43.18	$4.9x10^{-11}$ *	1.89
ML CN median and above	-0.096	0.174	0.308	0.579	0.908

6.5 Discussion

The basis of this study was to replicate the findings of Hardwick et al., 2012 that higher DEFB CNV is associated with increased HIV load prior to HAART (p=0.005) and poor immune reconstitution following initiation of HAART (p =0.003). This study was carried out on an African cohort (IAVI) and a Swiss cohort (SHCS). The table below compares the current cohorts with the cohort Hardwick et al., 2012 used.

Variable	**Hardwick et al., 2012**	**IAVI**	**SHCS**
Number of participants	1002	302	1525
Cohort origin	African	African	Swiss
Dependent variable	HIV viral load before the initiation of HAART	HIV viral load at set point	HIV viral load at set point

Participants of the IAVI cohort were of African origin; however the dependent variable analysed in this study was viral load at set point not as in Hardwick et al. 2012; which was VL before the initiation of HAART. Also, the fact that the IAVI cohort had a sample

size of 302 might has reduced the power of the study in detecting an association if an association existed. As for the SHCH cohort, the sample size in terms of population studies would have sufficed, however neither the dependent variable nor the origin of the population were as previously analysed. The population heritage might be a major cause in not having the data replicated because according to Petrovski et al. (2011), the subset of individuals that have Δ32 deletion or the rare m303T>A mutation in their *CCR5* gene conferring complete resistance to HIV-1 infection that use the *CCR5* as co-receptor is only found in individuals with northern European or Central Asian heritage. Another fact to factor in would be that Europeans and West Africans are exposed to HIV-2 subset rather than HIV-1. It could be possible that the cohort recruited contained the two subtypes of HIV and hence the effect has been masked. These two studies are not a true replication of the Hardwick et al., (2012) study.

As for the case-control study; cases had a mean DEFB CN of 4.28 whereas controls of 4.23. The p-values of 0.084, 0.214 and 0.725 for DEFB CN as calculated by raw PRT, rounded PRT and ML CN respectively of the association studies confirms that HIV infection does not favour high nor low DEFB CNV. With regards to survival analysis, it was evident from the results that DEFB CNV plays no role in the progression of an individual infected with HIV to AIDS regardless of DEFB CN division groups contradicting data published by (Mehlotra et al., 2012). Since the cohort used by Mehlotra et al., (2012) is predominately Caucasian, the findings together with the results of Hardwick et al., (2012) using a Sub-Saharan African Cohort and Milanese et al.,(2009) using a Brazilian Cohort indicate that the association of DEFB CNV and infection/disease progression may be population specific.

In order to conclude whether DEFB CNV is associated with HIV viral load prior to HAART and immune reconstitution after HAART, a true biological replicate of the cohort used by Hardwick et al., 2012 is necessary. It has to be a cohort of a sample size of 1000 and above, Sub-Saharan African origin, exposed by HIV-1 only and has matching clinical data readings of viral load.

7 Association of β-defensin copy number variation in other disease cohorts

7.1 Urinary Tract Infection

Paediatric urinary tract infections (UTI) account for 0.7% of physician office visits and 5–14% of emergency department visits by children annually in North America (Shaikh et al., 2008). In the USA, it is the most common infection in youngsters below the age of 6 years with an incidence rate higher in females (3-7%) than in males (1-2%) (Elder et al., 1997) and is mostly caused by *E.coli*, a gram-negative bacteria.

UTI in young children may be a marker for abnormalities of the urinary tract including vesicoureteral reflux (VUR) and reflux nephropathy. VUR is the commonest abnormality with a prevalence of 1% in all children and about 33% in children following first UTI (Carpenter et al. 2013; Koff et al., 1998; Weiss et al., 1992). VUR is characterised by the retrograde urine flow from the bladder to the ureters and kidneys (Chesney et al., 2008). Similarly, children with VUR can have recurrent UTIs which in turn lead to renal scarring and chronic kidney disease known as reflux nephropathy. Renal scarring is detected using the Dimercaptosuccinic acid (DMSA) radionuclide scan. DMSA scan requires 3-4 hours of time and the isotope used also emits radiation. Because of the length of the study, children often need sedation to take images during DMSA scan. Although lesser than CT scan, a substantial amount of radiation is involved in DMSA scanning. Plain ultrasound has a low sensitivity (47.2%) to detect renal scars (Moorthy et al., 2004).

Reflux nephropathy is the fourth leading cause of End Stage Renal Disease (ESRD) in the US (Spencer et al., 2011). Prevention of recurrent UTI and reduction in the rate and severity of renal scarring can be achieved by antibiotic prophylaxis, surgical interventions and follow-up (Hoberman et al., 2014; Brandström et al., 2010).

7.1.1 Study rationale

Valore et al. in 1998 discovered that defensins are expressed in the urinary tract of humans, specifically hBD-1 and hBD-2. hBD-1 mRNA is expressed constitutively by the epithelia lining of the loop of Henle, the distal tubule and the collecting duct of the kidney, and is always detected in urine (Becknell et al., 2013; Hiratsuka et al., 2000; Valore et al., 1998), whereas expression of hBD-2 was induced during urinary tract infection (Lehmann et al., 2002).

The first evidence which suggest that defensins play a role in the defence of the kidney come from studies of mouse Beta Defensin-1 (mBD-1) deficient mice. When compared to wild-type mice, the deficient mice had high numbers of colony forming units when their urine was plated out on agar, and about 30% of healthy mBD-1 knockout mice had *Staphylococcus* species in bladder urine (Morrison et al., 2002). The second evidence which demonstrated that was in a rat model infected with uropathogenic *E. coli* (UPEC), over-expression of exogenous hBD-2 not only significantly decreased the amounts of UPEC in bladder and urine, but also reduced infiltration of inflammatory cells, mucosal damage, and interstitial oedema in the bladder tissue (Zhao et al., 2011).

To date, however, no comparable study evaluating the association between DEFB CN and factors affecting VUR patients and recurrent UTIs. This study will result in novel information about copy number variation of the DEFB region in the consequences of VUR in patients.

7.1.2 Estimation of DEFB copy number in VUR and UTI samples

The Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR) study is a double-blind, randomized, placebo-controlled trial which recruited 607 children aged 2 to 72 months from 19 paediatric sites across North America assigned to receive daily doses of either antibiotics or placebo for 2 years. Blood was collected every 6 months throughout the trial to study genetic and biochemical determinants of VUR, recurrent UTI, and renal scarring (Keren et al., 2008). The study was approved by the Nationwide Children's Hospital Institutional Review Board, Nationwide Children's Hospital, Columbus, Ohio, USA.

DNA of 456 individuals from the RIVUR study was sent to us by our collaborator Dr. David Hains (Ohio, USA). 42 samples failed to type after two attempts, and this could be attributed to low DNA quality. 414 samples from the RIVUR cohort were successfully typed for DEFB CN using PRT and CN was deduced using both the ML approach and raw PRT calculations (2.6.4 2.6.5). The diploid CN of DEFB varied from 2 to 9 with a modal CN of 4 as demonstrated in Figure 56 below.

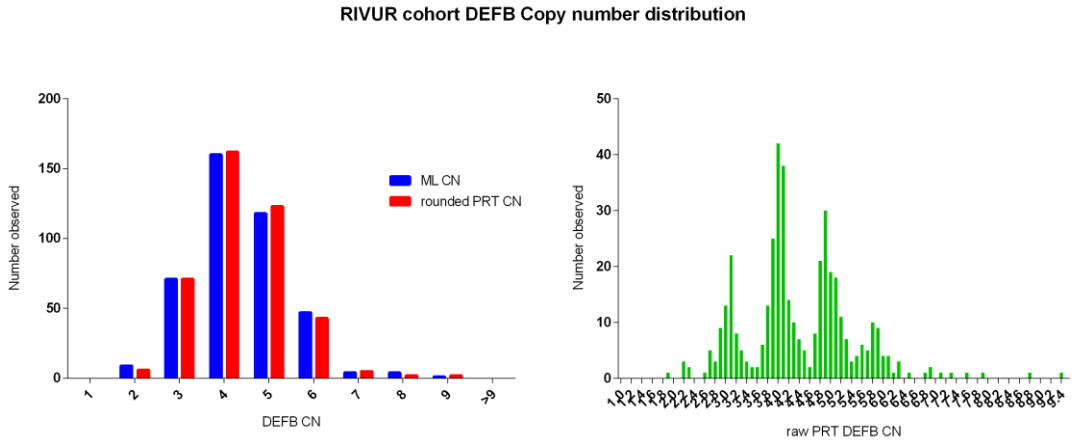


Figure 56: Distribution of RIVUR sample DEFB CN as estimated by ML approach, rounded PRT and raw PRT.

7.1.3 Association studies in VUR and UTI samples

For statistical analysis, only Non-Hispanic, Caucasian, female samples were analysed, as this subset had the full set of clinical information needed to be analysed. Therefore, a total of 290 samples were included in the association studies.

7.1.3.1 Number of breakthrough infections

The first outcome to be investigated was the number of UTI breakthroughs that occurred in the cohort. Among the participants, the minimum number of infections was 0 and the maximum was 4 as illustrated in Figure 57 below.

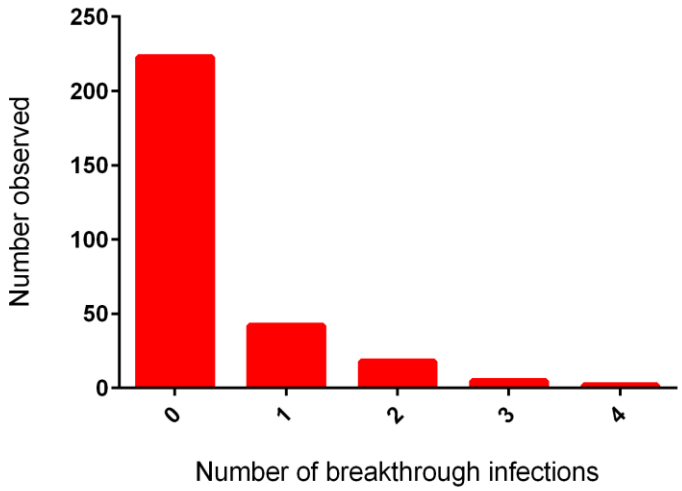


Figure 57: A histogram showing the number of UTI breakthroughs occurring in the RIVUR cohort plotted against the frequency of each.

A generalized linear model was constructed where the dependent variable (number of breakthrough infections) was modelled as Poisson distribution, under the log link function. In this model, treatment group and bowel and bladder dysfunction (BBD) were assigned as factors, age (in months) and DEFB CN as estimated by ML approach, raw PRT CN and rounded CN as covariates, and calculated using Type III sum of squares. A summary of the variables is shown in Table 17 below:

Table 17: A summary of the factors and covariates used in the association studies of the RIVUR cohort

Variable		Number	Percent
Treatment group	Antibiotics	147	50.7%
	Placebo	143	49.3%
BBD	Yes (1)	56	19.3%
	No (0)	113	39%
	No Information	121	41.7%
	Minimum	**Maximum**	**Mean**
Age	2	71	22
Raw PRT CN	1.82	8.73	4.36
Rounded PRT CN	2	9	4.40
ML CN	2	8	4.38

Results of the association between the number of UTI breakthroughs in VUR patients as an outcome and the predictors; treatment group, BBD, age and DEFB raw PRT CN,

DEFB rounded CN and DEFB ML CN respectively are summarized in the tables 18, 19 and 20 below.

Table 18: Results of the association between the number of infections in VUR patients as an outcome and the predictors; treatment group, BBD, age and DEFB raw PRT CN. ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: Number of breakthrough infections				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	-0.981	-1.54	-.423	$5.7x10^{-4}$ *
Treatment group: Placebo	0[a]			
BBD: no	-0.557	-1.04	-0.073	0.024*
BBD: yes	0[a]			
Age (months)	0.017	0.005	0.030	0.006*
DEFB raw PRT CN	0.209	-0.032	0.451	0.089

Table 19: Results of the association between the number of infections in VUR patients as an outcome and the predictors; treatment group, BBD, age and DEFB rounded CN. ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: Number of breakthrough infections				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	-0.975	-1.53	-0.419	$5.8x10^{-4}$ *
Treatment group: Placebo	0[a]			
BBD: no	-0.592	-1.077	-0.107	0.017*
BBD: yes	0[a]			
Age (months)	0.018	0.005	0.03	0.006*
DEFB rounded CN	0.231	-0.005	0.468	0.055

Table 20: Results of the association between the number of infections in VUR patients as an outcome and the predictors; treatment group, BBD, age and DEFB ML CN ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: Number of breakthrough infections				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	-0.982	-1.538	-0.427	$5.3x10^{-4}$ *
Treatment group: Placebo	0[a]			
BBD: no	-0.614	-1.099	-0.128	0.013*
BBD: yes	0[a]			
Age (months)	0.017	0.004	0.029	0.008*
DEFB ML CN	0.261	0.010	0.512	0.041*

The above results show that number of breakthrough infections decrease with antibiotics, decrease with children who are not potty trained, increases with age and increase with DEFB CN as estimated by ML approach. However, the results are barely significant with DEFB CN as estimated by raw PRT calculations and rounded PRT CN.

7.1.3.2 First infection caused by E.coli

The second outcome to be investigated was whether the first UTI a patient contracted in the RIVUR cohort was caused by *E.coli*. Clinical information for this analysis was available for 169 patients. *E.coli* was the cause of the first infection for 152 patients (90%) whereas the first infection of the remaining patients (10%) was caused by different organisms.

A generalized linear model was constructed where the dependent variable (first infection caused by *E.coli*) was modelled as binomial distribution, under the logit link function. In this model, treatment group and bowel and bladder dysfunction (BBD) were assigned as factors, age (in months) and DEFB CN as estimated by ML approach, raw PRT CN and rounded CN as covariates, and calculated using Type III sum of squares.

Table 21: Results of the association between *E.coli* as the entry infection organism in VUR patients and the predictors; treatment group, BBD, age and DEFB raw-PRT CN. ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: First infection caused by *E.coli*				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	-0.782	-1.841	0.278	0.148
Treatment group: Placebo	0[a]			
BBD: no	-0.076	-1.198	1.045	0.894
BBD: yes	0[a]			
Age (months)	-0.025	-0.052	0.002	0.069
DEFB raw-PRT CN	-0.034	-0.541	0.474	0.897

Table 22: Results of the association between *E.coli* as the entry infection organism in VUR patients and the predictors; treatment group, BBD, age and DEFB rounded CN. ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: First infection caused by *E.coli*				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	-0.770	-1.83	0.291	0.155
Treatment group: Placebo	0[a]			
BBD: no	-0.079	-1.201	1.043	0.890
BBD: yes	0[a]			
Age (months)	-0.025	-0.052	0.002	0.064
DEFB rounded CN	-0.124	-0.605	0.357	0.613

Table 23: Results of the association between *E.coli* as the entry infection organism in VUR patients and the predictors; treatment group, BBD, age and DEFB ML CN ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: First infection caused by *E.coli*				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	-0.781	-1.841	0.279	0.149
Treatment group: Placebo	0[a]			
BBD: no	-0.073	-1.194	1.048	0.898
BBD: yes	0[a]			
Age (months)	-0.025	-0.052	0.002	0.069
DEFB ML CN	-0.037	-0.544	0.469	0.886

None of the factors and covariates, including DEFB CN shows an association with whether the first infection was cause by *E.coli*.

7.1.3.3 Breakthrough infection caused by E.coli

The third outcome to be investigated was whether the breakthrough UTI of a patient was caused by *E.coli*. There were clinical data for 44 participants. The breakthrough infection was caused by *E.coli* in 86.4% of VUR patients in this cohort.

A generalized linear model was constructed where the dependent variable (Breakthrough infection caused by *E.coli*) was modelled as binomial distribution, under the logit link function. In this model, treatment group and bowel and bladder dysfunction (BBD) were assigned as factors, age (in months) and DEFB CN as estimated by ML approach, raw PRT CN and rounded CN as covariates, and calculated using Type III sum of squares.

Table 24: Results of the association between *E.coli* as the breakthrough infection organism in VUR patients and the predictors; treatment group, BBD, age and DEFB raw-PRT CN. ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: Breakthrough infection caused by *E.coli*				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	21.17	-42122	42164	0.999
Treatment group: Placebo	0[a]			
BBD: no	-1.418	-3.779	0.944	0.239
BBD: yes	0[a]			
Age (months)	-0.022	-0.075	0.030	0.406
DEFB raw-PRT CN	-0.477	-1.328	0.375	0.273

Table 25: Results of the association between *E.coli* as the breakthrough infection organism in VUR patients and the predictors; treatment group, BBD, age and DEFB rounded-PRT CN. ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: Breakthrough infection caused by *E.coli*				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	21.16	-41881	41923	0.999
Treatment group: Placebo	0[a]			
BBD: no	-1.429	-3.789	0.931	0.235
BBD: yes	0[a]			
Age (months)	-0.026	-0.078	0.026	0.330
DEFB rounded CN	-0.491	-1.347	0.365	0.261

Table 26: Results of the association between *E.coli* as the breakthrough infection organism in VUR patients and the predictors; treatment group, BBD, age and DEFB ML CN. ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: Breakthrough infection caused by *E.coli*				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	21.23	-41892	41934	0.999
Treatment group: Placebo	0[a]			
BBD: no	-1.405	-3.774	0.963	0.245
BBD: yes	0[a]			
Age (months)	-0.023	-0.075	0.029	0.388
DEFB ML CN	-0.511	-1.39	0.369	0.255

None of the factors and covariates is significantly associated with whether the breakthrough infection is caused by *E.coli*.

7.1.3.4 The development of new kidney scars

The last outcome that was investigated was whether patients in this cohort developed new kidney scars due to their VUR or not. Clinical information was available for 150 patients; only 10.7% of the RIVUR participants had developed new kidney scars in the course of the study.

A generalized linear model was constructed where the dependent variable (new kidney scar formation) was modelled as binomial distribution, under the logit link function. In this model, treatment group and bowel and bladder dysfunction (BBD) were assigned as factors, age (in months) and DEFB CN as estimated by ML approach, raw PRT CN and rounded CN as covariates, and calculated using Type III sum of squares.

Results of which are summarized in the tables 27, 28 and 29 below.

Table 27: Results of the association between the development of new kidney scars in VUR patients and the predictors; treatment group, BBD, age and DEFB raw-PRT CN. ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: New kidney scar formation				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	0.749	-0.385	1.882	0.169
Treatment group: Placebo	0[a]			
BBD: no	0.084	-1.067	1.235	0.887
BBD: yes	0[a]			
Age (months)	-0.025	-0.053	0.003	0.081
DEFB raw-PRT CN	-0.348	-0.913	0.217	0.228

Table 28: Results of the association between the development of new kidney scars in VUR patients and the predictors; treatment group, BBD, age and DEFB rounded CN. ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: New kidney scar formation				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	0.754	-0.382	1.89	0.193
Treatment group: Placebo	0[a]			
BBD: no	0.089	-1.063	1.241	0.879
BBD: yes	0[a]			
Age (months)	-0.026	-0.054	0.003	0.076
DEFB rounded CN	-0.385	-0.943	0.172	0.175

Table 29: Results of the association between the development of new kidney scars in VUR patients and the predictors; treatment group, BBD, age and DEFB ML CN. ([a]) Set to zero because this category is the reference for the specific parameter. (*) denotes a significant p-value < 0.05

Outcome: New kidney scar formation				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Treatment group: Antibiotics	0.780	-0.362	1.921	0.181
Treatment group: Placebo	0[a]			
BBD: no	0.153	-1.012	1.317	0.797
BBD: yes	0[a]			
Age (months)	-0.025	-0.053	0.003	0.085
DEFB ML CN	-0.454	-1.028	0.120	0.121

None of the factors and covariates is significantly associated with whether new kidney scars are formed in the patients in the course of the study.

7.1.4 Discussion

The present study was designed to test the hypothesis that CNV at the DEFB region might be associated with consequences of VUR patients. According to the above results, patients receiving antibiotics had a lower number of breakthrough infections than those receiving the placebo. As for the patient's bowel and bladder dysfunction, those who are unable to empty their bladder properly have less numbers of infections compared to those who are potty trained. When it comes to age, it can be deduced that the older the patient is; the number of infections increases whereas the odds of developing new kidney scars decreases.

Higher DEFB CN as estimated by ML approach showed an association with a higher number of breakthrough infections (p= 0.041). However, since the significance level is close to the 0.05 cut-off point and neither DEFB CN as estimated by raw PRT calculations (p=0.089) and rounded PRT CN (p=0.055) were found to be associated with the same dependent variable, a deeper investigation is required to confirm or refute the result. Also, given that the highest p value is given with DEFB CN as calculated using raw PRT data suggests this might be an artefact of CNV calling.

In conclusion, this work found no evidence that CNVs at the DEFB region affects susceptibility to VUR or UTIs. A replication study must be carried out with a larger cohort to minimize any CNV calling artefact. Also, the potential interactions between *DEFB* genes and environmental conditions (VUR, age, ethnicity… etc.) should explore more complex hypotheses, including gene–gene and gene–environment.

7.2 Hypertension

Blood pressure is usually measured in millimetres of mercury (mmHg) and recorded as a fraction with the maximum pressure following systole of the left ventricle of the heart over the minimum pressure that accompanies diastole hence; it varies depending on the cardiac output and the total vascular resistance (Frese et al., 2011). Blood pressure is also classified as systolic, which is the highest measurement of the pressure in arteries during heart contraction, whereas, diastolic blood pressure is the lowest measurement of pressure in arteries when the heart relaxes (Strandberg & Pitkala, 2003). Blood pressure readings vary between normal healthy individuals; however, an ideal pressure level in adults is suggested to be 120 systolic over 80 diastolic mmHg (Frese et al., 2011). Normal levels of systolic and diastolic blood pressure is important for the efficient function of vital organs such as the brain, kidneys, heart and overall health (WHO, 2013a).

Hypertension or high blood pressure is defined as a systolic blood pressure equal to or above 140 mm Hg and/or diastolic blood pressure equal to or above 90 mm Hg (WHO, 2013a). Globally, hypertension accounts for approximately 17 million deaths a year. Of these, complications of hypertension such as kidney disease, stroke, heart attack, embolism and aneurysms account for 9.4 million deaths worldwide annually (Lim et al., 2012).

A number of risk factors, including environmental, life style choices, medical conditions, and genetic factors have been identified for developing hypertension, and vary from poor nutritional habits, lack of exercise, smoking, obesity, high cholesterol levels, diabetes (Yusuf et al., 2004, 2014) and Left Ventricular Hypertrophy (LVH) (Boonpeng & Yusoff, 2013).

7.2.1 Study rationale

The basis of exploring the association of DEFB CNV in hypertension was formed on the recent discovery of β-Defensins in the heart. This included hBD-3 expression in adult human heart (Jia et al., 2001), pBD1 (porcine β-defensin 1) in pigs (Zhang et al., 1998), eBD1 (equine β-defensin 1) in the horse (Davis et al., 2004), Defb1 (murine β-defensin

1) in mice (Morrison et al., 2002), and rBD1 (rat β-defensin 1) in the rat (Page & Malik, 2003). Linde et al in 2013 documented that at least 7 different DEFB (i.e. rBD1/3/10/11/15/18 and 33) are constitutively expressed in the Rat myocardium and their expression is triggered by a high- fat-diet feeding. It was also published that these DEFBs influence local monocyte migration in the heart which suggest they respond to exogenous danger-signals, and act within the context of a myocardial "first-line-of-defence" (Linde et al., 2013).

In addition, published evidence suggests that a deleted region in DEFB repeat at the 8p23.1 chromosomal position is associated with congenital heart malformations and congenital diaphragmatic hernia (CDH) due to the loss of the *GATA4* gene. This in turn can cause any one of aortic outflow obstruction, severe mitral disease, left ventricular dysfunction or limited compliance of the systemic ventricle, all of which can alter the systolic blood pressure and skew normal blood pressure readings as reviewed in Hauser (2003).

All of the above suggests that DEFB is present in the cardiovascular system yet information on cardiac DEFB is still sparse and further research is needed to elucidate the actual role played by them in heart disease in general. However, the strongest evidence to date which gave rise to our interest to investigate possible connection between DEFB CNV and blood pressure is based on Liu et al., 2013's results which supported the hypothesis that hBD-2 was significantly downregulated in sera of patients with hypertension and that hBD-2 acted as an opener of Ca^{2+}-activated Potassium channels and induced vasodilation and hypotension in monkeys.

The study will explore whether copy number variation at the DEFB region is associated with blood pressure.

7.2.2 Estimation of DEFB copy number in hypertension cohorts

7.2.2.1 Young Men Cardiovascular Association (YMCA) cohorts

YMCA 1 consists of 1,157 biologically unrelated, healthy males recruited in Silesia (Southern Poland). YMCA 2 - an extension of YMCA 1- recruited an additional sample of unrelated 597 young men in Silesia (2.1.4.1).

A total of 1113 samples from the YMCA 1 cohort and 524 from YMCA 2 cohort were successfully typed for DEFB CN using PRT and CN was deduced using both the ML approach and raw PRT calculations. The diploid CN of DEFB varied from 1 to 8 for YMCA 1 and 2 to 8 for YMCA 2, with a modal CN of 4 for both as summarized in Figure 58 below.

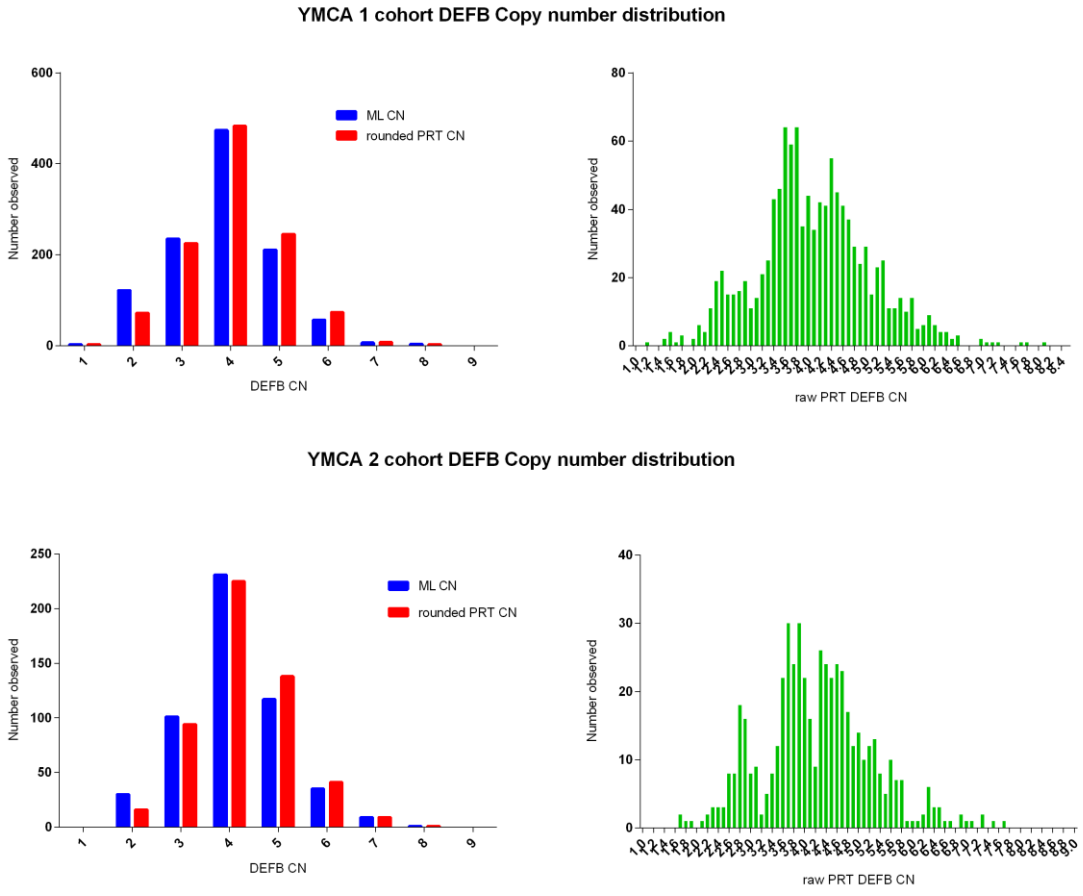


Figure 58: Distribution of YMCA cohorts DEFB CN as estimated by ML approach, rounded PRT and raw PRT.

7.2.2.2 Silesian Cardiovascular Study (SCS) cohort

The SCS is a cohort of 213 Polish families and 435 singletons recruited through probands with high cardiovascular risk (history of hypertension, coronary artery disease, and/or multiple cardiovascular risk factors), as previously described (Tomaszewski et al., 2009).

1054 samples from the SCS cohort were successfully typed for DEFB CN using PRT and CN was deduced using both the ML approach and raw PRT calculations. The diploid CN of DEFB varied from 1 to 9 with a modal CN of 4 as demonstrated in Figure 59 below

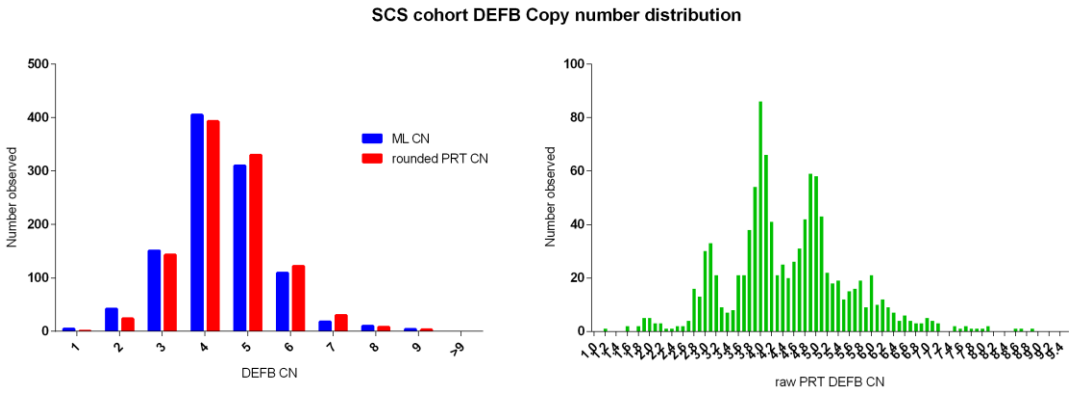


Figure 59: Distribution of SCS cohort DEFB CN as estimated by ML approach, rounded PRT and raw PRT.

7.2.3 Association studies in hypertension cohorts.

A total of 1113 samples from the YMCA 1 cohort and a total of 524 samples from the YMCA 2 cohort were analysed. As for the SCS cohort, a total of 732 unrelated samples had the required clinical information and were used for the analysis.

In all cohorts, a subset of men was on antihypertensive treatment in this study, their systolic and diastolic blood pressure needed to be corrected for blood pressure lowering effect of the antihypertensive treatment. This was done by adding a constant to their measured BP - 15 mmHg for SBP and 10 mmHg for DBP as suggested by Maciej Tomaszewski (personal communication, 03/02/2015) and previously used in (Tobin et al., 2005).

For YMCA 1 and YMCA 2, tests were conducted using Bonferroni correction where the p-value was adjusted to 0.025 per test (0.05/2).

The descriptive statistics for the analysed outcomes, factors and covariates for YMCA 1, YMCA 2 and SCS cohorts are summarized in Table 30 below

Table 30: The descriptive statistics for the analysed outcomes, factors and covariates for YMCA 1, YMCA 2 and SCS cohorts

		Age	BMI	SBP	DBP	Raw PRT	Rounded PRT	ML CN
YMCA 1	Min.	16	16	90	48.3	1.18	1	1
	Max.	44	36	200	128.3	8	8	8
	Mean	19	22.86	118.62	74.56	4.04	4.05	3.89
YMCA 2	Min.	16	16	85.10	56	1.63	2	2
	Max.	42	33	167.20	107.3	7.62	8	8
	Mean	19	22.61	119.34	74.69	4.19	4.24	4.11
SCS	Min.	18	16	80	45	1.10	1	1
	Max.	89	47	220	129	8.50	9	9
	Mean	55	27.48	140.41	81.10	4.46	4.49	4.38

7.2.3.1 Systolic Blood Pressure

The dependent variable; systolic blood pressure (Figure 60), was modelled as a normal distribution with an identity link function. The model used type III sum of squares ANOVA and goodness-of-fit was analysed using Wald statistics. Predictor variables and factors were; age (in years), body mass index (BMI), DEFB CN as estimated by ML approach, raw PRT CN and rounded CN as covariates for the YMCA cohorts whereas sex was added as a cofactor for the SCS cohort to correct for the fact that the cohort has both male and female participants.

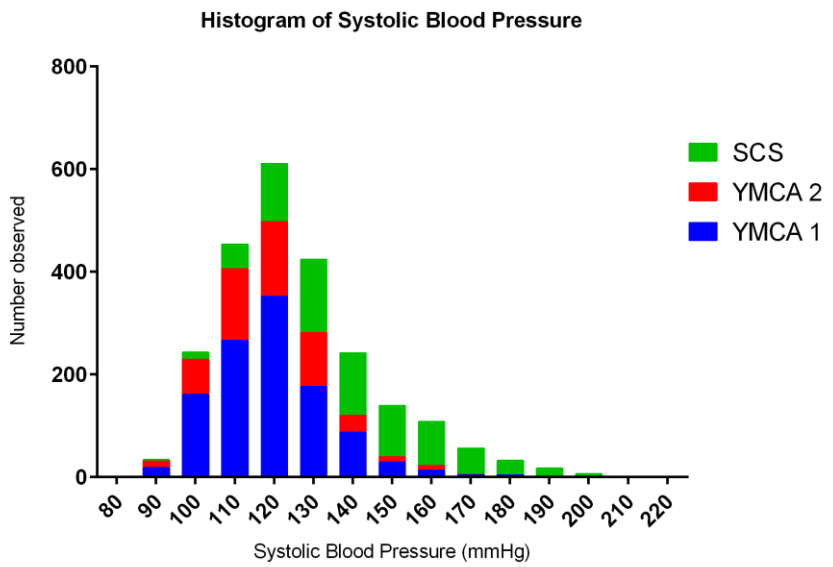


Figure 60: Histogram of the Systolic blood pressure for the YMCA 1, YMCA 2 and SCS cohort.

The results of the associations are presented in Table 31, Table 32 andTable 33.

Table 31: Results of the association between systolic blood pressure in YMCA 1 cohort patients and the predictors; age, BMI and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Bonferroni corrected probability for association. (*) denotes a significant p-value < 0.025 under Bonferroni correction

Outcome: Systolic blood pressure for YMCA 1				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value [a]**
		Lower	Upper	
Age	0.440	0.204	0.677	$2.6x10^{-4}$*
BMI	1.551	1.282	1.819	0.000*
DEFB raw-PRT CN	0.270	-0.533	1.073	0.509
Age	0.440	0.204	0.677	$2.6x10^{-4}$*
BMI	1.551	1.282	1.819	0.000*
DEFB rounded CN	0.270	-0.533	1.073	0.509
Age	0.439	0.203	0.676	$2.6x10^{-4}$*
BMI	1.548	1.281	1.816	<0.001*
DEFB ML CN	0.361	-0.368	1.090	0.332

Table 32: Results of the association between systolic blood pressure in YMCA 2 cohort patients and the predictors; age, BMI and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Bonferroni corrected probability for association. (*) denotes a significant p-value < 0.025 under Bonferroni correction

Outcome: Systolic blood pressure for YMCA 2				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value [a]**
		Lower	Upper	
Age	0.142	-0.213	0.497	0.433
BMI	0.979	0.577	1.380	$2x10^{-6}$*
DEFB raw-PRT CN	-1.608	-2.766	-0.450	0.007*
Age	0.133	-0.222	0.488	0.463
BMI	0.975	0.574	1.376	$2x10^{-6}$*
DEFB rounded CN	-1.721	-2.881	-0.561	0.004*
Age	0.172	-0.182	0.527	0.341
BMI	0.972	0.570	1.374	$2x10^{-6}$*
DEFB ML CN	-1.143	-2.259	-0.028	0.045

Table 33: Results of the association between systolic blood pressure in SCS cohort patients and the predictors; sex, age, BMI and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Set to zero because this category is the reference for the specific parameter (*) denotes a significant p-value < 0.05

Outcome: Systolic blood pressure for SCS				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Sex: Male	-4.276	-7.211	-1.341	0.004*
Sex: Female	0[a]			
Age	0.468	0.343	0.594	$2.7x10^{-13}$*
BMI	1.189	0.841	1.538	$2.2x10^{-11}$*
DEFB raw-PRT CN	0.090	-1.226	1.406	0.894
Sex: Male	-4.276	-7.211	-1.342	0.004*
Sex: Female	0[a]			
Age	0.468	0.343	0.594	$2.7x10^{-13}$*
BMI	1.189	0.841	1.538	$2.2x10^{-11}$*
DEFB rounded CN	0.050	-1.242	1.342	0.940
Sex: Male	-4.277	-7.212	-1.343	0.004*
Sex: Female	0[a]			
Age	0.468	0.343	0.594	$2.7x10^{-13}$*
BMI	1.189	0.841	1.537	$2.2x10^{-11}$*
DEFB ML CN	0.061	-1.183	1.305	0.924

7.2.3.2 Diastolic Blood Pressure

The second dependent variable to be investigated; diastolic blood pressure (Figure 61), was modelled as a normal distribution with an identity link function. The model used type III sum of squares ANOVA and goodness-of-fit was analysed using Wald statistics. Predictor variables and factors were; age (in years), body mass index (BMI), DEFB CN as estimated by ML approach, raw PRT CN and rounded CN as covariates for the YMCA cohorts whereas sex was added as a cofactor for the SCS cohort to correct for the fact that the cohort has both male and female participants.

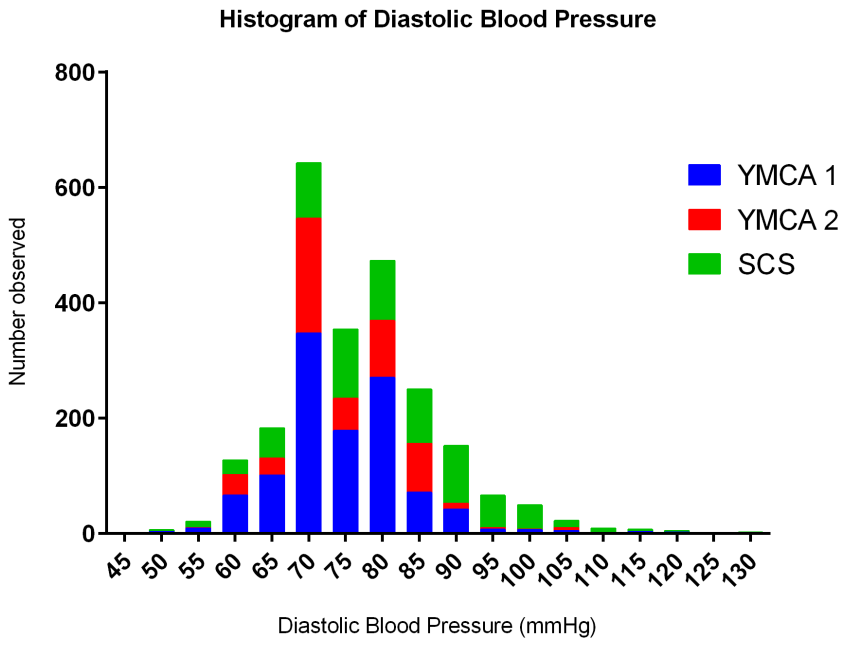


Figure 61: Histogram of the Diastolic blood pressure for the YMCA 1, YMCA 2 and SCS cohort.

The results of which are summarized in the Table 34, Table 35 and Table 36 below.

Table 34: Results of the association between diastolic blood pressure in YMCA 1 cohort patients and the predictors; age, BMI and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Bonferroni corrected probability for association. (*) denotes a significant p-value < 0.025 under Bonferroni correction

Outcome: Diastolic blood pressure for YMCA 1				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value [a]**
		Lower	Upper	
Age	0.283	0.136	0.431	$1.7x10^{-4}$*
BMI	0.704	0.537	0.871	$1.1x10^{-16}$*
DEFB raw-PRT CN	0.052	-0.448	0.552	0.839
Age	0.283	0.136	0.431	$1.7x10^{-4}$*
BMI	0.704	0.537	0.871	$1.1x10^{-16}$*
DEFB rounded CN	0.052	-0.448	0.552	0.839
Age	0.281	0.134	0.428	$1.8x10^{-4}$*
BMI	0.705	0.538	0.872	$1.1x10^{-16}$*
DEFB ML CN	0.166	-0.288	0.620	0.473

Table 35: Results of the association between diastolic blood pressure in YMCA 2 cohort patients and the predictors; age, BMI and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Bonferroni corrected probability for association. (*) denotes a significant p-value < 0.025 under Bonferroni correction

Outcome: Diastolic blood pressure for YMCA 2				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value [a]**
		Lower	Upper	
Age	0.055	-0.158	0.268	0.614
BMI	0.537	0.296	0.778	$1.2x10^{-5}$*
DEFB raw-PRT CN	-0.679	-1.375	0.016	0.056
Age	0.052	-0.161	0.265	0.632
BMI	0.536	0.295	0.776	$1.3X10^{-5}$*
DEFB rounded CN	-0.708	-1.405	-0.010	0.047
Age	0.065	-0.147	0.278	0.546
BMI	0.534	0.293	0.776	$1.4X10^{-5}$*
DEFB ML CN	-0.534	-1.202	0.135	0.118

Table 36: Results of the association between diastolic blood pressure in SCS cohort patients and the predictors; sex, age, BMI and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Set to zero because this category is the reference for the specific parameter (*) denotes a significant p-value < 0.05

Outcome: Diastolic blood pressure for SCS				
Parameter	B	95% Wald Confidence Interval		p-value
		Lower	Upper	
Sex: Male	0.324	-1.456	2.105	0.721
Sex: Female	0[a]			
Age	0.105	0.028	0.181	0.007*
BMI	0.600	0.389	0.811	$2.6X10^{-8}$*
DEFB raw-PRT CN	0.278	-0.520	1.076	0.495
Sex: Male	0.324	-1.457	2.104	0.721
Sex: Female	0[a]			
Age	0.105	0.029	0.181	0.007*
BMI	0.600	0.389	0.812	$2.6X10^{-8}$*
DEFB rounded CN	0.238	-0.546	1.022	0.551
Sex: Male	0.320	-1.461	2.100	0.725
Sex: Female	0[a]			
Age	0.105	0.028	0.181	0.007*
BMI	0.598	0.387	0.809	$2.9X10^{-8}$*
DEFB ML CN	0.194	-0.561	0.948	0.615

7.2.4 Discussion

The present study was designed to test the hypothesis that CNV at the DEFB region might be associated with hypertension. Looking at the statistical results of the YMCA 1 cohort, it can be deduced that SBP is affected by age and BMI; the older a person is and the larger their BMI is, the higher their SBP is likely to be by 0.4 and 1.5 mmHg respectively. As for the second outcome, as age and BMI increases, DBP is likely to increase by 0.3 and 0.8 mm/Hg respectively.

The same results were replicated among the SCS cohort. This cohort also shows that a high SBP value is 4 times less likely in males compared to females but had no association in DBP, whereas age and BMI affect both SBP and DBP but by varying magnitudes.

As for the YMCA 2 cohort, the results indicate that age plays no role in SBP and DBP but confirms that BMI does in the same direction as both YMCA 1 and SCS cohorts. The YMCA 2 cohort shows that the lower the DEFB CN is, the higher the SBP is and the result for DBP was significant only when tested with DEFB CN as estimated by raw PRT and rounded CN, and not with DEFB CN as calculated by ML approach. Since the results were not shown in both YMCA 1 and SCS cohorts which both have a higher number of participants, it is worth explaining that a p-value measures whether an observed result can be attributed to chance. Essentially the p-value can summarize the data assuming a specific null hypothesis. It cannot work backwards and make statements about the underlying reality because the odds that a real effect was there in the first place need to be known and in accordance to Goodman (2001), a p-value of 0.01 corresponds to a false-alarm probability of at least 11%, depending on the underlying probability that there is a true effect; a p-value of 0.05 raises that chance to at least 29%.

The fact that DEFB CN as calculated by raw PRT CN and rounded CN followed the same pattern, suggests this might be an artefact of CNV calling. Bonferroni correction was carried out on YMCA 1 and YMCA 2 to reduce the chances of obtaining false-positive results (type I errors), this combined with the explanation about p-value interpretation above could simply mean the results of significance in YMCA 2 cohort of SBP with DEFB CN as calculated by raw and rounded PRT could be explained by chance.

7.3 Obesity and metabolic syndrome

Obesity has emerged as one of the leading public-health issues in the past decades. According to the International Obesity Taskforce 1.1 billion people are overweight with a body mass index (BMI) over 25 kg/m^2 and 312 million are classified as obese (BMI > 30 kg/m^2) (Haslam & James, 2005). Though obesity itself is not a disease per se, it is a major risk factor for developing type II diabetes, cardiovascular disease and certain types of cancer at later ages (Khan, 2014; Pothiwala et al., 2009; Van Gaal et al., 2006).

The International Diabetes Federation published the definition of the metabolic syndrome describing a cluster of factors associated with an increased risk for atherosclerotic cardiovascular disease (CVD) and diabetes. For a person to be diagnosed with the metabolic syndrome the following criteria have been defined: central obesity measured by waist circumference plus two additional factors such as raised triglycerides (>150 mg/dl), raised blood pressure (130 mm Hg systolic or >85 mmHg diastolic) or raised fasting plasma glucose (>100mg/dl) (Grundy et al., 2004).

7.3.1 Study rationale

The causes of obesity are diverse. Although modern sedentary life style, and an unlimited offer of food, supports metabolic diseases, it cannot be neglected that genetic predisposition also plays a major role (O'Rahilly, 2009). Several genetic causes for obesity have been described, for example mutations in leptin (Montague et al., 1997) and the leptin receptor (Clement et al., 1998) or in a most recent research, DEFB knockout mice (knocked out for several of the genes orthologous to those in the human CNV) were found to have a distinct obesity phenotype (Dorin, unpublished), and DEFB knock-in mice studies implied that humans with higher copy number of DEFB tend to be leaner (Dorin, unpublished).

Because of Dorin's preliminary findings, we conducted an association study to investigate if CNV at the DEFB region in humans had an effect on BMI. The study was carried out on both YMCA cohorts and SCS cohort as they had the required clinical data.

7.3.2 Association studies of obesity

The descriptive statistics for the analysed outcomes, factors and covariates for YMCA 1, YMCA 2 and SCS cohorts are summarized in Table 37 below.

Table 37: descriptive statistics for the analysed outcomes, factors and covariates for YMCA 1, YMCA 2 and SCS cohorts

		Age	BMI	Log(TC:HDL)	Raw PRT	Rounded PRT	ML CN
YMCA 1 (n=1112)	Min.	16	16	0.36	1.18	1	1
	Max.	44	36	2.77	8	8	8
	Mean	19	22.86	1.33	4.04	4.05	3.89
YMCA 2 (n=524)	Min.	16	16	.64	1.63	2	2
	Max.	42	33	2.15	7.62	8	8
	Mean	19	22.61	1.24	4.19	4.24	4.11
SCS (n=741)	Min.	18	16	.36	1.10	1	1
	Max.	89	47	3.37	8.50	9	9
	Mean	55	27.48	1.70	4.46	4.49	4.38

7.3.2.1 Body mass index

The dependent variable; BMI (Figure 62), was modelled as a normal distribution with an identity link function. The model used type III sum of squares ANOVA and goodness-of-fit was analysed using Wald statistics. Predictor variables and factors were; age (in years), DEFB CN as estimated by ML approach, raw PRT CN and rounded CN as covariates for the YMCA cohorts whereas sex was added as a cofactor for the SCS cohort to correct for the fact that the cohort has both male and female participants.

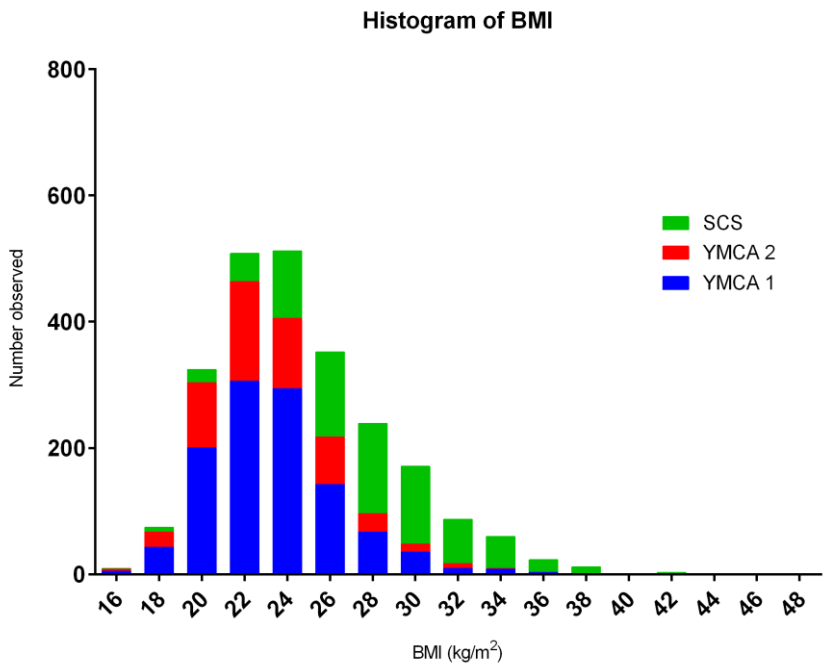


Figure 62: Histogram of the BMI for the YMCA 1, YMCA 2 and SCS cohort.

The results of which are summarized in the Table 38, Table 39 and Table 40 below.

Table 38: Results of the association between BMI in YMCA 1 cohort patients and the predictors; age, and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Bonferroni corrected probability for association. (*) denotes a significant p-value < 0.025 under Bonferroni correction

Outcome: BMI for YMCA 1				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value [a]**
		Lower	Upper	
Age	0.236	0.186	0.286	<0.001*
DEFB raw-PRT CN	-0.198	-0.374	-0.023	0.027
Age	0.236	0.186	0.286	<0.001*
DEFB rounded CN	-0.198	-0.374	-0.023	0.027
Age	0.234	0.184	0.284	<0.001*
DEFB ML CN	-0.072	-0.232	0.089	0.380

Table 39: Results of the association between BMI in YMCA 2 cohort patients and the predictors; age, and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Bonferroni corrected probability for association. (*) denotes a significant p-value < 0.025 under Bonferroni correction

Outcome: BMI for YMCA 2				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value [a]**
		Lower	Upper	
Age	0.262	0.190	0.335	$1.2x10^{-12*}$
DEFB raw-PRT CN	0.032	-0.215	0.279	0.801
Age	0.262	0.190	0.335	$1.2x10^{-12*}$
DEFB rounded CN	0.032	-0.215	0.279	0.801
Age	0.260	0.188	0.333	$1.4x10^{-12}$*
DEFB ML CN	-0.003	-0.240	0.234	0.981

Table 40: Results of the association between BMI in SCS cohort patients and the predictors; sex, age, and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Set to zero because this category is the reference for the specific parameter (*) denotes a significant p-value < 0.05

Outcome: BMI for SCS				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Sex: Male	0.190	-0.422	0.802	0.542
Sex: Female	0[a]			
Age	0.053	0.027	0.079	$6.4x10^{-5}$*
DEFB raw-PRT CN	-0.134	-0.407	0.139	0.336
Sex: Male	0.190	-0.422	0.802	0.542
Sex: Female	0[a]			
Age	0.053	0.027	0.079	$6.6x10^{-5}$*
DEFB rounded CN	-0.144	-0.412	0.125	0.294
Sex: Male	0.194	-0.419	0.806	0.536
Sex: Female	0[a]			
Age	0.053	0.027	0.079	$6.2x10^{-5}$*
DEFB ML CN	-0.025	-0.284	0.234	0.851

The same statistical test was repeated but this time only people aged 23 and below were in the inclusion criterion for the YMCA cohorts and age was no longer set as a cofactor. As for the SCS cohort, the association was carried out on each sex alone

therefore sex was no longer set as a cofactor. Hence the total number of participants became as summarized below.

Cohort	Participants	Sex
YMCA 1	1068	Male
YMCA 2	498	Male
SCS	432	Male
	309	Female

The results of BMI and DEFB CN as calculated by ML, raw PRT and rounded PRT showed no association across the cohorts.

In order to minimize association speculations of DEFB CN and BMI in YMCA 2 and SCS cohorts, power analysis was carried out using simulation was based on Bolker, (2006). Power is defined as the probability of detecting a "true" effect, when the effect exists. Most recommendations for power fall between 80% and 90%.

Power was estimated by a simulated study based of equivalent size to the YMCA 1 cohort (n=1112) with a similar distribution of DEFB CN to that observed in YMCA 1. For the current sample size of YMCA 2, the power to detect an effect of BMI change due to DEFB CN was 34% whereas it was 49% for SCS cohort. Both results are below the recommended minimum of experimental power of 80%.

7.3.2.2 Total cholesterol to high-density lipoprotein ratio (TC: HDL)

It has been demonstrated by cross-sectional and longitudinal studies that among adults, blood cholesterol concentration increases with age (Verschuren et al., 1994; The Lipid Research Clinics Program Epidemiology Committee, 1979). All available results also suggest that the increase in total cholesterol with age is different for men and women (Wenger, 1999). In addition, there is a general consensus that the total cholesterol level rises as the body mass index rises (Alexander, 2001; Brown et al., 2000).

In light of this information, the second outcome to be studied that shows a good association with obesity is log transformed values of the total cholesterol to High-Density Lipoprotein ratio (LogTC:HDL)(Figure 63).

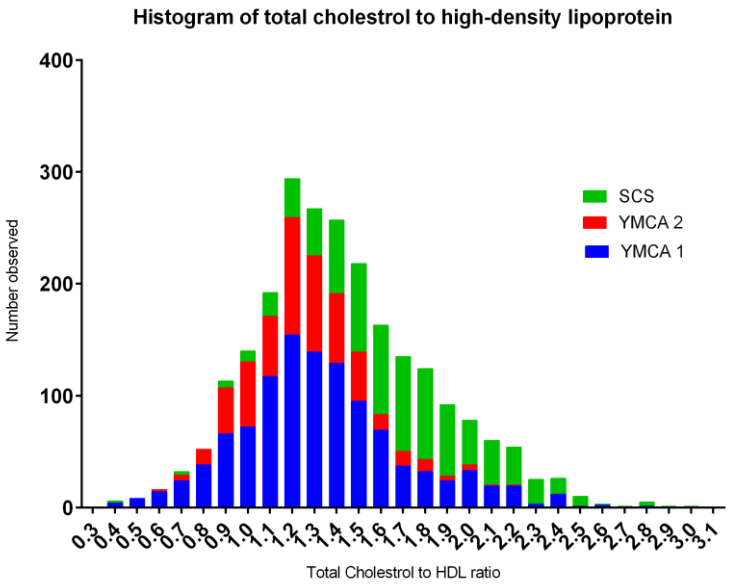


Figure 63: Histogram of the log transformed values of TC: HDL ratio for the YMCA 1, YMCA 2 and SCS cohort.

Here, the dependent variable was modelled as a normal distribution with an identity link function. The model used type III sum of squares ANOVA and goodness-of-fit was analysed using Wald statistics. Predictor variables and factors were; age (in years), DEFB CN as estimated by ML approach, raw PRT CN and rounded CN as covariates for the YMCA cohorts whereas sex was added as a cofactor for the SCS cohort to correct for the fact that the cohort has both male and female participants. The results are summarized in Table 41, Table 42 and 43 below.

Table 41: Results of the association between log(TC:HDL) in YMCA 1 cohort patients and the predictors; age, and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Bonferroni corrected probability for association. (*) denotes a significant p-value < 0.025 under Bonferroni correction

Outcome: Log(TC:HDL) for YMCA 1				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value [a]**
		Lower	Upper	
Age	0.009	0.003	0.016	0.004*
DEFB raw-PRT CN	-0.024	-0.046	-0.001	0.039
Age	0.009	.003	0.016	0.005*
DEFB rounded CN	-0.022	-0.044	-0.001	0.038
Age	0.009	0.003	0.016	0.005*
DEFB ML CN	-0.021	-0.041	-0.001	0.043

For YMCA 1, a slight positive association exists between age and TC: HDL and no association exists between DEFB CN and TC: HDL under Bonferroni corrected p-value cut-off.

Table 42: Results of the association between log(TC:HDL) in YMCA 2 cohort patients and the predictors; age, and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Bonferroni corrected probability for association. (*) denotes a significant p-value < 0.025 under Bonferroni correction

Outcome: Log(TC:HDL) for YMCA 2				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value [a]**
		Lower	Upper	
Age	0.026	0.020	0.032	<0.001*
DEFB raw-PRT CN	0.015	-0.006	0.035	0.154
Age	0.026	0.020	0.032	<0.001*
DEFB rounded CN	0.013	-0.007	0.034	0.202
Age	0.026	0.020	0.032	<0.001*
DEFB ML CN	0.015	-0.005	0.034	0.147

YMCA 2 cohort shows an association with age; however no association exists with DEFB CN.

Table 43: Results of the association between TC: HDL ratio in SCS cohort patients and the predictors; sex, age, and DEFB raw-PRT CN, DEFB rounded CN and DEFB ML CN respectively. ([a]) Set to zero because this category is the reference for the specific parameter (*) denotes a significant p-value < 0.05.

Outcome: Log(TC:HDL) for SCS				
Parameter	**B**	**95% Wald Confidence Interval**		**p-value**
		Lower	Upper	
Sex: Female	-0.107	-0.160	-0.053	$9x10^{-5}$*
Sex: Male	0[a]			
Age	$-2.6x10^{-4}$	-0.003	0.002	0.820
DEFB raw-PRT CN	0.007	-0.017	0.031	0.547
Sex: Female	-0.107	-0.160	-0.053	$9x10^{-5}$*
Sex: Male	0[a]			
Age	$-2.6x10^{-4}$	-0.003	0.002	0.823
DEFB rounded CN	0.009	-0.015	0.032	0.459
Sex: Female	-0.107	-0.160	-0.053	$9.2x10^{-5}$*
Sex: Male	0[a]			
Age	$-2.6x10^{-4}$	-0.003	0.002	0.821
DEFB ML CN	0.009	-0.013	0.032	0.428

In the SCS cohort, a negative association exists between being female and TC: HDL. Neither age nor DEFB CN shows an association.

Power was estimated by a simulated study based of equivalent size to the YMCA 1 cohort (n=1112) with a similar distribution of DEFB CN to that observed in YMCA 1. For the current sample size of YMCA 2, the power to detect an effect of log(TC: HDL) change due to DEFB CN was 9% whereas it was 10% for SCS cohort. Both results are below the recommended minimum of experimental power of 80%.

7.3.3 Discussion

According to the above results of YMCA 1, DEFB CN was not found to have an effect on BMI under Bonferroni correction (p-value cut off of 0.025) however, assuming no multiple testing correction was carried out, DEFB CN was found to have an effect on BMI in the same direction as that indicated by Dorin's preliminary results. YMCA 2 under Bonferroni correction and without it and SCS cohorts showed no association

with BMI. This could be due to a low sample size (524 and 741 respectively) with power too small to pick up a real effect in the population. The same trend was found when the test was carried out using only participants below the age of 24. For YMCA 2 and SCS cohorts to pick up a real effect; the minimum sample size has to be 1624.

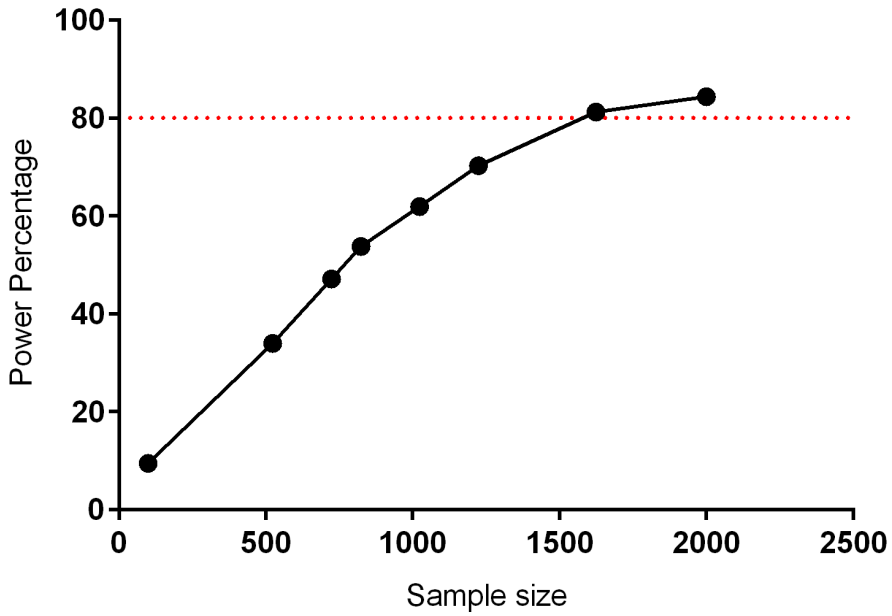


Figure 64: a graph showing the power of the study for each sample size to pick up an effect in BMI change due to DEFB CN based on the YMCA 1 parameters. The red line set at 80% shows the minimum accepted power for a study to pick up a true effect if an effect is present in a population.

As for the log transformed values for total cholesterol to HDL ratio, the higher the ratio was, the lower the DEFB CN would have been in YMCA 1 cohort if no Bonferroni correction was applied. Under the Bonferroni correction, there was no association between DEFB CN and log transformed values for total cholesterol to HDL ratio. The effect was not detected in YMCA 2 (with and without Bonferroni correction) or SCS cohort as the power of the study was 32% and 44% respectively. For YMCA 2 and SCS cohorts to pick up a real effect; the minimum sample size has to be 1800.

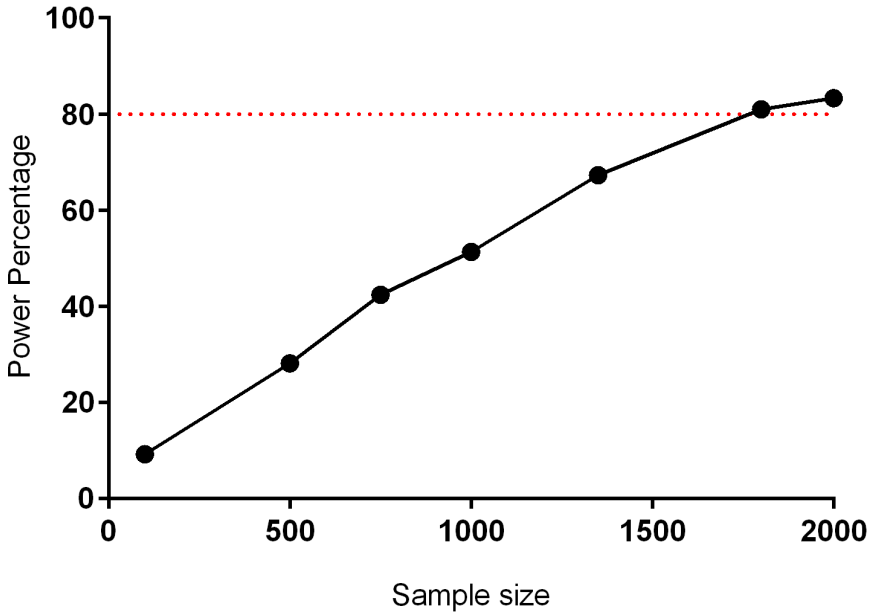


Figure 65: a graph showing the power of the study for each sample size to pick up an effect in log transformed values of TC: HDL ratio change due to DEFB CN based on the YMCA 1 parameters. The red line set at 80% shows the minimum accepted power for a study to pick up a true effect if an effect is present in a population.

In order to conclude if DEFB CN has an effect on BMI and log (TC: HDL), a replication study true to the discovery study; YMCA 1 cohort, not under Bonferroni correction, needs to be carried out, with a minimum sample size of 1800. The criteria for recruitment should be only males, aged 17 years and above, and of Eastern European origin.

8 Discussion

My PhD thesis explored the structural variability of the β-defensin copy number variable region shedding light on the size of the repeat, its relationship to the polymorphic inversion at the 8p23.1 genomic location and whether SNPs and/or SNP haplotypes were in LD with the region. The thesis also compared and contrasted the different methods used for calling DEFB CNVs and highlighted the best method to use in genotyping copy number for the DEFB region. The role of copy number variations of DEFB in various diseases was also explored using several case-control studies and an attempt to develop a model system to investigate whether DEFB expression levels differ with CN in response to treatment with pneumolysin by using Normal Human Bronchial Epithelial (NHBE) cells was carried out.

8.1 Size of the contiguous DEFB CNV region is 322Kb.

Using NimbleGen aCGH data we confirm that heterogeneity at the DEFB region was limited to the presence/absence polymorphism of an endogenous retrovirus, HERV-k115 in the intron of *DEFB107* confirming the results of Turner et al., (2001), however contradicting its prevalence of 16%. Our data showed a loss in *DEFB107* in one of the regions in 30% of the samples. NimbleGen aCGH data was also used to deduce that the size of the contiguous DEFB CNV region to be 322Kb with two copies assembled in hg18, agreeing with the findings of Hollox et al., (2003), with gene content of CNV confirmed by Forni et al., (2015).

The identification of the CNV boundary in humans has two important consequences. First, identification of the boundary within an LTR as part of a large antiparallel segmentally duplicated region suggests a link to the frequent polymorphic inversion at 8p23.1, where the 5-Mb region between REPP and REPD is inverted. This is because inversions are sponsored by recurrent non-allelic homologous recombination between antiparallel repeats. Analysis of potentially recombinant BAC sequences from this region has previously suggested multiple inversion breakpoints, particularly centred on the olfactory repeat region, but the closest is at least 50kb away from the DEFB CNV

boundaries (Salm et al. 2012). So although it looks very likely that the entire DEFB copy number region was repeatedly carried between REPP and REPD by inversion formation (Abu Bakar et al. 2009; Salm et al. 2012), our findings confirm no correlation was found between DEFB CN and inversion status. The second important consequence of the identification of the CNV boundary is that non defensin genes within the repeat unit that are confirmed as copy number variable may show alteration of expression levels concomitant with the CNV.

SNP haplotype analysis showed that the TAG and GGA haplotype show significant weak correlation with DEFB CN. The presence of the TAG haplotype is associated with an increase in CN and presence of the GGA haplotype is associated with a decrease in CN. These results could perhaps be best explained by the differences in mutation rates. Slowest, representing the oldest variation are SNPs and their haplotypes, followed by the inversion which is recurrent and leaves a faint yet detectable signal of association with the SNP haplotype, and finally there is the CNV, which mutates too quickly to leave signal with a flanking SNP allele and seemingly too quickly to leave a signal with the inversion. It also sheds light on the fact that diseases, like Psoriasis, are associated with DEFB CNV regardless of their orientation and inversion status.

8.2 PRT is the best method to genotype DEFB CNV.

Agilent aCGH and NimbleGen aCGH first developed CGH as a method for copy number comparison between differentially labelled target and reference DNAs by measuring the fluorescence ratio along the length of each chromosomal region, indicating relative losses or gains in a target sample using fluorescence in situ hybridization. CGH arrays use arrays of bacterial artificial chromosomes, cDNA, or long synthetic oligonucleotides to probe specific regions of interest for copy number assessment (Li & Olivier, 2013). These probes can either be selected for an even genome-wide distribution or for unique, repetitive regions. NimbleGen aCGH and Agilent aCGH platforms contain substantial amount of probes in SD regions. In general, SDs show higher levels of false positives and false negative call rates in comparison to unique regions of the genome (Pinto et al., 2011). The success for resolving copy number polymorphism (CNP) into

discrete copy number classes depends on the genetic property of the CNP, the density and performance of the probe and the parameters used for normalization. It is worth mentioning that probe number does not mean improved coverage or resolution (Cantsilieris et al., 2013). When it comes to DEFB CN calling, NimbleGen aCGH has 15,266 probes across the DEFB region and Agilent aCGH has 62, and none provide as good a clustering around an integer number call as PRT does. A major drawback of using NimbleGen aCGH and Agilent aCGH in calling DEFB CN is that estimation of the copy number call is relative and depends on how well the reference sample is characterised. For example, a loss in the reference sample can be interpreted as a gain in the test sample even though the test sample may have a diploid CN of two (Cantsilieris et al., 2013). As far as the nCounter is concerned, the ability to call absolute DEFN CN is very specific especially among the lower copy number calls. We can conclude that nCounter and PRT are equally good in calling copy number, however when looking at one gene at a time, nCounter is not feasible as it is capable of profiling up to 800 regions of the human genome in a single reaction making its running cost high. The DEFB CN calls as typed by ddPCR were accurate and reproducible when run in duplicate and the threshold of quadruplets was tweaked manually. This step is time consuming since each sample has to be looked at individually to set the threshold at the right position according to the user, hence typing large cohorts becomes prohibitively time-consuming. When a copy number is high for example, ddPCR would require an extra step in sample preparation such as restriction enzymes to efficiently separate linked copies of the target gene. As for genome STRiP method, the results when plotted against PRT had a regression coefficient of 0.98 and almost 95% identity. This proved to be a good calling method for DEFB CN however; it requires sequenced genomes to generate the data which is not always possible due to ethical and financial considerations.

In summary, triplex PRT proved to be the best method for typing DEFB CN. PRT uses a small amount of DNA of 5ng/µl, uses one pair of primers to amplify test and reference which increases consistency and reproducibility. During this PhD, 7173 samples were genotyped for DEFB CNV using PRT, 443 failed to type from the first run. i.e. PRT has a high pass rate of almost 94% in returning an integer value, with a significant p-value

from the first run, rapid and suitable for typing thousands of samples in large case-control association studies.

8.3 DEFB CN and lung function.

8.3.1 NHBE cell lines having different DEFB CN have different basal DEFB mRNA levels and treatment with pneumolysin decrease hBD-2 expression.

The original experiment was carried out on an NHBE cell line (cell line 1) that carried a DEFB CN of 4. The replication experiment (cell line 2) was carried out on an NHBE cell line that had a DEFB CN of 5. qPCR results show that basal mRNA levels for cell line 2 were higher than those in cell line 1 suggesting that DEFB CN has an effect on the amount of mRNA stored *DEFB4*. Nonetheless, in response to treatment with Pn, the fold change in expression of *DEFB4* relative to reference *PPIA* and *UBC* genes in both cell lines were similar; suggesting that amount of *DEFB4* mRNA released upon infection is not controlled by the DEFB CN.

As far as hBD-2 concentration levels are concerned, treatment with Pn reduced the amount of hBD-2 in the supernatant. These results contradict previous findings of Kim et al., (2013) and Scharf et al., (2012). The contradicting results could be explained by the use of different cell lines in each of the experiments; NHBE cell lines in ours, Human alveolar epithelial A549 cells in Kim's experiment and Primary human small airway epithelial cells in Scharf's experiment. It has been previously confirmed by Hellmann et al., (2001) that A549 cell lines, being carcinoma cell lines had expression differences across 17 genes expression differences, either upregulation or downregulation relative to NHBE cells. Hence, the results obtained from carcinoma cell lines do not necessarily have the same effect on normal human bronchial epithelial cells and cannot mimic *in vivo* studies.

hBD-2 concentration levels were not significantly different in cells that have a DEFB CN of 4 and 5, but hBD-2 concentration levels decreased as levels of treatment with Pn increased in each cell line (Figure 41 and Figure 44). The explanation behind why

mRNA levels showed no change upon treatment with Pn could suggest that hBD-2 is released at sublytic (pro-inflammatory) concentrations of Pn as an early indicator that the epithelial cells are stressed and return to basal levels. Therefore the reduction in hBD-2 levels may be an acute response to treatment with Pn.

8.4 No evidence of association of DEFB CN with HIV viral load.

DEFB CN was determined in a cohort of 302 people with African descent and in a separate cohort of 3155 people with European descent using PRT. The distributions of CN observed in each cohort ranged from 1 – 9, with a modal CN of 4 consistent with previous data (Hardwick et al., 2012). Association with log(spVL) was tested using a generalised linear model, with sex, age, and the first three principal components of genome-wide SNP genotype data as covariates. No association with DEFB CN was found in the IAVI cohort (b=0.007, 95%CI -0.064 to 0.077, p=0.853) nor the Swiss cohort (b=-0.02, 95%CI -0.06 to 0.021, p=0.335).

Following the footsteps of previous analyses, the association of VL with DEFB CN by dividing the copy number distribution, ranging from 1 to 9, into two discrete categories, 4 or more copies and fewer than 4 copies was carried out. This has the potential to increase power, as a linear response CN is not assumed. However, using the same covariates as above, neither the SHC cohort nor the IAVI cohort showed an association (reference category copy number <4, b=-0.015, 95%CI -0.117 to 0.087, p=0.773 for the SHC cohort, b=-0.22, 95%CI -0.567 to 0.127, p=0.214 for the IAVI cohort).

Another clinical variable which shows evidence of association with host genetic variation is the rate of progression of HIV from seroconversion to death due to AIDS. Progression data for 229 individuals from the Swiss cohort were available. Using a Cox regression model, we found no statistically significant association with DEFB CN (Exp(B)=1.146, p=0.081), although the direction of effect suggested an increase in the hazard ratio with increasing DEFB CN. Given that this is consistent with previous results (Hardwick *et al.*, 2012), and that rate of progression analyses the effect of host variation across the whole timescale of HIV infection including the later stages of AIDS,

we suggest that this result should be examined further with a large cohort where progression data is available.

We finally investigated whether there was evidence of association of DEFB CN and risk of acquiring HIV by constructing a case-control study in which the cases are from the Swiss cohort and the controls are individuals of European descent of unknown HIV status previously typed as part of other studies. For controls, we used 1156 individuals from a population cohort from Nottingham, UK and 695 individuals from Leicester, UK (Wain *et al.*, 2014), combined with 183 UK individuals from the ECACC Human Random Controls cohort (Hardwick *et al.*, 2011). These individuals were of unknown HIV status. Using logistic regression with case/control as the binary outcome variable, we found no association with DEFB CN (b=0.009, 95%CI -0.042 to 0.061, p=0.725).

In summary, no evidence was found for association of DEFB CN with HIV susceptibility or spVL. We also find no evidence of a strong effect on HIV progression rate, although it should be noted that the small sample size makes it unlikely that we could detect a small- or medium-sized effect.

8.5 Association of DEFB CN variation in other disease cohorts

8.5.1 No evidence that DEFB CNV plays a role in VUR patients

DEFB CN was determined in a cohort of 414 individuals from the RIVUR study using PRT. The distributions of CN observed in each cohort ranged from 2 – 9, with a modal CN of 4. Association of raw PRT DEFB CN, rounded PRT DEFB CN and ML DEFB CN respectively with number of breakthrough infections, whether first infection caused by *E.coli*, whether breakthrough infection was caused by *E.coli* and whether new kidney scars developed were carried out.

No association of raw PRT DEFB CN (b=0.209, 95%CI -0.032 to 0.451, p=0.089) and rounded PRT DEFB CN (b=0.231, 95%CI -0.005 to 0.468, p=0.055) was found with number of breakthrough infections. However a barely significant association was found with DEFB CN as calculated by ML (b=0.261, 95%CI 0.01 to 0.512, p=0.041). Higher DEFB CN as estimated by ML approach showed an association with a higher

number of breakthrough infections (p= 0.041). However, since the significance level is close to the 0.05 cut-off point and neither DEFB CN as estimated by raw PRT calculations (p=0.089) and rounded PRT CN (p=0.055) were found to be associated with the same dependent variable, a deeper investigation is required to confirm or refute the result. Also, given that the highest p value is given with DEFB CN as calculated using raw PRT data suggests this might be an artefact of CNV calling.

No association of raw PRT DEFB CN (b=-0.034, 95%CI -0.541 to 0.474, p=0.897), rounded PRT DEFB CN (b=-0.124, 95%CI -0.605 to 0.357, p=0.613) and DEFB CN as calculated by ML (b=-0.037, 95%CI -0.544 to 0.469, p=0.886) was found with whether first infection caused by *E.coli.*

No association of raw PRT DEFB CN (b=-0.477, 95%CI -1.33 to 0.375, p=0.273), rounded PRT DEFB CN (b=-0.491, 95%CI -1.35 to 0.365, p=0.261) and DEFB CN as calculated by ML (b=-0.511, 95%CI -1.39 to 0.369, p=0.255) was found with whether the breakthrough infection was caused by *E.coli*

Lastly, No association of raw PRT DEFB CN (b=-0.348, 95%CI -0.913 to 0.217, p=0.228), rounded PRT DEFB CN (b=-0.385, 95%CI -0.943 to 0.172, p=0.175) and DEFB CN as calculated by ML (b=-0.545, 95%CI -1.03 to 0.120, p=0.121) was found with the development of new kidney scars.

In conclusion, this work found no evidence that CNVs at the DEFB region affects susceptibility to VUR or UTIs. A replication study must be carried out with a larger cohort to minimize any CNV calling artefact. Also, the potential interactions between DEFB genes and environmental conditions (VUR, age, ethnicity… etc.) should explore more complex hypotheses, including gene–gene and gene–environment.

8.5.2 No evidence that DEFB CNV affects blood pressure

DEFB CN was determined in three cohorts; YMCA 1 of 1113 individuals, YMCA 2 of 524 individuals and SCS of 741 individuals using PRT. The distributions of CN observed in each cohort ranged from 1 – 8, 2 – 8 and 1 – 9 respectively with a modal CN of 4.

Association of raw PRT DEFB CN, rounded PRT DEFB CN and ML DEFB CN respectively with systolic blood pressure and diastolic blood pressure was carried out. Bonferroni correction was carried out on YMCA 1 and YMCA 2 to reduce the chances of obtaining false-positive results (type I errors).

No association of systolic blood pressure and diastolic blood pressure was found in YMCA 1 and SCS cohorts and DEFB CN. The YMCA 2 cohort shows that the lower the DEFB CN is, the higher the SBP is (for raw PRT; b=-1.61, 95%CI -2.77 to -0.45, p=0.007) and the result for DBP was significant only when tested with DEFB CN as estimated by raw PRT and rounded CN, and not with DEFB CN as calculated by ML approach. The results were not shown in both YMCA 1 and SCS cohorts which both have a higher number of participants. The fact that DEFB CN as calculated by raw PRT CN and rounded CN followed the same pattern, suggests this might be an artefact of CNV calling and could simply mean the results of significance in YMCA 2 cohort of SBP with DEFB CN as calculated by raw and rounded PRT could be explained by chance.

8.5.3 No association of DEFB CNV and obesity

Studies of obesity were carried out using the YMCA 1, YMCA 2 and SCS cohorts as the required descriptive statistics were available for analysis. According to results of YMCA 1, DEFB CN was not found to have an effect on BMI under Bonferroni correction (p-value cut off of 0.025) however, assuming no multiple testing correction was carried out, DEFB CN was found to have an effect on BMI in the same direction as that indicated by Dorin's preliminary results on knockout mice which is: lower DEFB CN is associated with a higher BMI. YMCA 2 under Bonferroni correction and without it and SCS cohorts showed no association with BMI. This could be due to a low sample size (524 and 741 respectively) with power too small to pick up a real effect in the population. The same trend was found when the test was carried out using only participants below the age of 24. For YMCA 2 and SCS cohorts to pick up a real effect; the minimum sample size has to be 1624.

As for the log transformed values for total cholesterol to HDL ratio, the higher the ratio was, the lower the DEFB CN would have been in YMCA 1 cohort if no Bonferroni

correction was applied. Under the Bonferroni correction, there was no association between DEFB CN and log transformed values for total cholesterol to HDL ratio. The effect was not detected in YMCA 2 (with and without Bonferroni correction) or SCS cohort as the power of the study was 32% and 44% respectively. For YMCA 2 and SCS cohorts to pick up a real effect; the minimum sample size has to be 1800.

In order to conclude if DEFB CN has an effect on BMI and log (TC: HDL), a replication study true to the discovery study; YMCA 1 cohort, not under Bonferroni correction, needs to be carried out, with a minimum sample size of 1800. The criteria for recruitment should be only males, aged 17 years and above, and of Eastern European origin.

In summary, results from this thesis confirm that the DEFB CNV region is 322kb in length, with a polymorphic inversion that occurs at a prevalence of 30% at the 8p23.1 genomic location that is independent of the DEFB CN. Paralogue Ratio Test (PRT) proved to be the best method of genotyping DEFB CNV especially in larger cohorts. Work from this thesis also founded the basis of developing an *in vitro* model system to investigate whether DEFB expression levels differ with CN in response to treatment with pneumolysin by using Normal Human Bronchial Epithelial (NHBE) cells. As far as case/control and cohort studies are concerned; results from this thesis does not support previous results that present an association between HIV viral load and DEFB CN. DEFB CN was also found not to be associated with recurrent UTIs in VUR patients, nor with hypertension. Data suggested that DEFB CN might be associated with BMI but this has not been reproduced in a smaller cohort.

9 Appendices

Appendix 1: Samples that showed heterogeneity upon analysing NimbleGen aCGH data

Samples	Position	Length (bp)	Genes	Gain or Loss	Repeat Elements Involved
C0053	7146450-7175250	28801	FAM66B/DEFB1091B	Gain	Various
	7175550-7187850	12301	FAM66B/USP17L1P/USP17L4	Loss	Simple
C0075	7175550-7183350	7801	FAM66B/USP17L1P/USP17L4	Loss	Simple
C0096	7175550-7187550	12001	FAM66B/USP17L1P/USP17L4	Loss	Simple
	7250250-7252350			Gain	LTR:HERVH
C0140	7250250-7252950			Gain	LTR:HERVH
C0195	7247250-7253250			Gain	LTR: HERVH
C0766	7247250-7252950			Gain	LTR:HERVH
C0877	7175550-7187550	12001	FAM66B/USP17L1P/USP17L4	Loss	Simple
	7247250-7252950			Gain	LTR:HERVH
C0888	7344150-7349850	5701	DEFB107A/B	Loss	LTR: HERVK
NA06994	7175550-7187850	12301	FAM66B/USP17L1P/USP17L4	Loss	Simple
NA07000	7175550-7187850	12301	FAM66B/USP17L1P/USP17L4	Loss	Simple
NA07019	7247250-7252950			Gain	LTR:HERVH
NA07029	7175550-7187850	12301	FAM66B/USP17L1P/USP17L4	Loss	Simple
NA07055	7175550-7187550	12001	FAM66B/USP17L1P/USP17L4	Loss	Simple
NA07345	7175550-7187850	12301	FAM66B/USP17L1P/USP17L4	Loss	
NA10831	7343850-7354050	10201	DEFB107A/B	Loss	LTR: HERVK
NA10835	7175550-7183050	7501	FAM66B/USP17L1P/USP17L4	Loss	Simple
NA10838	7175550-7187550	12001	FAM66B/USP17L1P/USP17L4	Loss	Simple

NA10846	7097250-7143450	46201	LOC349196	Loss	Various
	7175550-7187850	12301	FAM66B/USP17L1P/USP17L4	Loss	Simple
NA10851	7175550-7187550	12001	FAM66B/USP17L1P/USP17L4	Loss	Simple
	7247250-7252950			Gain	LTR:HERVH
NA10860	7246950-7252950			Gain	LTR:HERVH
NA11840	7175550-7187850	12301	FAM66B/USP17L1P/USP17L4	Loss	Simple
	7342950-7351650	8701	DEFB107A/B	Loss	LTR: HERVK
NA11986	7247250-7252950			Gain	LTR:HERVH
NA11991	7344150-7351350	7201	DEFB107A/B	Loss	LTR: HERVK
NA11996	7306950-7308450	1501	HE2/SPAG11B	Gain	
NA12005	7175550-7187550	12001	FAM66B/USP17L1P/USP17L4	Loss	Simple
	7343850-7351050	7201	DEFB107A/B	Loss	LTR: HERVK
NA12144	7247250-7252950			Gain	LTR:HERVH
NA12154	7328850-7329150	301	DEFB106A/B	Gain	Simple
NA12234	7247250-7252950			Gain	LTR:HERVH
	7328850-7330050	1201	DEFB106A/B	Gain	SINE+Simple
NA12239	7342950-7351950	9001	DEFB107A/B	Loss	LTR: HERVK
NA12753	7175550-7187850	12301	FAM66B/USP17L1P/USP17L4	Loss	Simple
NA12864	7247250-7252950			Gain	LTR:HERVH
NA12865	7176450-7187850	11401	FAM66B/USP17L1P/USP17L4	Loss	Simple
NA12873	7246950-7253850			Gain	LTR:HERVH
	7328850-7330050	1201	DEFB106A/B	Gain	LINE + Simple
NA18562	7343850-7349550	5701	DEFB107A/B	Loss	LTR: HERVK
NA18858	7342950-7351950	9001	DEFB107A/B	Loss	LTR: HERVK
NA18860	7247250-7252950			Loss	LTR:HERVH
	7342950-7351950	9001	DEFB107A/B	Loss	LTR: HERVK

NA19003	7342950-7352250	9301	DEFB107A/B	Loss	LTR: HERVK
	7352700-7389300	36601	DEFB107A/B	Gain	Various
NA19140	7175550-7187850	12301	FAM66B/USP17L1P/USP17L4	Loss	Simple
	7343250-7351050	7801	DEFB107A/B	Loss	LTR: HERVK
NA19153	7342950-7351950	9001	DEFB107A/B	Loss	LTR: HERVK
NA19204	7175550-7183350	7801	FAM66B/USP17L1P/USP17L4	Loss	Simple
	7342950-7352250	9301	DEFB107A/B	Loss	LTR:HERVK

Appendix 2: A list showing the different techniques in which the samples were genotyped for DEFB CN.

Sample name	PRT	NimbleGen	Agilent	nCounter	ddPCR	genome STRiP
co002	X				X	
co006	X				X	
co007	X	X			X	
co008	X				X	
co009	X				X	
co010	X				X	
co016	X				X	
co018	X				X	
co022	X				X	
co027	X				X	
co029	X				X	
co030	X				X	
co034	X				X	
co035	X				X	
co036	X				X	
co038	X				X	
co040	X				X	
co045	X				X	
co047	X				X	
co053	X	X			X	
co055	X				X	
co058	X				X	
co060	X				X	
co063	X				X	
co065	X				X	
co066	X				X	
co068	X				X	
co073	X				X	
co075	X	X				
co080	X				X	
co081	X				X	
co084	X				X	
co085	X				X	
co088	X	X			X	
co090	X				X	
co091	X				X	
co095	X				X	
CO096	X	X				
co097	X				X	
co098	X				X	

co100	X				X	
co1006	X				X	
co1008	X				X	
co1010	X				X	
co1011	X				X	
co106	X				X	
co107	X				X	
co108	X				X	
co111	X				X	
co121	X				X	
co123	X				X	
co126	X				X	
co136	X				X	
co137	X				X	
co139	X				X	
co143	X				X	
co145	X				X	
co147	X				X	
co149	X				X	
co150	X				X	
co152	X				X	
co154	X				X	
co156	X				X	
co157	X				X	
co160	X				X	
co166	X				X	
co167	X				X	
co168	X				X	
co176	X				X	
co178	X				X	
co180	X				X	
co182	X				X	
co183	X				X	
co184	X				X	
co185	X				X	
co186	X				X	
co187	X	X			X	
co188	X				X	
co189	X				X	
co190	X				X	
co191	X				X	
co192	X				X	
co194	X				X	
co195	X	X			X	

co196	X				X	
co197	X				X	
co201	X				X	
co203	X				X	
co204	X				X	
co207	X				X	
co208	X				X	
co210	X				X	
co215	X				X	
co722	X				X	
co723	X				X	
co724	X				X	
co725	X				X	
co728	X				X	
co730	X				X	
co731	X				X	
co735	X				X	
co739	X				X	
co741	X				X	
co744	X				X	
co747	X				X	
co748	X	X			X	
co749	X				X	
co750	X				X	
co753	X				X	
co755	X				X	
co766	X	X			X	
co781	X				X	
co786	X				X	
co832	X				X	
co848	X				X	
co849	X				X	
co850	X				X	
co851	X				X	
co854	X				X	
co855	X				X	
co856	X				X	
co857	X				X	
co858	X				X	
co861	X				X	
co862	X				X	
co863	X				X	
co864	X				X	
co870	X				X	

co871	X				X	
co877	X	X			X	
co880	X				X	
co881	X				X	
co882	X				X	
co883	X				X	
co884	X				X	
co886	X				X	
co888	X	X			X	
co891	X				X	
co892	X				X	
co893	X				X	
co894	X				X	
co895	X				X	
co896	X				X	
co897	X				X	
co898	X				X	
co899	X				X	
co901	X				X	
co902	X				X	
co904	X				X	
co906	X				X	
co907	X				X	
co908	X				X	
co909	X	X			X	
co913	X				X	
co917	X	X			X	
co920	X				X	
co921	X				X	
co937	X	X			X	
co938	X				X	
co940	X				X	
co941	X				X	
co953	X				X	
co956	X				X	
co958	X				X	
co959	X				X	
co960	X				X	
co968	X				X	
co969	X				X	
co977	X				X	
co978	X				X	
co994	X				X	
co996	X				X	

co997	X				X	
NA06985	X		X	X		
NA06991	X	X				
NA06993	X	X	X	X		
NA06994	X	X	X	X		X
NA07000	X	X	X	X		X
NA07019	X	X				
NA07022	X	X	X	X		
NA07029	X	X				
NA07034	X	X	X	X		
NA07048	X	X				X
NA07055	X	X	X	X		
NA07056	X	X	X	X		X
NA07345	X	X	X	X		
NA07348	X	X				
NA07357	X		X	X		X
NA10830	X	X				
NA10831	X	X				
NA10835	X	X				
NA10838	X	X				
NA10839	X	X				
NA10846	X	X				
NA10847	X	X				X
NA10851	X	X				X
NA10854	X	X				
NA10860	X	X				
NA10861	X	X				
NA10863	X	X				
NA11829	X		X	X		X
NA11830	X		X	X		X
NA11831	X	X	X	X		X
NA11832	X	X	X	X		
NA11839	X		X	X		
NA11840	X	X	X	X		
NA11881	X		X	X		
NA11882	X		X	X		
NA11892	X		X	X		
NA11893	X		X	X		
NA11894	X		X	X		
NA11992	X					X
NA11993	X					X
NA11994	X					X
NA11995	X	X	X	X		X
NA12003	X		X	X		X

NA12004	X		X	X		X
NA12005	X	X	X	X		
NA12006	X		X	X		X
NA12043	X		X	X		X
NA12044	X		X	X		X
NA12056	X		X	X		
NA12057	X		X	X		
NA12144	X	X	X	X		X
NA12145	X		X	X		
NA12146	X	X	X	X		
NA12154	X	X	X	X		X
NA12155	X		X	X		X
NA12156	X		X	X		
NA12234	X	X	X	X		
NA12236	X		X	X		
NA12239	X	X	X	X		
NA12248	X		X	X		
NA12249	X	X	X	X		X
NA12264	X		X	X		
NA12707	X	X				
NA12716	X	X	X	X		X
NA12717	X		X	X		
NA12750	X		X	X		X
NA12751	X		X	X		X
NA12752	X	X				
NA12760	X	X	X	X		
NA12761	X		X	X		X
NA12762	X		X	X		
NA12763	X		X	X		X
NA12801	X	X				
NA12802	X	X				
NA12812	X		X	X		
NA12813	X		X	X		
NA12814	X		X	X		
NA12815	X		X	X		
NA12864	X	X				
NA12865	X	X				
NA12872	X		X	X		
NA12873	X	X	X	X		
NA12874	X		X	X		
NA12875	X		X	X		
NA12891	X		X	X		
NA12892	X		X	X		
NA18500	X	X				

NA18501	X		X	X		X
NA18502	X	X	X	X		
NA18504	X		X	X		X
NA18505	X		X	X		X
NA18507	X		X	X		X
NA18508	X		X	X		X
NA18516	X		X	X		X
NA18517	X		X	X		X
NA18522	X		X	X		X
NA18523	X		X	X		
NA18524	X		X	X		
NA18526	X		X	X		X
NA18529	X		X	X		
NA18532	X		X	X		X
NA18537	X		X	X		X
NA18540	X		X	X		
NA18542	X		X	X		X
NA18545	X		X	X		X
NA18547	X		X	X		X
NA18550	X		X	X		X
NA18552	X		X	X		X
NA18555	X		X	X		X
NA18558	X		X	X		X
NA18561	X		X	X		X
NA18562	X	X	X	X		X
NA18563	X		X	X		
NA18564	X		X	X		X
NA18566	X		X	X		X
NA18570	X		X	X		X
NA18571	X		X	X		X
NA18572	X		X	X		X
NA18573	X		X	X		X
NA18576	X		X	X		X
NA18577	X		X	X		X
NA18579	X		X	X		X
NA18582	X		X	X		X
NA18592	X		X	X		X
NA18593	X		X	X		X
NA18594	X		X	X		
NA18603	X		X	X		X
NA18605	X		X	X		X
NA18608	X	X	X	X		X
NA18609	X		X	X		X
NA18611	X		X	X		X

NA18612	X		X	X		X
NA18620	X		X	X		X
NA18621	X		X	X		X
NA18622	X		X	X		X
NA18623	X		X	X		X
NA18624	X		X	X		X
NA18632	X		X	X		X
NA18633	X		X	X		X
NA18635	X		X	X		
NA18636	X		X	X		X
NA18637	X		X	X		X
NA18852	X		X	X		
NA18853	X		X	X		X
NA18855	X		X	X		
NA18856	X		X	X		X
NA18858	X	X	X	X		X
NA18859	X		X	X		
NA18861	X		X	X		X
NA18862	X		X	X		
NA18870	X		X	X		X
NA18871	X		X	X		X
NA18912	X		X	X		X
NA18913	X		X	X		
NA18940	X					X
NA18942	X					X
NA18943	X					X
NA18944	X					X
NA18945	X					X
NA18947	X					X
NA18948	X					X
NA18949	X					X
NA18951	X					X
NA18952	X					X
NA18953	X					X
NA18956	X					X
NA18959	X					X
NA18960	X					X
NA18961	X					X
NA18964	X					X
NA18965	X					X
NA18968	X					X
NA18971	X					X
NA18973	X					X
NA18974	X					X

NA18975	X					X
NA18976	X					X
NA18980	X					X
NA18981	X					X
NA18987	X					X
NA18990	X					X
NA18999	X					X
NA19000	X					X
NA19003	X	X				X
NA19005	X					X
NA19007	X					X
NA19012	X					X
NA19092	X		X	X		
NA19093	X		X	X		X
NA19098	X		X	X		X
NA19099	X		X	X		X
NA19101	X		X	X		
NA19102	X		X	X		X
NA19116	X		X	X		X
NA19119	X		X	X		X
NA19127	X		X	X		
NA19128	X		X	X		
NA19129	X					X
NA19130	X		X	X		
NA19131	X		X	X		X
NA19137	X		X	X		
NA19140	X	X	X	X		
NA19141	X		X	X		
NA19143	X		X	X		
NA19144	X		X	X		
NA19152	X		X	X		X
NA19153	X	X	X	X		
NA19159	X		X	X		
NA19160	X		X	X		X
NA19171	X		X	X		X
NA19172	X		X	X		X
NA19192	X		X	X		
NA19193	X		X	X		
NA19200	X		X	X		X
NA19201	X		X	X		
NA19203	X		X	X		
NA19204	X	X	X	X		X
NA19206	X		X	X		
NA19207	X		X	X		X

NA19209	X		X	X		X
NA19210	X		X	X		
NA19222	X		X	X		
NA19223	X		X	X		
NA19238	X		X	X		
NA19239	X		X	X		
PK1	X				X	
PK17	X				X	
PK18	X				X	
pk19	X				X	
PK2	X				X	
PK20	X				X	
PK21	X				X	
PK22	X				X	
PK23	X				X	
PK24	X				X	
PK25	X				X	
PK26	X				X	
PK27	X				X	
PK3	X				X	
PK32	X				X	
PK34	X				X	
PK36	X				X	
PK37	X				X	
PK4	X				X	
PK40	X				X	
PK41	X				X	
PK42	X				X	
PK43	X				X	
PK44	X				X	
PK46	X				X	
PK47	X				X	
PK48	X				X	
PK5	X				X	
PK50	X				X	
PK51	X				X	
PK52	X				X	
PK53	X				X	
PK54	X				X	
PK55	X				X	
PK56	X				X	
PK57	X				X	
PK58	X				X	
PK59	X				X	

PK6	X				X	
PK60	X				X	
PK61	X				X	
PK62	X				X	
PK63	X				X	
PK64	X				X	
PK65	X				X	
PK66	X				X	
PK67	X				X	
PK68	X				X	
PK69	X				X	
PK7	X				X	
PK70	X				X	
PK71	X				X	
PK74	X				X	
PK75	X				X	
PK76	X				X	
PK77	X				X	
PK78	X				X	
PK79	X				X	
PK8	X				X	
PK80	X				X	
PK81	X				X	
PK82	X				X	
PK84	X				X	
PK85	X				X	
PK87	X				X	
PK88	X				X	
PK89	X				X	
PK92	X				X	
SG0042	X				X	
SG0121	X				X	
SG0125	X				X	

10 Bibliography

Abe, S., Miura, K., Kinoshita, A., Mishima, H., Miura, S., Yamasaki, K., Hasegawa, Y., Higashijima, A., Jo, O., Sasaki, K., Yoshida, A., Yoshiura, K. and Masuzaki, H. (2013) 'Copy number variation of the antimicrobial-gene, defensin beta 4, is associated with susceptibility to cervical cancer.', *Journal of human genetics*, Nature Publishing Group, 58(5), pp. 250–3.

Abu Bakar, S., Hollox, E. and Armour, J. (2009) 'Allelic recombination between distinct genomic locations generates copy number diversity in human beta-defensins.', *Proceedings of the National Academy of Sciences of the United States of America*, 106(3), pp. 853–8.

Ahn, J. W., Bint, S., Bergbaum, A., Mann, K., Hall, R. P. and Ogilvie, C. M. (2013) 'Array CGH as a first line diagnostic test in place of karyotyping for postnatal referrals - results from four years' clinical application for over 8,700 patients', *Molecular Cytogenetics*, Molecular Cytogenetics, 6(1), p. 16.

Ahn, K., Gotay, N., Andersen, T. M., Anvari, A. A., Gochman, P., Lee, Y., Sanders, S., Guha, S., Darvasi, A., Glessner, J. T., Hakonarson, H., Lencz, T., State, M. W., Shugart, Y. Y. and Rapoport, J. L. (2014) 'High rate of disease-related copy number variations in childhood onset schizophrenia', *Mol Psychiatry*, JOUR, Macmillan Publishers Limited, 19(5), pp. 568–572.

Ahuja, S. K., Kulkarni, H., Catano, G., Agan, B. K., Camargo, J. F., He, W., O'Connell, R. J., Marconi, V. C., Delmar, J., Eron, J., Clark, R. A., Frost, S., Martin, J., Ahuja, S. S., Deeks, S. G., Little, S., Richman, D., Hecht, F. M. and Dolan, M. J. (2008) 'CCL3L1-CCR5 genotype influences durability of immune recovery during antiretroviral therapy of HIV-1–infected individuals', *Nature Medicine*, 14(4), pp. 413–420.

Aklillu, E., Odenthal-Hesse, L., Bowdrey, J., Habtewold, A., Ngaimisi, E., Yimer, G., Amogne, W., Mugusi, S., Minzi, O., Makonnen, E., Janabi, M., Mugusi, F., Aderaye, G., Hardwick, R., Fu, B., Viskaduraki, M., Yang, F. and Hollox, E. J. (2013) 'CCL3L1 copy number, HIV load, and immune reconstitution in sub-Saharan Africans', *BMC Infectious Diseases*, BMC Infectious Diseases, 13(1), p. 536.

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

disease.', *Human molecular genetics*, 19(24), pp. 4930–8.

Alexander, J. K. (2001) 'Obesity and coronary heart disease.', *The American journal of the medical sciences*, Journal Article, Review, United States, 321(4), pp. 215–224.

Andresen, E., Günther, G., Bullwinkel, J., Lange, C. and Heine, H. (2011) 'Increased expression of beta-defensin 1 (DEFB1) in chronic obstructive pulmonary disease', *PLoS ONE*, 6(7), pp. 1–10.

Antonacci, F., Kidd, J. M., Marques-Bonet, T., Ventura, M., Siswara, P., Jiang, Z. and Eichler, E. E. (2009) 'Characterization of six human disease-associated inversion polymorphisms', *Human Molecular Genetics*, 18(14), pp. 2555–2566.

Antoni, T., Blasi, F., Dartois, N. and Akova, M. (2015) 'Which individuals are at increased risk of pneumococcal disease and why? Impact of COPD, asthma, smoking, diabetes, and/or chronic heart disease on community-acquired pneumonia and invasive pneumococcal disease', *Thorax*, 70(10), pp. 984–989.

Arlt, M. F., Wilson, T. E. and Glover, T. W. (2012) 'Replication stress and mechanisms of CNV formation', *Current Opinion in Genetics & Development*, 22(3), pp. 204–210.

Armour, J. a, Sismani, C., Patsalis, P. C. and Cross, G. (2000) 'Measurement of locus copy number by hybridisation with amplifiable probes.', *Nucleic acids research*, 28(2), pp. 605–9.

Armour, J., Palla, R., Zeeuwen, P., den Heijer, M., Schalkwijk, J. and Hollox, E. (2007) 'Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats.', *Nucleic acids research*, 35(3), p. e19.

Artigas, M. S., Loth, D. W., Wain, L. V, Gharib, S. A., Obeidat, M., Tang, W., Zhai, G., Zhao, J. H., Smith, A. V., Huffman, J. E., Albrecht, E., Jackson, C. M., Evans, D. M., Cadby, G., Fornage, M., Manichaikul, A., Lopez, L. M., Johnson, T., Aldrich, M. C., Aspelund, T., Barroso, I., Campbell, H., Cassano, P. A., Couper, D. J., Eiriksdottir, G., Franceschini, N., Garcia, M., Gieger, C., Gislason, G. K., Grkovic, I., Hammond, C. J., Hancock, D. B., Harris, T. B., Ramasamy, A., Heckbert, S. R., Heliövaara, M., Homuth, G., Hysi, P. G., James, A. L., Jankovic, S., Joubert, B. R., Karrasch, S., Klopp, N., Koch, B., Kritchevsky, S. B., Launer, L. J., Liu, Y., Loehr, L. R., Lohman, K., Loos, R. J. F., Lumley, T., Al Balushi, K. A., Ang, W. Q., Barr, R. G., Beilby, J., Blakey, J. D., Boban, M., Boraska, V., Brisman, J., Britton, J. R., Brusselle, G. G., Cooper, C., Curjuric, I., Dahgam, S., Deary, I. J., Ebrahim, S., Eijgelsheim, M., Francks, C., Gaysina, D., Granell, R., Gu, X., Hankinson, J. L., Hardy, R., Harris, S. E.,

Henderson, J., Henry, A., Hingorani, A. D., Hofman, A., Holt, P. G., Hui, J., Hunter, M. L., Imboden, M., Jameson, K. A., Kerr, S. M., Kolcic, I., Kronenberg, F., Liu, J. Z., Marchini, J., McKeever, T., Morris, A. D., Olin, A.-C., Porteous, D. J., Postma, D. S., Rich, S. S., Ring, S. M., Rivadeneira, F., Rochat, T., Sayer, A. A., Sayers, I., Sly, P. D., Smith, G. D., Sood, A., Starr, J. M., Uitterlinden, A. G., Vonk, J. M., Wannamethee, S. G., Whincup, P. H., Wijmenga, C., Williams, O. D., Wong, A., Mangino, M., Marciante, K. D., McArdle, W. L., Meibohm, B., Morrison, A. C., North, K. E., Omenaas, E., Palmer, L. J., Pietiläinen, K. H., Pin, I., Pola[sbreve]ek, O., Pouta, A., Psaty, B. M., Hartikainen, A.-L., Rantanen, T., Ripatti, S., Rotter, J. I., Rudan, I., Rudnicka, A. R., Schulz, H., Shin, S.-Y., Spector, T. D., Surakka, I., Vitart, V., Völzke, H., Wareham, N. J., Warrington, N. M., Wichmann, H.-E., Wild, S. H., Wilk, J. B., Wjst, M., Wright, A. F., Zgaga, L., Zemunik, T., Pennell, C. E., Nyberg, F., Kuh, D., Holloway, J. W., Boezen, H. M., Lawlor, D. A., Morris, R. W., Probst-Hensch, N., Kaprio, J., Wilson, J. F., Hayward, C., Kähönen, M., Heinrich, J., Musk, A. W., Jarvis, D. L., Gläser, S., Järvelin, M.-R., Ch Stricker, B. H., Elliott, P., O'Connor, G. T., Strachan, D. P., London, S. J., Hall, I. P., Gudnason, V. and Tobin, M. D. (2011) 'Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function', *Nature Genetics*, 43(11), pp. 1082–1090.

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

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

Bals, R. (2000) 'Epithelial antimicrobial peptides in host defense against infection.', *Respiratory research*, 1, pp. 141–150.

Bals, R., Wang, X., Wu, Z., Freeman, T., Bafna, V., Zasloff, M. and Wilson, J. M. (1998) 'Human beta-defensin 2 is a salt-sensitive peptide antibiotic expressed in human lung.', *Journal of Clinical Investigation*, Journal Article, Research Support, Non-U.S. Gov't, Research Support, U.S. Gov't, P.H.S., UNITED STATES, 102(5), pp. 874–880.

Barber, J., Bunyan, D., Curtis, M., Robinson, D., Morlot, S., Dermitzel, A., Liehr, T., Alves, C., Trindade, J., Paramos, A., Cooper, C., Ocraft, K., Taylor, E. J. and Maloney, V. (2010) '8P23.1 Duplication Syndrome Differentiated From Copy Number Variation of the Defensin Cluster At Prenatal Diagnosis in Four New Families.', *Molecular Cytogenetics*, 3,

p. 3.

Barber, J., Joyce, C., Collinson, M., Nicholson, J., Willatt, L., Dyson, H., Bateman, M., Green, A., Yates, J. and Dennis, N. (1998) 'Duplication of 8p23.1: a cytogenetic anomaly with no established clinical significance.', *Journal of medical genetics*, 35(6), pp. 491–6.

Barber, J., Maloney, V., Huang, S., Bunyan, D., Cresswell, L., Kinning, E., Benson, A., Cheetham, T., Wyllie, J., Lynch, S. A., Zwolinski, S., Prescott, L., Crow, Y., Morgan, R. and Hobson, E. (2008) '8p23.1 duplication syndrome; a novel genomic condition with unexpected complexity revealed by array CGH.', *European journal of human genetics : EJHG*, 16(1), pp. 18–27.

Barber, J., Rosenfeld, J. A., Graham, J. M., Kramer, N., Lachlan, K. L., Bateman, M. S., Collinson, M. N., Stadheim, B. F., Turner, C. L. S., Gauthier, J. N., Reimschisel, T. E., Qureshi, A. M., Dabir, T. A., Humphreys, M. W., Marble, M., Huang, T., Beal, S. J., Massiah, J., Taylor, E.-J. and Wynn, S. L. (2015) 'Inside the 8p23.1 duplication syndrome; eight microduplications of likely or uncertain clinical significance.', *American journal of medical genetics. Part A*, (June), pp. 2052–2064.

Baris, H. N., Tan, W., Kimonis, V. E. and Irons, M. B. (2007) 'Diagnostic Utility of Array-Based Comparative Genomic Hybridization in a Clinical Setting', *American Journal of Medical Genetics Part*, 143(A), pp. 2523–2533.

Bassuk, A. G., Geraghty, E., Wu, S., Mullen, S. A., Berkovic, S. F., Scheffer, I. E. and Mefford, H. C. (2013) 'Deletions of 16p11.2 and 19p13.2 in a family with intellectual disability and generalized epilepsy', *American Journal of Medical Genetics Part A*, 161(7), pp. 1722–1725.

Bayés, M., Magano, L. F., Rivera, N., Flores, R. and Pérez Jurado, L. (2003) 'Mutational Mechanisms of Williams-Beuren Syndrome Deletions', *The American Journal of Human Genetics*, 73(1), pp. 131–151.

Becknell, B., Spencer, J. D., Carpenter, A. R., Chen, X., Singh, A., Ploeger, S., Kline, J., Ellsworth, P., Li, B., Proksch, E., Schwaderer, A. L., Hains, D. S., Justice, S. S. and McHugh, K. M. (2013) 'Expression and Antimicrobial Function of Beta-Defensin 1 in the Lower Urinary Tract', *PLoS ONE*, 8(10), pp. 1–10.

Bensch, K. W., Raida, M., Hans-jfirgen, M. and Schulz-knappe, P. (1995) 'hBD-1: a novel Beta defensin from human plasma', *FEBS Letters*, 368, pp. 331–335.

Bentley, R. W., Pearson, J., Gearry, R. B., Barclay, M. L., McKinney, C., Merriman, T. R. and Roberts, R. L. (2010) 'Association of higher DEFB4 genomic copy number with Crohn's disease.', *The American journal of gastroenterology*, Nature Publishing Group, 105(2), pp. 354–9.

Bolker, B. (2006) 'Stochastic simulation and power analysis', In *Ecological Models and Data in R*, Princeton University Press.

Bondeson, M. L., Malmgren, H., Dahl, N., Carlberg, B. M. and Pettersson, U. (1995) 'Presence of an IDS-related locus (IDS2) in Xq28 complicates the mutational analysis of Hunter syndrome.', *European journal of human genetics*, 3(4), pp. 219–27.

Boonpeng, H. and Yusoff, K. (2013) 'The utility of copy number variation (CNV) in studies of hypertension-related left ventricular hypertrophy (LVH): rationale, potential and challenges.', *Molecular cytogenetics*, Molecular Cytogenetics, 6(1), p. 8.

Bosch, N., Escaramís, G., Mercader, J. M., Armengol, L. and Estivill, X. (2008) 'Analysis of the multi-copy gene family FAM90A as a copy number variant in different ethnic backgrounds.', *Gene*, 420(2), pp. 113–7.

Bosch, N., Morell, M., Ponsa, I., Mercader, J. M., Armengol, L. and Estivill, X. (2009) 'Nucleotide, cytogenetic and expression impact of the human chromosome 8p23.1 inversion polymorphism', *PLoS ONE*, 4(12).

Braff, M. H. and Gallo, R. L. (2006) *Antimicrobial Peptides and Human Disease*, Shafer, W. M. (ed.), *Antimicrobial Peptides and Human Disease*, Current Topics in Microbiology and Immunology, Springer Berlin Heidelberg.

Brandström, P., Esbjörner, E., Herthelius, M., Swerkersson, S., Jodal, U. and Hansson, S. (2010) 'The Swedish Reflux Trial in Children: III. Urinary Tract Infection Pattern', *The Journal of Urology*, 184(1), pp. 286–291.

Broad Institute (2015) 'genome STRiP overview', [online] Available from: http://www.broadinstitute.org/software/genomestrip/ (Accessed 1 October 2015).

Brockus, C. W., Jackwood, M. W. and Harmon, B. G. (1998) 'Characterization of beta-defensin prepropeptide mRNA from chicken and turkey bone marrow.', *Animal genetics*, 29(4), pp. 283–9.

Brown, C. D., Higgins, M., Donato, K. A., Rohde, F. C., Garrison, R., Obarzanek, E., Ernst, N. D. and Horan, M. (2000) 'Body mass index and the prevalence of hypertension and

dyslipidemia.', *Obesity research*, Journal Article, Research Support, U.S. Gov't, P.H.S., United States, 8(9), pp. 605–619.

Buchbinder, S. P., Mehrotra, D. V, Duerr, A., Fitzgerald, D. W., Mogg, R., Li, D., Gilbert, P. B., Lama, J. R., Marmor, M., Del Rio, C., McElrath, M. J., Casimiro, D. R., Gottesdiener, K. M., Chodakewitz, J. A., Corey, L. and Robertson, M. N. (2008) 'Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial.', *Lancet*, 372(9653), pp. 1881–93.

Burrows, J. F., McGrattan, M. J. and Johnston, J. A. (2005) 'The DUB/USP17 deubiquitinating enzymes, a multigene family within a tandemly repeated sequence.', *Genomics*, Journal Article, United States, 85(4), pp. 524–529.

Bustin, S. a, Benes, V., Garson, J. a, Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J. and Wittwer, C. T. (2009) 'The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.', *Clinical chemistry*, 55(4), pp. 611–22.

Cahan, P., Li, Y., Izumi, M. and Graubert, T. A. (2009) 'The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells.', *Nature genetics*, Journal Article, Research Support, N.I.H., Extramural, Research Support, Non-U.S. Gov't, United States, 41(4), pp. 430–437.

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

Cantsilieris, S., Baird, P. N. and White, S. J. (2013) 'Molecular methods for genotyping complex copy number polymorphisms', *Genomics*, Elsevier Inc., 101(2), pp. 86–93.

Carpenter, M. a, Hoberman, A., Mattoo, T. K., Mathews, R., Keren, R., Chesney, R. W., Moxey-Mims, M. and Greenfield, S. P. (2013) 'The RIVUR trial: profile and baseline clinical associations of children with vesicoureteral reflux.', *Pediatrics*, 132(1), pp. e34-45.

Carter, N. P. (2007) 'Methods and strategies for analyzing copy number variation using DNA microarrays.', *Nature genetics*, 39(7 Suppl), pp. S16-21.

Castaldi, P. J., Cho, M. H., Litonjua, A. A., Bakke, P., Gulsvik, A., Lomas, D. A., Anderson, W., Beaty, T. H., Hokanson, J. E., Crapo, J. D., Laird, N. and Silverman, E. K. (2011) 'The association of genome-wide significant spirometric loci with chronic obstructive

pulmonary disease susceptibility', *American Journal of Respiratory Cell and Molecular Biology*, 45(6), pp. 1147–1153.

Chang, T. L.-Y., François, F., Mosoian, A. and Klotman, M. E. (2003) 'CAF-mediated human immunodeficiency virus (HIV) type 1 transcriptional inhibition is distinct from alpha-defensin-1 HIV inhibition.', *Journal of virology*, 77(12), pp. 6777–84.

Charchar, F., Tomaszewski, Lacka, B., Zakrzewski, J., Zukowska-Szczechowska, E., Grzeszczak, W. and Dominiczak, A. (2004) 'Association of the human Y chromosome with cholesterol levels in the general population.', *Arteriosclerosis, thrombosis, and vascular biology*, Journal Article, Research Support, Non-U.S. Gov't, United States, 24(2), pp. 308–12.

Chen, H., Xu, Z., Peng, L., Fang, X., Yin, X., Xu, N. and Cen, P. (2006) 'Recent advances in the research and development of human defensins.', *Peptides*, 27(4), pp. 931–40.

Chen, Q., Book, M., Fang, X., Hoeft, A. and Stuber, F. (2006) 'Screening of copy number polymorphisms in human beta-defensin genes using modified real-time quantitative PCR.', *Journal of immunological methods*, Evaluation Studies, Journal Article, Research Support, Non-U.S. Gov't, Netherlands, 308(1–2), pp. 231–240.

Chesney, R., Carpenter, M., Moxey-Mims, M., Nyberg, L., Greenfield, S., Hoberman, A., Keren, R., Matthews, R., Mattoo, T. and RIVUR steering Committee (2008) 'Randomized Intervention for Children with Vesicoureteral Reflux (RIVUR): Background Commentary of RIVUR Investigators', *Pediatrics*, 122(0 5), pp. S233–S239.

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D. and Todd, J. A. (2005) 'Population structure, differential bias and genomic control in a large-scale, case-control association study.', *Nature genetics*, Comparative Study, Journal Article, Research Support, Non-U.S. Gov't, United States, 37(11), pp. 1243–1246.

Clement, K., Vaisse, C., Lahlou, N., Cabrol, S., Pelloux, V., Cassuto, D., Gourmelen, M., Dina, C., Chambaz, J., Lacorte, J. M., Basdevant, A., Bougneres, P., Lebouc, Y., Froguel, P. and Guy-Grand, B. (1998) 'A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction.', *Nature*, Journal Article, Research Support, Non-U.S. Gov't, ENGLAND, 392(6674), pp. 398–401.

Cole, A. M., Wang, W., Waring, A. J. and Lehrer, R. I. (2004) 'Retrocyclins: using past as prologue.', *Current protein & peptide science*, 5(5), pp. 373–81.

Conrad, D. F., Bird, C., Blackburne, B., Lindsay, S., Mamanova, L., Lee, C., Turner, D. J. and Hurles, M. E. (2010) 'Mutation spectrum revealed by breakpoint sequencing of human germline CNVs', *Nature Genetics*, 42(5), pp. 385–391.

Conrad, D. F. and Hurles, M. E. (2007) 'The population genetics of structural variation', *Nature Genetics*, 39(7s), pp. S30–S36.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., MacArthur, D. G., MacDonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W. and Hurles, M. E. (2010) 'Origins and functional impact of copy number variation in the human genome', *Nature*, Nature Publishing Group, 464(7289), pp. 704–712.

Cookson, W. O. C. and Moffatt, M. F. (2011) 'Genetics of complex airway disease.', *Proceedings of the American Thoracic Society*, 8(2), pp. 149–153.

Correa, P. G. and Oguiura, N. (2013) 'Phylogenetic analysis of beta-defensin-like genes of Bothrops, Crotalus and Lachesis snakes.', *Toxicon : official journal of the International Society on Toxinology*, Journal Article, Research Support, Non-U.S. Gov't, England, 69, pp. 65–74.

Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D. F., Giannoulatou, E., Holmes, C., Marchini, J. L., Stirrups, K., Tobin, M. D., Wain, L. V, Yau, C., Aerts, J., Ahmad, T., Andrews, T. D., Arbury, H., Attwood, A., Auton, A., Ball, S. G., Balmforth, A. J., Barrett, J. C., Barroso, I., Barton, A., Bennett, A. J., Bhaskar, S., Blaszczyk, K., Bowes, J., Brand, O. J., Braund, P. S., Bredin, F., Breen, G., Brown, M. J., Bruce, I. N., Bull, J., Burren, O. S., Burton, J., Byrnes, J., Caesar, S., Clee, C. M., Coffey, A. J., Connell, J. M. C., Cooper, J. D., Dominiczak, A. F., Downes, K., Drummond, H. E., Dudakia, D., Dunham, A., Ebbs, B., Eccles, D., Edkins, S., Edwards, C., Elliot, A., Emery, P., Evans, D. M., Evans, G., Eyre, S., Farmer, A., Ferrier, I. N., Feuk, L., Fitzgerald, T., Flynn, E., Forbes, A., Forty, L., Franklyn, J. a, Freathy, R. M., Gibbs, P., Gilbert, P., Gokumen, O., Gordon-Smith, K., Gray, E., Green, E., Groves, C. J., Grozeva, D., Gwilliam, R., Hall, A., Hammond, N., Hardy, M., Harrison, P., Hassanali, N., Hebaishi, H., Hines, S., Hinks, A., Hitman, G. a, Hocking, L., Howard, E., Howard, P., Howson, J. M. M., Hughes, D., Hunt, S., Isaacs, J. D., Jain, M., Jewell, D. P., Johnson, T., Jolley, J. D., Jones, I. R., Jones, L. a, Kirov, G., Langford, C. F., Lango-Allen, H., Lathrop, G. M., Lee, J., Lee, K. L., Lees, C., Lewis, K., Lindgren, C. M., Maisuria-Armer, M., Maller, J., Mansfield, J., Martin,

P., Massey, D. C. O., McArdle, W. L., McGuffin, P., McLay, K. E., Mentzer, A., Mimmack, M. L., Morgan, A. E., Morris, A. P., Mowat, C., Myers, S., Newman, W., Nimmo, E. R., O'Donovan, M. C., Onipinla, A., Onyiah, I., Ovington, N. R., Owen, M. J., Palin, K., Parnell, K., Pernet, D., Perry, J. R. B., Phillips, A., Pinto, D., Prescott, N. J., Prokopenko, I., Quail, M. a, Rafelt, S., Rayner, N. W., Redon, R., Reid, D. M., Renwick, Ring, S. M., Robertson, N., Russell, E., St Clair, D., Sambrook, J. G., Sanderson, J. D., Schuilenburg, H., Scott, C. E., Scott, R., Seal, S., Shaw-Hawkins, S., Shields, B. M., Simmonds, M. J., Smyth, D. J., Somaskantharajah, E., Spanova, K., Steer, S., Stephens, J., Stevens, H. E., Stone, M. a, Su, Z., Symmons, D. P. M., Thompson, J. R., Thomson, W., Travers, M. E., Turnbull, C., Valsesia, A., Walker, M., Walker, N. M., Wallace, C., Warren-Perry, M., Watkins, N. a, Webster, J., Weedon, M. N., Wilson, A. G., Woodburn, M., Wordsworth, B. P., Young, A. H., Zeggini, E., Carter, N. P., Frayling, T. M., Lee, C., McVean, G., Munroe, P. B., Palotie, A., Sawcer, S. J., Scherer, S. W., Strachan, D. P., Tyler-Smith, C., Brown, M. a, Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Gough, S. C. L., Hall, A. S., Hattersley, A. T., Hill, A. V. S., Mathew, C. G., Pembrey, M., Satsangi, J., Stratton, M. R., Worthington, J., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W., Parkes, M., Rahman, N., Todd, J. a, Samani, N. J. and Donnelly, P. (2010) 'Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls.', *Nature*, Nature Publishing Group, 464(7289), pp. 713–20.

Cukier, H. N., Pericak-Vance, M. A., Gilbert, J. R. and Hedges, D. J. (2009) 'Sample degradation leads to false-positive copy number variation calls in multiplex real-time polymerase chain reaction assays', *Analytical Biochemistry*, 386(2), pp. 288–290.

D'haene, B., Vandesompele, J. and Hellemans, J. (2010) 'Accurate and objective copy number profiling using real-time quantitative PCR.', *Methods (San Diego, Calif.)*, Journal Article, Research Support, Non-U.S. Gov't, Review, United States, 50(4), pp. 262–270.

Davis, E. G., Sang, Y. and Blecha, F. (2004) 'Equine beta-defensin-1: full-length cDNA sequence and tissue expression.', *Veterinary immunology and immunopathology*, Journal Article, Research Support, Non-U.S. Gov't, Netherlands, 99(1–2), pp. 127–132.

Dereli-Oz, A., Versini, G. and Halazonetis, T. (2011) 'Studies of genomic copy number changes in human cancers reveal signatures of DNA replication stress.', *Molecular oncology*, Journal Article, Research Support, Non-U.S. Gov't, Review, Netherlands, 5(4), pp. 308–314.

Devriendt, K., De Mars, K., De Cock, P., Gewillig, M. and Fryns, J. P. (1995) 'Terminal deletion in

chromosome region 8p23.1-8pter in a child with features of velo-cardio-facial syndrome.', *Annales de génétique*, 38(4), pp. 228–30.

Devriendt, K., Matthijs, G., Van Dael, R., Gewillig, M., Eyskens, B., Hjalgrim, H., Dolmer, B., McGaughran, J., Bröndum-Nielsen, K., Marynen, P., Fryns, J. P. and Vermeesch, J. R. (1999) 'Delineation of the critical deletion region for congenital heart defects, on chromosome 8p23.1.', *American journal of human genetics*, 64(4), pp. 1119–26.

Dunsche, a, Açil, Y., Siebert, R., Harder, J., Schröder, J. M. and Jepsen, S. (2001) 'Expression profile of human defensins and antimicrobial proteins in oral tissues.', *Journal of oral pathology & medicine*, 30(3), pp. 154–8.

Elder, J. S., Peters, C. A., Arant, B. S., Ewalt, D. H., Hawtrey, C. E., Hurwitz, R. S., Parrott, T. S., Snyder, H. M., Weiss, R. A., Woolf, S. H. and Hasselblad, V. (1997) 'Pediatric vesicoureteral reflux guidelines panel summary report on the management of primary vesicoureteral reflux in children', *Journal of Urology*, 157(5), pp. 1846–1851.

EMBL-EBI (n.d.) 'Expression Atlas', [online] Available from: http://www.ebi.ac.uk/gxa/home.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C. and Gibbs, R. A. (2012) 'Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology', Liu, Z. (ed.), *PLoS ONE*, 7(11), p. e47768.

Fanales-Belasio, E., Raimondo, M., Suligoi, B. and Butto, S. (2010) 'HIV virology and pathogenetic mechanisms of infection: a brief overview.', *Annali dell'Istituto superiore di sanita*, Journal Article, Review, Italy, 46(1), pp. 5–14.

Farzan, M., Choe, H., Martin, K. A., Sun, Y., Sidelko, M., Mackay, C. R., Gerard, N. P., Sodroski, J. and Gerard, C. (1997) 'HIV-1 entry and macrophage inflammatory protein-1beta-mediated signaling are independent functions of the chemokine receptor CCR5.', *The Journal of biological chemistry*, 272(11), pp. 6854–6857.

Fellay, J., Ge, D., Shianna, K. V, Colombo, S., Ledergerber, B., Cirulli, E. T., Urban, T. J., Zhang, K., Gumbs, C. E., Smith, J. P., Castagna, A., Cozzi-Lepri, A., De Luca, A., Easterbrook, P., Günthard, H. F., Mallal, S., Mussini, C., Dalmau, J., Martinez-Picado, J., Miro, J. M., Obel, N., Wolinsky, S. M., Martinson, J. J., Detels, R., Margolick, J. B., Jacobson, L. P., Descombes, P., Antonarakis, S. E., Beckmann, J. S., O'Brien, S. J., Letvin, N. L., McMichael, A. J., Haynes, B. F., Carrington, M., Feng, S., Telenti, A. and Goldstein, D. B. (2009) 'Common genetic variation and the control of HIV-1 in humans.', *PLoS genetics*, 5(12), p.

e1000791.

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

Fernandez-Jimenez, N., Castellanos-Rubio, A., Plaza-Izurieta, L., Gutierrez, G., Irastorza, I., Castaño, L., Vitoria, J. C. and Bilbao, J. R. (2011) 'Accuracy in copy number calling by qPCR and PRT: a matter of DNA.', *PloS one*, 6(12), p. e28910.

Feuk, L., Carson, A. R. and Scherer, S. W. (2006) 'Structural variation in the human genome.', *Nature reviews. Genetics*, 7(2), pp. 85–97.

Fode, P., Jespersgaard, C., Hardwick, R. J., Bogle, H., Theisen, M., Dodoo, D., Lenicek, M., Vitek, L., Vieira, A., Freitas, J., Andersen, P. S. and Hollox, E. J. (2011) 'Determination of beta-defensin genomic copy number in different populations: a comparison of three methods.', *PloS one*, 6(2), p. e16768.

Forni, D., Martin, D., Abujaber, R., Sharp, A. J., Sironi, M. and Hollox, E. (2015) 'Determining multiallelic complex copy number and sequence variation from high coverage exome sequencing data', *BMC Genomics*, BMC Genomics, 16(1), p. 891.

Frank, B., Bermejo, J. L., Hemminki, K., Sutter, C., Wappenschmidt, B., Meindl, a., Kiechle-Bahat, M., Bugert, P., Schmutzler, R. K., Bartram, C. R. and Burwinkel, B. (2007) 'Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk', *Carcinogenesis*, 28(7), pp. 1442–1445.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. a, Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter, N. P., Scherer, S. W. and Lee, C. (2006) 'Copy number variation: new insights in genome diversity.', *Genome research*, 16(8), pp. 949–61.

Frese, E. M., Fick, A. and Sadowsky, H. S. (2011) 'Blood pressure measurement guidelines for physical therapists.', *Cardiopulmonary physical therapy journal*, 22(2), pp. 5–12.

Funderburg, N., Lederman, M. M., Feng, Z., Drage, M. G., Jadlowsky, J., Harding, C. V., Weinberg, A. and Sieg, S. F. (2007) 'Human -defensin-3 activates professional antigen-presenting cells via Toll-like receptors 1 and 2', *Proceedings of the National Academy of*

Sciences, 104(47), pp. 18631–18635.

Furukawa, M., Wheeler, S., Clark, A. M. and Wells, A. (2015) 'Lung Epithelial Cells Induce Both Phenotype Alteration and Senescence in Breast Cancer Cells', Samant, R. (ed.), *PLOS ONE*, 10(1), p. e0118060.

Van Gaal, L. F., Mertens, I. L. and De Block, C. E. (2006) 'Mechanisms linking obesity with cardiovascular disease', *Nature*, JOUR, 444(7121), pp. 875–880.

Ganz, T. (2003) 'Defensins: antimicrobial peptides of innate immunity.', *Nature reviews. Immunology*, 3(9), pp. 710–20.

Ganz, T. and Lehrer, R. (1994) 'Defensins.', *Current opinion in immunology*, 6(4), pp. 584–9.

Ganz, T., Selsted, M. E. and Lehrer, R. I. (2009) 'Defensins', *European Journal of Haematology*, Comparative Study, Journal Article, Review, DENMARK, 44(1), pp. 1–8.

Garcia, J. R., Jaumann, F., Schulz, S., Krause, A., Rodriguez-Jimenez, J., Forssmann, U., Adermann, K., Kluver, E., Vogelmeier, C., Becker, D., Hedrich, R., Forssmann, W. G. and Bals, R. (2001) 'Identification of a novel, multifunctional beta-defensin (human beta-defensin 3) with specific antimicrobial activity. Its interaction with plasma membranes of Xenopus oocytes and the induction of macrophage chemoattraction.', *Cell and tissue research*, Journal Article, Research Support, Non-U.S. Gov't, Germany, 306(2), pp. 257–264.

Gelderblom, H. R. (1997) 'Fine structure of HIV and SIV', *Los Alamos National Laboratory Web site*, Berlin, [online] Available from: http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/1997/partIII/Gelderblom.pdf (Accessed 25 August 2015).

Ghanem, N., Uring-Lambert, B., Abbal, M., Hauptmann, G., Lefranc, M. P. and Lefranc, G. (1988) 'Polymorphism of MHC class III genes: definition of restriction fragment linkage groups and evidence for frequent deletions and duplications.', *Human genetics*, Journal Article, Research Support, Non-U.S. Gov't, GERMANY, WEST, 79(3), pp. 209–218.

Giglio, S., Broman, K., Matsumoto, N., Calvari, V., Gimelli, G., Neumann, T., Ohashi, H., Voullaire, L., Larizza, D., Giorda, R., Weber, J., Ledbetter, D. and Zuffardi, O. (2001) 'Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements.', *American journal of human genetics*, 68(4), pp. 874–83.

Goldman, M. J., Anderson, G. M., Stolzenberg, E. D., Kari, U. P., Zasloff, M. and Wilson, J. M.

(1997) ‘Human beta-defensin-1 is a salt-sensitive antibiotic in lung that is inactivated in cystic fibrosis.’, *Cell*, Journal Article, Research Support, Non-U.S. Gov’t, UNITED STATES, 88(4), pp. 553–560.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J., Murthy, K. K., Rovin, B. H., Bradley, W., Clark, R. A., Anderson, S. A., O’connell, R. J., Agan, B. K., Ahuja, S. S., Bologna, R., Sen, L., Dolan, M. J. and Ahuja, S. K. (2005) ‘The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility.’, *Science*, Journal Article, Research Support, Non-U.S. Gov’t, Research Support, U.S. Gov’t, Non-P.H.S., Research Support, U.S. Gov’t, P.H.S., United States, 307(5714), pp. 1434–1440.

Goodman, S. (2001) ‘Of P-values and Bayes: a modest proposal.’, *Epidemiology*, 12(3), pp. 295–297.

Goossens, M., Dozy, A. M., Emburyt, S. H., Zachariadest, Z., Hadjiminast, M. G., Stamatoyannopoulos, G. and Kan, Y. W. a I. (1980) ‘Triplicated a-globin loci’, *Genetics*, 77(1), pp. 518–521.

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

Grundy, S. M., Brewer, H. B., Cleeman, J. I., Smith, S. C. and Lenfant, C. (2004) ‘Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition.’, *Arteriosclerosis, thrombosis, and vascular biology*, 24(2), pp. e13–e18.

Guo, C.-J., Tan, N., Song, L., Douglas, S. D. and Ho, W.-Z. (2004) ‘Alpha-defensins inhibit HIV infection of macrophages through upregulation of CC-chemokines.’, *AIDS (London, England)*, 18(8), pp. 1217–8.

Hancock, D. B., Eijgelsheim, M., Wilk, J. B., Gharib, S. a, Loehr, L. R., Marciante, K. D., Franceschini, N., van Durme, Y. M. T. a, Chen, T.-H., Barr, R. G., Schabath, M. B., Couper, D. J., Brusselle, G. G., Psaty, B. M., van Duijn, C. M., Rotter, J. I., Uitterlinden, A. G., Hofman, A., Punjabi, N. M., Rivadeneira, F., Morrison, A. C., Enright, P. L., North, K. E., Heckbert, S. R., Lumley, T., Stricker, B. H. C., O’Connor, G. T. and London, S. J. (2010) ‘Meta-analyses of genome-wide association studies identify multiple loci associated with

pulmonary function.', *Nature genetics*, 42(1), pp. 45–52.

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M. and McCarroll, S. a (2015) 'Large multiallelic copy number variations in humans.', *Nature genetics*, Nature Publishing Group, (January), pp. 1–10.

Handsaker, R. E., Korn, J. M., Nemesh, J. and McCarroll, S. a (2011) 'Discovery and genotyping of genome structural polymorphism by sequencing on a population scale.', *Nature genetics*, Nature Publishing Group, 43(3), pp. 269–76.

Harder, J., Bartels, J., Christophers, E. and Schroder, J. (2001) 'Isolation and Characterization of Human beta-Defensin-3, a Novel Human Inducible Peptide Antibiotic', *Journal of Biological Chemistry*, 276(8), pp. 5707–5713.

Harder, J., Meyer-Hoffert, U., Teran, L. M., Schwichtenberg, L., Bartels, J., Maune, S. and Schroder, J. M. (2000) 'Mucoid Pseudomonas aeruginosa, TNF-alpha, and IL-1beta, but not IL-6, induce human beta-defensin-2 in respiratory epithelia.', *American journal of respiratory cell and molecular biology*, Journal Article, Research Support, Non-U.S. Gov't, UNITED STATES, 22(6), pp. 714–721.

Hardwick, R. J., Amogne, W., Mugusi, S., Yimer, G., Ngaimisi, E., Habtewold, A., Minzi, O., Makonnen, E., Janabi, M., Machado, L. R., Viskaduraki, M., Mugusi, F., Aderaye, G., Lindquist, L., Hollox, E. J. and Aklillu, E. (2012) 'β-defensin genomic copy number is associated with HIV load and immune reconstitution in sub-saharan Africans.', *The Journal of infectious diseases*, 206(7), pp. 1012–9.

Hardwick, R. J., Machado, L. R., Zuccherato, L. W., Antolinos, S., Xue, Y., Shawa, N., Gilman, R. H., Cabrera, L., Berg, D. E., Tyler-Smith, C., Kelly, P., Tarazona-Santos, E. and Hollox, E. J. (2011) 'A worldwide analysis of beta-defensin copy number variation suggests recent selection of a high-expressing DEFB103 gene copy in East Asia.', *Human mutation*, 32(7), pp. 743–50.

Hardwick, R. J., Ménard, A., Sironi, M., Milet, J., Garcia, A., Sese, C., Yang, F., Fu, B., Courtin, D. and Hollox, E. J. (2014) 'Haptoglobin (HP) and Haptoglobin-related protein (HPR) copy number variation, natural selection, and trypanosomiasis', *Human Genetics*, 133(1), pp. 69–83.

Haridan, U., Mokhtar, U., Machado, L., Abdul Aziz, A. T., Shueb, R., Zaid, M., Sim, B., Mustafa, M., Nik Yusof, N., Lee, C., Abu Bakar, S., AbuBakar, S., Hollox, E. and Boon Peng, H. (2015) 'A Comparison of Assays for Accurate Copy Number Measurement of the Low-Affinity Fc

Gamma Receptor Genes FCGR3A and FCGR3B.', *PloS one*, 10(1), p. e0116791.

Harwig, S. S., Swiderek, K. M., Kokryakov, V. N., Tan, L., Lee, T. D., Panyutich, E. A., Aleshina, G. M., Shamova, O. V and Lehrer, R. I. (1994) 'Gallinacins: cysteine-rich antimicrobial peptides of chicken leukocytes.', *FEBS letters*, Comparative Study, Journal Article, Research Support, U.S. Gov't, P.H.S., NETHERLANDS, 342(3), pp. 281–285.

Haslam, D. W. and James, W. P. T. (2005) 'Obesity.', *Lancet*, Elsevier, 366(9492), pp. 1197–209.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M. and Ira, G. (2009) 'Mechanisms of change in gene copy number.', *Nature reviews. Genetics*, 10(8), pp. 551–64.

Hauser, M. (2003) 'Exercise blood pressure in congenital heart disease and in patients after coarctation repair.', *Heart*, 89(2), pp. 125–126.

Haynes, R. J., Tighe, P. J. and Dua, H. S. (1999) 'Antimicrobial defensin peptides of the human ocular surface.', *The British journal of ophthalmology*, 83(6), pp. 737–41.

Hazlett, L. and Wu, M. (2011) 'Defensins in innate immunity.', *Cell and tissue research*, 343(1), pp. 175–88.

Hellmann, G. M., Fields, W. R. and Doolittle, D. J. (2001) 'Gene expression profiling of cultured human bronchial epithelial and lung carcinoma cells.', *Toxicological sciences*, 61(1), pp. 154–63.

Henrard, D., Phillips, J., Muenz, L., Blattner, W., Wiesner, D., Eyster, M. and Goedert, J. (1995) 'Natural history of HIV-1 cell-free viremia.', *JAMA*, 274(7), pp. 554–8.

Henrichsen, C. N., Vinckenbosch, N., Zollner, S., Chaignat, E., Pradervand, S., Schutz, F., Ruedi, M., Kaessmann, H. and Reymond, A. (2009) 'Segmental copy number variation shapes tissue transcriptomes.', *Nature genetics*, Journal Article, Research Support, N.I.H., Extramural, Research Support, Non-U.S. Gov't, United States, 41(4), pp. 424–429.

HGNC (n.d.) 'Defensins, beta (DEFB)', [online] Available from: http://www.genenames.org/cgi-bin/genefamilies/set/503.

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

Hills, A., Ahn, J., Donaghue, C., Thomas, H., Mann, K. and Ogilvie, C. (2010) 'MLPA for confirmation of array CGH results and determination of inheritance', *Molecular Cytogenetics*, BioMed Central Ltd, 3(1), p. 19.

Hindson, B., Ness, K. D., Masquelier, D. a, Belgrader, P., Heredia, N. J., Makarewicz, A. J., Bright, I. J., Lucero, M. Y., Hiddessen, A. L., Legler, T. C., Kitano, T. K., Hodel, M. R., Petersen, J. F., Wyatt, P. W., Steenblock, E. R., Shah, P. H., Bousse, L. J., Troup, C. B., Mellen, J. C., Wittmann, D. K., Erndt, N. G., Cauley, T. H., Koehler, R. T., So, A. P., Dube, S., Rose, K. a, Montesclaros, L., Wang, S., Stumbo, D. P., Hodges, S. P., Romine, S., Milanovich, F. P., White, H. E., Regan, J. F., Karlin-Neumann, G. a, Hindson, C. M., Saxonov, S. and Colston, B. W. (2011) 'High-throughput droplet digital PCR system for absolute quantitation of DNA copy number.', *Analytical chemistry*, 83(22), pp. 8604–10.

Hiratsuka, T., Mukae, H., Iiboshi, H., Ashitani, J., Nabeshima, K., Minematsu, T., Chino, N., Ihi, T., Kohno, S. and Nakazato, M. (2003) 'Increased concentrations of human beta-defensins in plasma and bronchoalveolar lavage fluid of patients with diffuse panbronchiolitis.', *Thorax*, Journal Article, Research Support, Non-U.S. Gov't, England, 58(5), pp. 425–430.

Hiratsuka, T., Nakazato, M., Ihi, T., Minematsu, T., Chino, N., Nakanishi, T., Shimizu, A., Kangawa, K. and Matsukura, S. (2000) 'Structural Analysis of Human β-Defensin-1 and Its Significance in Urinary Tract Infection', *Nephron*, JOUR, 85(1), pp. 34–40.

Hirst, R., Kadioglu, A., O'callaghan, C. and Andrew, P. (2004) 'The role of pneumolysin in pneumococcal pneumonia and meningitis.', *Clinical and experimental immunology*, 138(2), pp. 195–201.

Ho, D. (1996) 'Viral counts count in HIV infection.', *Science*, 272(5265), pp. 1124–5.

Hoberman, A., Greenfield, S., Mattoo, T., Keren, R., Mathews, R., Pohl, H., Kropp, B., Skoog, S., Nelson, C., Moxey-Mims, M., Chesney, R., Carpenter, M. and The RIVUR Trial Investigators (2014) 'Antimicrobial Prophylaxis for Children with Vesicoureteral Reflux', *New England Journal of Medicine*, 370(25), pp. 2367–2376.

van Hoek, M. (2014) 'Antimicrobial Peptides in Reptiles', *Pharmaceuticals*, 7(6), pp. 723–753.

Hollox, E. (2008) 'Copy number variation of beta-defensins and relevance to disease.', *Cytogenetic and genome research*, 123(1–4), pp. 148–55.

Hollox, E. (2010) 'Beta-defensins and Crohn's disease: confusion from counting copies.', *The American journal of gastroenterology*, 105(2), pp. 360–2.

Hollox, E. (2012) 'The challenges of studying complex and dynamic regions of the human genome.', In Feuk, L. (ed.), *Genomic Structural Variants, Methods and Protocols*, Springer, pp. 187–207.

Hollox, E., Armour, J. and Barber, J. (2003) 'Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster.', *American journal of human genetics*, 73(3), pp. 591–600.

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

Hollox, E., Davies, J., Griesenbach, U., Burgess, J., Alton, E. and Armour, J. (2005) 'Beta-defensin genomic copy number is not a modifier locus for cystic fibrosis.', *Journal of negative results in biomedicine*, 4, p. 9.

Hollox, E., Huffmeier, U., Zeeuwen, P., Palla, R., Lascorz, J., Rodijk-Olthuis, Kerkhof, van de, Traupe, H., de Jongh, G., den Heijer, M., Reis, A., Armour, J. and Schalkwijk, J. (2008) 'Psoriasis is associated with increased beta-defensin genomic copy number.', *Nature genetics*, 40(1), pp. 23–5.

Hollox, E. J., Akrami, S. M. and Armour, J. a L. (2002) 'DNA copy number analysis by MAPH: molecular diagnostic applications', *Expert Review of Molecular Diagnostics*, 2(4), pp. 370–378.

Hollox, E. J. and Hoh, B.-P. (2014) 'Human gene copy number variation and infectious disease', *Human Genetics*, 133(10), pp. 1217–1233.

von Horsten, H. H., Schafer, B. and Kirchhoff, C. (2004) 'SPAG11/isoform HE2C, an atypical anionic beta-defensin-like peptide.', *Peptides*, Comparative Study, Journal Article, Research Support, Non-U.S. Gov't, United States, 25(8), pp. 1223–1233.

Huntley, S., Baggott, D. M., Hamilton, A. T., Tran-Gyamfi, M., Yang, S., Kim, J., Gordon, L., Branscomb, E. and Stubbs, L. (2006) 'A comprehensive catalog of human KRAB-associated zinc finger genes: Insights into the evolutionary history of a large family of transcriptional repressors', *Genome Research*, 16(5), pp. 669–677.

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

IAVI (2006) 'IAVI Observational Epidemiology', [online] Available from: http://www.iavi.org/who-we-are/experts/our-studies/observational-epidemiology/observational-epidemiology?start=5.

Inoue, K. and Lupski, J. R. (2002) 'Molecular mechanisms for genomic disorders.', *Annual review of genomics and human genetics*, Journal Article, Research Support, Non-U.S. Gov't, Research Support, U.S. Gov't, P.H.S., Review, United States, 3, pp. 199–242.

International Human Genome Sequencing Consortium (2004) 'Finishing the euchromatic sequence of the human genome', *Nature*, 431(7011), pp. 931–945.

Jaillard, S., Drunat, S., Bendavid, C., Aboura, A., Etcheverry, A., Journel, H., Delahaye, A., Pasquier, L., Bonneau, D., Toutain, A., Burglen, L., Guichet, A., Pipiras, E., Gilbert-Dussardier, B., Benzacken, B., Martin-Coignard, D., Henry, C., David, A., Lucas, J., Mosser, J., David, V., Odent, S., Verloes, A. and Dubourg, C. (2010) 'Identification of gene copy number variations in patients with mental retardation using array-CGH: Novel syndromes in a large French series', *European Journal of Medical Genetics*, 53(2), pp. 66–75.

Jansen, P., Rodijk-Olthuis, D., Hollox, E., Kamsteeg, M., Tjabringa, G., de Jongh, G., van Vlijmen-Willems, I., Bergboer, J., van Rossum, M., de Jong, E., den Heijer, M., Evers, A., Bergers, M., Armour, J., Zeeuwen, P. and Schalkwijk, J. (2009) 'Beta-defensin-2 protein is a serum biomarker for disease activity in psoriasis and reaches biologically relevant concentrations in lesional skin.', *PloS one*, 4(3), p. e4725.

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

Janssens, W., Nuytten, H., Dupont, L. J., Van Eldere, J., Vermeire, S., Lambrechts, D., Nackaerts, K., Decramer, M., Cassiman, J.-J. and Cuppens, H. (2010) 'Genomic copy number determines functional expression of {beta}-defensin 2 in airway epithelial cells and associates with chronic obstructive pulmonary disease.', *American journal of respiratory and critical care medicine*, Journal Article, Research Support, Non-U.S. Gov't, United States, 182(2), pp. 163–9.

Jaradat, S. W., Cubillos, S., Krieg, N., Lehmann, K., Issa, B., Piehler, S., Wehner-Diab, S., Hipler, U.-C. and Norgauer, J. (2015) 'Low DEFB4 Copy Number and High Systemic hBD-2 and IL-22 Levels Are Associated with Dermatophytosis', *J Invest Dermatol*, Elsevier Masson SAS, 135(3), pp. 750–758.

Jaradat, S. W., Hoder-Przyrembel, C., Cubillos, S., Krieg, N., Lehmann, K., Piehler, S., Sigusch, B. W. and Norgauer, J. (2013) 'Beta-defensin-2 genomic copy number variation and chronic periodontitis.', *Journal of dental research*, 92(11), pp. 1035–40.

Jia, H., Schutte, B., Schudy, A., Linzmeier, R., Guthmiller, J., Johnson, G. K., Tack, B. F., Mitros, J. P., Rosenthal, A., Ganz, T. and McCray, P. B. (2001) 'Discovery of new human β-defensins using a genomics-based approach', *Gene*, 263(1–2), pp. 211–218.

Jones, E. a, Kananurak, A., Bevins, C. L., Hollox, E. J. and Bakaletz, L. O. (2014) 'Copy number variation of the beta defensin gene cluster on chromosome 8p influences the bacterial microbiota within the nasopharynx of otitis-prone children.', *PloS one*, 9(5), p. e98269.

Karlen, Y., McNair, A., Perseguers, S., Mazza, C. and Mermod, N. (2007) 'Statistical significance of quantitative PCR.', *BMC bioinformatics*, Journal Article, Research Support, Non-U.S. Gov't, Validation Studies, England, 8, p. 131.

Keren, R., Carpenter, M. a, Hoberman, A., Shaikh, N., Matoo, T. K., Chesney, R. W., Matthews, R., Gerson, A. C., Greenfield, S. P., Fivush, B., McLurie, G. a, Rushton, H. G., Canning, D., Nelson, C. P., Greenbaum, L., Bukowski, T., Primack, W., Sutherland, R., Hosking, J., Stewart, D., Elder, J., Moxey-Mims, M. and Nyberg, L. (2008) 'Rationale and design issues of the Randomized Intervention for Children With Vesicoureteral Reflux (RIVUR) study.', *Pediatrics*, 122 Suppl, pp. S240-50.

Khan, M. Z. (2014) 'Mechanism linking diabetes mellitus and obesity', pp. 587–591.

Kim, Y.-J., Shin, H.-S., Lee, J.-H., Jung, Y. W., Kim, H.-B. and Ha, U.-H. (2013) 'Pneumolysin-mediated expression of β-defensin 2 is coordinated by p38 MAP kinase-MKP1 in human airway cells.', *Journal of microbiology (Seoul, Korea)*, 51(2), pp. 194–9.

Kraus, D., Deschner, J., Jager, A., Wenghoefer, M., Bayer, S., Jepsen, S., Allam, J. P., Novak, N., Meyer, R. and Winter, J. (2012) 'Human beta-defensins differently affect proliferation, differentiation, and mineralization of osteoblast-like MG63 cells.', *Journal of cellular physiology*, Comparative Study, Journal Article, Research Support, Non-U.S. Gov't, United States, 227(3), pp. 994–1003.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L.,

Mercer, S., Milne, S., Mullikin, J., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R., Wilson, R., Hillier, L., McPherson, J., Marra, M., Mardis, E., Fulton, L., Chinwalla, A., Pepin, K., Gish, W., Chissoe, S., Wendl, M., Delehaunty, K., Miner, T., Delehaunty, A., Kramer, J., Cook, L., Fulton, R., Johnson, D., Minx, P., Clifton, S., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R., Muzny, D., Scherer, S., Bouck, J., Sodergren, E., Worley, K., Rives, C., Gorrell, J., Metzker, M., Naylor, S., Kucherlapati, R., Nelson, D., Weinstock, G., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M., Peterson, J., Felsenfeld, A., Wetterstrand, K., Myers, R., Schmutz, J., Dickson, M., Grimwood, J., Cox, D., Olson, M., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G., Athanasiou, M., Schultz, R., Patrinos, A. and Morgan, M. (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921.

Larsen, M. H., Thorner, L. W., Zinyama, R., Amstrup, J., Kallestrup, P., Gerstoft, J., Gomo, E., Erikstrup, C. and Ullum, H. (2012) 'CCL3L gene copy number and survival in an HIV-1 infected Zimbabwean population.', *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, Journal Article, Research Support, Non-U.S. Gov't, Netherlands, 12(5), pp. 1087–1093.

Lee, H.-Y., Andalibi, A., Webster, P., Moon, S.-K., Teufert, K., Kang, S.-H., Li, J.-D., Nagura, M.,

Ganz, T. and Lim, D. J. (2004) 'Antimicrobial activity of innate immune molecules against Streptococcus pneumoniae, Moraxella catarrhalis and nontypeable Haemophilus influenzae.', *BMC infectious diseases*, 4, p. 12.

Lehmann, J., Retz, M., Harder, J., Krams, M., Kellner, U., Hartmann, J., Hohgräwe, K., Raffenberg, U., Gerber, M., Loch, T., Weichert-jacobsen, K. and Stöckle, M. (2002) 'Expression of human bate-defensins 1 and 2 in kidneys with chronic bacterial infection', *BMC infectious Disease*, 10, pp. 1–10.

Levy, H., Raby, B. A., Lake, S., Tantisira, K. G., Kwiatkowski, D., Lazarus, R., Silverman, E. K., Richter, B., Klimecki, W. T., Vercelli, D., Martinez, F. D. and Weiss, S. T. (2005) 'Association of defensin beta-1 gene polymorphisms with asthma.', *The Journal of allergy and clinical immunology*, 115(2), pp. 252–8.

Li, D., Zhang, L., Yin, H., Xu, H., Satkoski Trask, J., Smith, D. G., Li, Y., Yang, M. and Zhu, Q. (2014) 'Evolution of primate α and θ defensins revealed by analysis of genomes.', *Molecular biology reports*, 41(6), pp. 3859–66.

Li, Wenli and Olivier, M. (2013) 'Current analysis platforms and methods for detecting copy number variation', *Physiological Genomics*, 45(1), pp. 1–16.

Li, W. and Olivier, M. (2013) 'Current analysis platforms and methods for detecting copy number variation', *Physiological Genomics*, 45(1), pp. 1–16.

Liao, Z., Dong, J., Hu, X., Wang, T., Wan, C., Li, X., Li, L., Guo, L., Xu, D. and Wen, F. (2012) 'Enhanced expression of human ??-defensin 2 in peripheral lungs of patients with chronic obstructive pulmonary disease', *Peptides*, 38(2), pp. 350–356.

Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., Amann, M., Anderson, H. R., Andrews, K. G., Aryee, M., Atkinson, C., Bacchus, L. J., Bahalim, A. N., Balakrishnan, K., Balmes, J., Barker-Collo, S., Baxter, A., Bell, M. L., Blore, J. D., Blyth, F., Bonner, C., Borges, G., Bourne, R., Boussinesq, M., Brauer, M., Brooks, P., Bruce, N. G., Brunekreef, B., Bryan-Hancock, C., Bucello, C., Buchbinder, R., Bull, F., Burnett, R. T., Byers, T. E., Calabria, B., Carapetis, J., Carnahan, E., Chafe, Z., Charlson, F., Chen, H., Chen, J. S., Cheng, A. T.-A., Child, J. C., Cohen, A., Colson, K. E., Cowie, B. C., Darby, S., Darling, S., Davis, A., Degenhardt, L., Dentener, F., Des Jarlais, D. C., Devries, K., Dherani, M., Ding, E. L., Dorsey, E. R., Driscoll, T., Edmond, K., Ali, S. E., Engell, R. E., Erwin, P. J., Fahimi, S., Falder, G., Farzadfar, F., Ferrari, A., Finucane, M. M., Flaxman, S., Fowkes, F. G. R., Freedman, G., Freeman, M. K., Gakidou, E., Ghosh, S., Giovannucci, E., Gmel, G., Graham,

K., Grainger, R., Grant, B., Gunnell, D., Gutierrez, H. R., Hall, W., Hoek, H. W., Hogan, A., Hosgood, H. D., Hoy, D., Hu, H., Hubbell, B. J., Hutchings, S. J., Ibeanusi, S. E., Jacklyn, G. L., Jasrasaria, R., Jonas, J. B., Kan, H., Kanis, J. A., Kassebaum, N., Kawakami, N., Khang, Y.-H., Khatibzadeh, S., Khoo, J.-P., Kok, C., Laden, F., Lalloo, R., Lan, Q., Lathlean, T., Leasher, J. L., Leigh, J., Li, Y., Lin, J. K., Lipshultz, S. E., London, S., Lozano, R., Lu, Y., Mak, J., Malekzadeh, R., Mallinger, L., Marcenes, W., March, L., Marks, R., Martin, R., McGale, P., McGrath, J., Mehta, S., Mensah, G. A., Merriman, T. R., Micha, R., Michaud, C., Mishra, V., Mohd Hanafiah, K., Mokdad, A. A., Morawska, L., Mozaffarian, D., Murphy, T., Naghavi, M., Neal, B., Nelson, P. K., Nolla, J. M., Norman, R., Olives, C., Omer, S. B., Orchard, J., Osborne, R., Ostro, B., Page, A., Pandey, K. D., Parry, C. D. H., Passmore, E., Patra, J., Pearce, N., Pelizzari, P. M., Petzold, M., Phillips, M. R., Pope, D., Pope, C. A., Powles, J., Rao, M., Razavi, H., Rehfuess, E. A., Rehm, J. T., Ritz, B., Rivara, F. P., Roberts, T., Robinson, C., Rodriguez-Portales, J. A., Romieu, I., Room, R., Rosenfeld, L. C., Roy, A., Rushton, L., Salomon, J. A., Sampson, U., Sanchez-Riera, L., Sanman, E., Sapkota, A., Seedat, S., Shi, P., Shield, K., Shivakoti, R., Singh, G. M., Sleet, D. A., Smith, E., Smith, K. R., Stapelberg, N. J. C., Steenland, K., Stöckl, H., Stovner, L. J., Straif, K., Straney, L., Thurston, G. D., Tran, J. H., Van Dingenen, R., van Donkelaar, A., Veerman, J. L., Vijayakumar, L., Weintraub, R., Weissman, M. M., White, R. A., Whiteford, H., Wiersma, S. T., Wilkinson, J. D., Williams, H. C., Williams, W., Wilson, N., Woolf, A. D., Yip, P., Zielinski, J. M., Lopez, A. D., Murray, C. J. L., Ezzati, M., AlMazroa, M. A. and Memish, Z. A. (2012) 'A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010.', *Lancet*, Elsevier, 380(9859), pp. 2224–60.

Linde, A., Ross, C., Blecha, F., Lushington, G. and Melgarejo, T. (2013) 'Cardiac beta-defensins upregulate with a high fat diet and influence monocyte migration *', *Open Journal of Internal Medicine*, 3, pp. 81–94.

Linzmeier, R., Ho, C. H., Hoang, B. V and Ganz, T. (1999) 'A 450-kb contig of defensin genes on human chromosome 8p23.', *Gene*, 233(1–2), pp. 205–11.

Linzmeier, R. M. and Ganz, T. (2005) 'Human defensin gene copy number polymorphisms: comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23.', *Genomics*, 86(4), pp. 423–30.

Liu, R., Zhang, Z., Liu, H., Hou, P., Lang, J., Wang, S., Yan, H., Li, P., Huang, Z., Wu, H., Rong, M., Huang, J., Wang, H., Lv, L., Qiu, M., Ding, J. and Lai, R. (2013) 'Human β-defensin 2 is a

novel opener of Ca2+-activated potassium channels and induces vasodilation and hypotension in monkeys.', *Hypertension*, 62(2), pp. 415–25.

Liu, S., Yao, L., Ding, D. and Zhu, H. (2010) 'CCL3L1 Copy Number Variation and Susceptibility to HIV-1 Infection: A Meta-Analysis', *PLoS ONE*, JOUR, Public Library of Science, 5(12), p. e15778.

Lupski, J. R. (1998) 'Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits', *Trends in Genetics*.

Machado, L. R., Bowdrey, J., Ngaimisi, E., Habtewold, A., Minzi, O., Makonnen, E., Yimer, G., Amogne, W., Mugusi, S., Janabi, M., Aderaye, G., Mugusi, F., Viskaduraki, M., Aklillu, E. and Hollox, E. J. (2013) 'Copy Number Variation of Fc Gamma Receptor Genes in HIV-Infected and HIV-Tuberculosis Co-Infected Individuals in Sub-Saharan Africa', He, W. (ed.), *PLoS ONE*, 8(11), p. e78165.

Machado, L. R. and Ottolini, B. (2015) 'An evolutionary history of defensins: a role for copy number variation in maximizing host innate and adaptive immune responses.', *Frontiers in immunology*, 6(March), p. 115.

Mackewicz, C. E., Yuan, J., Tran, P., Diaz, L., Mack, E., Selsted, M. E. and Levy, J. A. (2003) 'alpha-Defensins can have anti-HIV activity but are not CD8 cell anti-HIV factors.', *AIDS (London, England)*, 17(14), pp. F23-32.

Marques, F. Z., Prestes, P. R., Pinheiro, L. B., Scurrah, K., Emslie, K. R., Tomaszewski, M., Harrap, S. B. and Charchar, F. J. (2014) 'Measurement of absolute copy number variation reveals association with essential hypertension', *BMC Medical Genomics*, 7(1), p. 44.

Marshall, C. R. and Scherer, S. W. (2012) 'Detection and characterization of copy number variation in autism spectrum disorder.', *Methods in molecular biology (Clifton, N.J.)*, Journal Article, Review, United States, 838, pp. 115–135.

Maxwell, A. ., Morrison, G. . and Dorin, J. . (2003) 'Rapid sequence divergence in mammalian β-defensins by adaptive evolution', *Molecular Immunology*, 40(7), pp. 413–421.

McCarroll, S. a and Altshuler, D. M. (2007) 'Copy-number variation and association studies of human disease.', *Nature genetics*, 39(7 Suppl), pp. S37-42.

McCray, P. B. and Bentley, L. (1997) 'Human airway epithelia express a beta-defensin.', *American Journal of Respiratory Cell and Molecular Biology*, 16(3), pp. 343–349.

McLaren, P. J. and Carrington, M. (2015) 'The impact of host genetic variation on infection with

HIV-1', *Nature Immunology*, 16(6), pp. 577–583.

Medvedev, P., Stanciu, M. and Brudno, M. (2009) 'Computational methods for discovering structural variation with next-generation sequencing', *Nature Methods*, 6(11s), pp. S13–S20.

Medzhitov, R. and Janeway, C. J. (2000) 'Innate immunity.', *The New England journal of medicine*, Journal Article, Research Support, Non-U.S. Gov't, Research Support, U.S. Gov't, P.H.S., Review, UNITED STATES, 343(5), pp. 338–344.

Mefford, H. C., Muhle, H., Ostertag, P., von Spiczak, S., Buysse, K., Baker, C., Franke, A., Malafosse, A., Genton, P., Thomas, P., Gurnett, C. A., Schreiber, S., Bassuk, A. G., Guipponi, M., Stephani, U., Helbig, I. and Eichler, E. E. (2010) 'Genome-Wide Copy Number Variation in Epilepsy: Novel Susceptibility Loci in Idiopathic Generalized and Focal Epilepsies', *PLoS Genet*, JOUR, Public Library of Science, 6(5), p. e1000962.

Mehlotra, R., Dazard, J. E., John, B., Zimmerman, P., Weinberg, A. and Jurevic, R. (2012) 'Copy Number Variation within Human β-Defensin Gene Cluster Influences Progression to AIDS in the Multicenter AIDS Cohort Study', *Journal of AIDS & Clinical Research*, 3(10), pp. 1199–1216.

Mehlotra, R., Zimmerman, P., Weinberg, A. and Jurevic, R. (2012) 'Variation in human β-defensin genes: new insights from a multi-population study.', *International journal of immunogenetics*.

Mellors, J. W., Muñoz, A., Giorgi, J. V, Margolick, J. B., Tassoni, C. J., Gupta, P., Kingsley, L. A., Todd, J. A., Saah, A. J., Detels, R., Phair, J. P. and Rinaldo, C. R. (1997) 'Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection.', *Annals of internal medicine*, 126(12), pp. 946–54.

Milanese, M., Segat, L., Arraes, L. C., Garzino-Demo, A. and Crovella, S. (2009) 'Copy number variation of defensin genes and HIV infection in Brazilian children.', *Journal of acquired immune deficiency syndromes*, Journal Article, Research Support, Non-U.S. Gov't, United States, 50(3), pp. 331–333.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, Kai, Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y. K., Leng, J., Li, R., Li, Y., Lin, C.-Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stütz, A. M.,

Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, Kenny, Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. A. and Korbel, J. O. (2011) 'Mapping copy number variation by population-scale genome sequencing.', *Nature*, 470(7332), pp. 59–65.

Moffatt, M. F., Gut, I. G., Demenais, F., Strachan, D. P., Bouzigon, E., Heath, S., von Mutius, E., Farrall, M., Lathrop, M., Cookson, W. O. C. M. and GABRIEL Consortium (2010) 'A large-scale, consortium-based genomewide association study of asthma.', *The New England journal of medicine*, 363(13), pp. 1211–21.

Montague, C. T., Farooqi, I. S., Whitehead, J. P., Soos, M. A., Rau, H., Wareham, N. J., Sewter, C. P., Digby, J. E., Mohammed, S. N., Hurst, J. A., Cheetham, C. H., Earley, A. R., Barnett, A. H., Prins, J. B. and O'Rahilly, S. (1997) 'Congenital leptin deficiency is associated with severe early-onset obesity in humans.', *Nature*, Case Reports, Journal Article, Research Support, Non-U.S. Gov't, ENGLAND, 387(6636), pp. 903–908.

Moorthy, I., Wheat, D. and Gordon, I. (2004) 'Ultrasonography in the evaluation of renal scarring using DMSA scan as the gold standard.', *Pediatric nephrology*, Comparative Study, Evaluation Studies, Journal Article, Germany, 19(2), pp. 153–156.

Morrison, G., Kilanowski, F., Davidson, D. and Dorin, J. (2002) 'Characterization of the mouse [beta]-defensin 1, Defb1, mutant mouse model', *Infect. Immun.*, 70(6), pp. 3053–3060.

Nakashima, H., Yamamoto, N., Masuda, M. and Fujii, N. (1993) 'Defensins inhibit HIV replication in vitro.', *AIDS (London, England)*, 7(8), p. 1129.

NanoString Technologies (2013) 'nCounter Workflow', [online] Available from: http://www.nanostring.com/products/workflow (Accessed 2 June 2013).

Nguewa, P. a, Agorreta, J., Blanco, D., Lozano, M. D., Gomez-Roman, J., Sanchez, B. a, Valles, I., Pajares, M. J., Pio, R., Rodriguez, M. J., Montuenga, L. M. and Calvo, A. (2008) 'Identification of importin 8 (IPO8) as the most accurate reference gene for the clinicopathological analysis of lung specimens.', *BMC molecular biology*, 9, p. 103.

Nguyen, T. X., Cole, A. M. and Lehrer, R. I. (2003) 'Evolution of primate theta-defensins: a serpentine path to a sweet tooth.', *Peptides*, Journal Article, Research Support, U.S. Gov't, P.H.S., United States, 24(11), pp. 1647–1654.

Nix, M. A., Kaelin, C. B., Ta, T., Weis, A., Morton, G. J., Barsh, G. S. and Millhauser, G. L. (2013) 'Molecular and Functional Analysis of Human β-Defensin 3 Action at Melanocortin

Receptors', *Chemistry & Biology*, 20(6), pp. 784–795.

Nuytten, H., Wlodarska, I., Nackaerts, K., Vermeire, S., Vermeesch, J., Cassiman, J.-J. and Cuppens, H. (2009) 'Accurate determination of copy number variations (CNVs): Application to the α- and β-defensin CNVs', *Journal of Immunological Methods*, Comparative Study, Evaluation Studies, Journal Article, Research Support, Non-U.S. Gov't, Netherlands, 344(1), pp. 35–44.

O'Rahilly, S. (2009) 'Human genetics illuminates the paths to metabolic disease', *Nature*, JOUR, Macmillan Publishers Limited. All rights reserved, 462(7271), pp. 307–314.

Oswald, N. (2007) 'The Invention of PCR', *BitesizeBio*, [online] Available from: http://bitesizebio.com/13505/the-invention-of-pcr/ (Accessed 1 October 2015).

Ottolini, B., Hornsby, M., Abujaber, R., MacArthur, J., Badge, R., Schwarzacher, T., Albertson, D., Bevins, C., Solnick, J. and Hollox, E. (2014) 'Evidence of convergent evolution in humans and macaques supports an adaptive role for copy number variation of the β-defensin-2 gene.', *Genome biology and evolution*, 6(11), pp. 3025–38.

Ouahchi, K., Lindeman, N. and Lee, C. (2006) 'Copy number variants and pharmacogenomics', *Pharmacogenomics*, 7(1), pp. 25–29.

Page, R. A. and Malik, A. N. (2003) 'Elevated levels of beta defensin-1 mRNA in diabetic kidneys of GK rats.', *Biochemical and biophysical research communications*, Journal Article, Research Support, Non-U.S. Gov't, United States, 310(2), pp. 513–521.

Palmer, L. J., Celed??n, J. C., Chapman, H. A., Speizer E., F. E., Weiss, S. T. and Silverman, E. K. (2003) 'Genome-wide linkage analysis of bronchodilator responsiveness and post-bronchodilator spirometric phenotypes in chronic obstructive pulmonary disease', *Human Molecular Genetics*, 12(10), pp. 1199–1210.

Park, J. J., Oh, B. R., Kim, J. Y., Park, J. A., Kim, C., Lee, Y. J., Song, Y. W., Armour, J. A. L. and Lee, E. B. (2011) 'Copy number variation of β-defensin genes in Behçet's disease.', *Clinical and experimental rheumatology*, 29(4 Suppl 67), pp. S20-3.

Pazgier, M., Hoover, D. M., Yang, D., Lu, W. and Lubkowski, J. (2006) 'Human beta-defensins.', *Cellular and molecular life sciences : CMLS*, 63(11), pp. 1294–313.

Pedersen, K., Wiechec, E., Madsen, B. E., Overgaard, J. and Hansen, L. (2010) 'A simple way to evaluate self-designed probes for tumor specific Multiplex Ligation-dependent Probe Amplification (MLPA)', *BMC Research Notes*, 3(1), p. 179.

Pelak, K., Need, A. C., Fellay, J., Shianna, K. V, Feng, S., Urban, T. J., Ge, D., De Luca, A., Martinez-Picado, J., Wolinsky, S. M., Martinson, J. J., Jamieson, B. D., Bream, J. H., Martin, M. P., Borrow, P., Letvin, N. L., McMichael, A. J., Haynes, B. F., Telenti, A., Carrington, M., Goldstein, D. B. and Alter, G. (2011) 'Copy number variation of KIR genes influences HIV-1 control.', *PLoS biology*, 9(11), p. e1001208.

Pennisi, E. (2012) 'ENCODE Project Writes Eulogy for Junk DNA', *Science*, 337(6099), pp. 1159–1161.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C. and Stone, A. C. (2007) 'Diet and the evolution of human amylase gene copy number variation.', *Nature genetics*, 39(10), pp. 1256–60.

Petrovski, S., Fellay, J., Shianna, K. V, Carpenetti, N., Kumwenda, J., Kamanga, G., Kamwendo, D. D., Letvin, N. L., McMichael, A. J., Haynes, B. F., Cohen, M. S. and Goldstein, D. B. (2011) 'Common human genetic variants and HIV-1 susceptibility: a genome-wide survey in a homogeneous African population.', *AIDS*, 25(4), pp. 513–518.

Pierson, T. C. and Doms, R. W. (2003) 'HIV-1 entry and its inhibition.', *Current topics in microbiology and immunology*, Journal Article, Review, Germany, 281, pp. 1–27.

Pinheiro, L. B., Coleman, V. a, Hindson, C. M., Herrmann, J., Hindson, B. J., Bhat, S. and Emslie, K. R. (2012) 'Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification.', *Analytical chemistry*, 84(2), pp. 1003–11.

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A. C., Thiruvahindrapuram, B., Macdonald, J. R., Mills, R., Prasad, A., Noonan, K., Gribble, S., Prigmore, E., Donahoe, P. K., Smith, R. S., Park, J. H., Hurles, M. E., Carter, N. P., Lee, C., Scherer, S. W. and Feuk, L. (2011) 'Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants.', *Nature biotechnology*, 29(6), pp. 512–520.

Piotrowski, A., Bruder, C. E. G., Andersson, R., de Ståhl, T. D., Menzel, U., Sandgren, J., Poplawski, A., von Tell, D., Crasto, C., Bogdan, A., Bartoszewski, R., Bebok, Z., Krzyzanowski, M., Jankowski, Z., Partridge, E. C., Komorowski, J. and Dumanski, J. P. (2008) 'Somatic mosaicism for copy number variation in differentiated human tissues', *Human Mutation*, 29(9), pp. 1118–1124.

Polan, M. B., Pastore, M. T., Steingass, K., Hashimoto, S., Thrush, D. L., Pyatt, R., Reshmi, S.,

Gastier-Foster, J. M., Astbury, C. and McBride, K. L. (2014) 'Neurodevelopmental disorders among individuals with duplication of 4p13 to 4p12 containing a GABAA receptor subunit gene cluster', *European Journal of Human Genetics*, 22(1), pp. 105–109.

Pothiwala, P., Jain, S. K. and Yaturu, S. (2009) 'Metabolic syndrome and cancer.', *Metabolic syndrome and related disorders*, 7(4), pp. 279–288.

Quiñones-Mateu, M., Lederman, M., Feng, Z., Chakraborty, B., Weber, J., Rangel, H., Marotta, M., Mirza, M., Jiang, B., Kiser, P., Medvik, K., Sieg, S. and Weinberg, A. (2003) 'Human epithelial beta-defensins 2 and 3 inhibit HIV-1 replication.', *AIDS (London, England)*, 17(16), pp. F39–F48.

Raap, a K. (1998) 'Advances in fluorescence in situ hybridization.', *Mutation research*, 400(1–2), pp. 287–98.

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. J., Zerjal, T., Zhang, J. J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W. and Hurles, M. E. (2006) 'Global variation in copy number in the human genome', *Nature*, 444(7118), pp. 444–454.

Repapi, E., Sayers, I., Wain, L. V, Burton, P. R., Johnson, T., Obeidat, M., Zhao, J. H., Ramasamy, A., Zhai, G., Vitart, V., Huffman, J. E., Igl, W., Albrecht, E., Deloukas, P., Henderson, J., Granell, R., McArdle, W. L., Rudnicka, A. R., Barroso, I., Loos, R. J. F., Wareham, N. J., Mustelin, L., Rantanen, T., Surakka, I., Imboden, M., Wichmann, H. E., Grkovic, I., Jankovic, S., Zgaga, L., Hartikainen, A.-L., Peltonen, L., Gyllensten, U., Johansson, A., Zaboli, G., Campbell, H., Wild, S. H., Wilson, J. F., Gläser, S., Homuth, G., Völzke, H., Mangino, M., Soranzo, N., Spector, T. D., Polasek, O., Rudan, I., Wright, A. F., Heliövaara, M., Ripatti, S., Pouta, A., Naluai, A. T., Olin, A.-C., Torén, K., Cooper, M. N., James, A. L., Palmer, L. J., Hingorani, A. D., Wannamethee, S. G., Whincup, P. H., Smith, G. D., Ebrahim, S., McKeever, T. M., Pavord, I. D., MacLeod, A. K., Morris, A. D., Porteous, D. J., Cooper, C., Dennison, E., Shaheen, S., Karrasch, S., Schnabel, E., Schulz, H., Grallert, H., Bouatia-Naji, N., Delplanque, J., Froguel, P., Blakey, J. D., Britton, J. R., Morris, R. W., Holloway, J. W., Lawlor, D. A., Hui, J., Nyberg, F., Jarvelin, M.-R., Jackson, C., Kähönen, M., Kaprio, J., Probst-Hensch, N. M., Koch, B., Hayward, C., Evans, D. M., Elliott, P., Strachan, D. P., Hall,

I. P. and Tobin, M. D. (2010) ‘Genome-wide association study identifies five loci associated with lung function.’, *Nature genetics*, 42(1), pp. 36–44.

Ribeiro, C. M., Romano, R. M. and Avellar, M. C. W. (2012) ‘Beta-defensins in the epididymis : clues to multifunctional roles’, *Anim.Reprod*, 9(4), pp. 751–759.

Rodríguez-Jiménez, F.-J., Krause, A., Schulz, S., Forssmann, W.-G., Conejo-Garcia, J.-R., Schreeb, R. and Motzkus, D. (2003) ‘Distribution of new human β-defensin genes clustered on chromosome 20 in functionally different segments of epididymis’, *Genomics*, 81(2), pp. 175–183.

Rogers, M. F., White, C. R., Sanders, R., Schable, C., Ksell, T. E., Wasserman, R. L., Bellanti, J. A., Peters, S. M. and Wray, B. B. (1990) ‘Lack of transmission of human immunodeficiency virus from infected children to their household contacts.’, *Pediatrics*, Journal Article, UNITED STATES, 85(2), pp. 210–214.

Rutledge, R. G. and Cote, C. (2003) ‘Mathematics of quantitative kinetic PCR and the application of standard curves.’, *Nucleic acids research*, Journal Article, Research Support, Non-U.S. Gov’t, England, 31(16), p. e93.

Sailani, M. R., Makrythanasis, P., Valsesia, A., Santoni, F. A., Deutsch, S., Popadin, K., Borel, C., Migliavacca, E., Sharp, A. J., Duriaux Sail, G., Falconnet, E., Rabionet, K., Serra-Juhé, C., Vicari, S., Laux, D., Grattau, Y., Dembour, G., Megarbane, A., Touraine, R., Stora, S., Kitsiou, S., Fryssira, H., Chatzisevastou-Loukidou, C., Kanavakis, E., Merla, G., Bonnet, D., Pérez-Jurado, L. A., Estivill, X., Delabar, J. M., Antonarakis, S. E., Serra-Juhe, C., Vicari, S., Laux, D., Grattau, Y., Dembour, G., Megarbane, A., Touraine, R., Stora, S., Kitsiou, S., Fryssira, H., Chatzisevastou-Loukidou, C., Kanavakis, E., Merla, G., Bonnet, D., Perez-Jurado, L. A., Estivill, X., Delabar, J. M., Antonarakis, S. E., Serra-Juhé, C., Vicari, S., Laux, D., Grattau, Y., Dembour, G., Megarbane, A., Touraine, R., Stora, S., Kitsiou, S., Fryssira, H., Chatzisevastou-Loukidou, C., Kanavakis, E., Merla, G., Bonnet, D., Pérez-Jurado, L. A., Estivill, X., Delabar, J. M. and Antonarakis, S. E. (2013) ‘The complex SNP and CNV genetic architecture of the increased risk of congenital heart defects in Down syndrome.’, *Genome research*, 23(9), pp. 1410–21.

Salm, M. P. A., Horswell, S. D., Hutchison, C. E., Speedy, H. E., Yang, X., Liang, L., Schadt, E. E., Cookson, W. O., Wierzbicki, A. S., Naoumova, R. P. and Shoulders, C. C. (2012) ‘The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism.’, *Genome research*, 22(6), pp. 1144–53.

Scharf, S., Zahlten, J., Szymanski, K., Hippenstiel, S., Suttorp, N. and N'Guessan, P. D. (2012) 'Streptococcus pneumoniae induces human β-defensin-2 and -3 in human lung epithelium', *Experimental Lung Research*, 38(2), pp. 100–110.

Schneider, G. F. and Dekker, C. (2012) 'DNA sequencing with nanopores', *Nature Biotechnology*, 30(4), pp. 326–328.

Schneider, J. J., Unholzer, A., Schaller, M., Schäfer-Korting, M. and Korting, H. C. (2005) 'Human defensins.', *Journal of molecular medicine (Berlin, Germany)*, 83(8), pp. 587–95.

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

Schröder, J. and Harder, J. (1999) 'Human beta-defensin-2', *The international journal of biochemistry & cell …*, 31, pp. 645–651.

ScienCell Research Laboratories (2016) 'Human Bronchial Epithelial Cells', [online] Available from: http://www.sciencellonline.com/PS/3210.pdf.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A. and Wigler, M. (2004) 'Large-scale copy number polymorphism in the human genome.', *Science*, 305(5683), pp. 525–8.

Semple, C. A. M., Rolfe, M. and Dorin, J. R. (2003) 'Duplication and selection in the evolution of primate beta-defensin genes.', *Genome biology*, BioMed Central Ltd, 4(5), p. R31.

Semple, F. and Dorin, J. R. (2012) 'ß-Defensins: Multifunctional Modulators of Infection, Inflammation and More?', *Journal of Innate Immunity*, 4(4), pp. 337–348.

Seo, S., Ahn, S., Hong, C. and Ro, B. (2001) 'Expressions of beta-defensins in human keratinocyte cell lines.', *Journal of dermatological science*, 27(3), pp. 183–91.

Shaikh, N., Morone, N. E., Bost, J. E. and Farrell, M. H. (2008) 'Prevalence of urinary tract infection in childhood: a meta-analysis.', *The Pediatric infectious disease journal*, 27(4), pp. 302–308.

Sharp, P. M. and Hahn, B. H. (2011) 'Origins of HIV and the AIDS pandemic.', *Cold Spring Harbor perspectives in medicine*, 1(1), p. a006841.

Shlien, A. and Malkin, D. (2009) 'Copy number variations and cancer', *Genome Medicine*, 1(6), p. 62.

Silverman, E. K., Chapman, H. A., Drazen, J. M., Weiss, S. T., Rosner, B., Campbell, E. J., O'Donnell, W. J., Reilly, J. J., Ginns, L., Mentzer, S., Wain, J. and Speizer, F. E. (1998) 'Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease: Risk to relatives for airflow obstruction and chronic bronchitis', *American Journal of Respiratory and Critical Care Medicine*, 157(6 I), pp. 1770–1778.

Silverman, E. K., Mosley, J. D., Palmer, L. J., Barth, M., Senter, J. M., Brown, A., Drazen, J. M., Kwiatkowski, D. J., Chapman, H. a, Campbell, E. J., Province, M. a, Rao, D. C., Reilly, J. J., Ginns, L. C., Speizer, F. E. and Weiss, S. T. (2002) 'Genome-wide linkage analysis of severe, early-onset chronic obstructive pulmonary disease: airflow obstruction and chronic bronchitis phenotypes.', *Human molecular genetics*, 11(6), pp. 623–32.

Small, K. and Warren, S. T. (1998) 'Emerin deletions occurring on both Xq28 inversion backgrounds.', *Human molecular genetics*, Journal Article, Research Support, Non-U.S. Gov't, ENGLAND, 7(1), pp. 135–139.

Soler Artigas, M., Loth, D. W., Wain, L. V, Gharib, S. A., Obeidat, M., Tang, W., Zhai, G., Zhao, J. H., Smith, A. V., Huffman, J. E., Albrecht, E., Jackson, C. M., Evans, D. M., Cadby, G., Fornage, M., Manichaikul, A., Lopez, L. M., Johnson, T., Aldrich, M. C., Aspelund, T., Barroso, I., Campbell, H., Cassano, P. A., Couper, D. J., Eiriksdottir, G., Franceschini, N., Garcia, M., Gieger, C., Gislason, G. K., Grkovic, I., Hammond, C. J., Hancock, D. B., Harris, T. B., Ramasamy, A., Heckbert, S. R., Heliövaara, M., Homuth, G., Hysi, P. G., James, A. L., Jankovic, S., Joubert, B. R., Karrasch, S., Klopp, N., Koch, B., Kritchevsky, S. B., Launer, L. J., Liu, Y., Loehr, L. R., Lohman, K., Loos, R. J. F., Lumley, T., Al Balushi, K. A., Ang, W. Q., Barr, R. G., Beilby, J., Blakey, J. D., Boban, M., Boraska, V., Brisman, J., Britton, J. R., Brusselle, G. G., Cooper, C., Curjuric, I., Dahgam, S., Deary, I. J., Ebrahim, S., Eijgelsheim, M., Francks, C., Gaysina, D., Granell, R., Gu, X., Hankinson, J. L., Hardy, R., Harris, S. E., Henderson, J., Henry, A., Hingorani, A. D., Hofman, A., Holt, P. G., Hui, J., Hunter, M. L., Imboden, M., Jameson, K. A., Kerr, S. M., Kolcic, I., Kronenberg, F., Liu, J. Z., Marchini, J., McKeever, T., Morris, A. D., Olin, A.-C., Porteous, D. J., Postma, D. S., Rich, S. S., Ring, S. M., Rivadeneira, F., Rochat, T., Sayer, A. A., Sayers, I., Sly, P. D., Smith, G. D., Sood, A., Starr, J. M., Uitterlinden, A. G., Vonk, J. M., Wannamethee, S. G., Whincup, P. H., Wijmenga, C., Williams, O. D., Wong, A., Mangino, M., Marciante, K. D., McArdle, W. L., Meibohm, B., Morrison, A. C., North, K. E., Omenaas, E., Palmer, L. J., Pietiläinen, K. H., Pin, I., Pola Sbreve Ek, O., Pouta, A., Psaty, B. M., Hartikainen, A.-L., Rantanen, T., Ripatti, S., Rotter, J. I., Rudan, I., Rudnicka, A. R., Schulz, H., Shin, S.-Y., Spector, T. D., Surakka, I.,

Vitart, V., Völzke, H., Wareham, N. J., Warrington, N. M., Wichmann, H.-E., Wild, S. H., Wilk, J. B., Wjst, M., Wright, A. F., Zgaga, L., Zemunik, T., Pennell, C. E., Nyberg, F., Kuh, D., Holloway, J. W., Boezen, H. M., Lawlor, D. A., Morris, R. W., Probst-Hensch, N., Kaprio, J., Wilson, J. F., Hayward, C., Kähönen, M., Heinrich, J., Musk, A. W., Jarvis, D. L., Gläser, S., Järvelin, M.-R., Ch Stricker, B. H., Elliott, P., O'Connor, G. T., Strachan, D. P., London, S. J., Hall, I. P., Gudnason, V. and Tobin, M. D. (2011) 'Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function.', *Nature genetics*, 43(11), pp. 1082–90.

Spencer, J. D., Schwaderer, A., McHugh, K., Vanderbrink, B., Becknell, B. and Hains, D. S. (2011) 'The demographics and costs of inpatient vesicoureteral reflux management in the USA', *Pediatric Nephrology*, 26(11), pp. 1995–2001.

Spiess, A.-N. (2014) 'Package "qpcR"', [online] Available from: https://cran.r-project.org/web/packages/qpcR/qpcR.pdf.

Strandberg, T. E. and Pitkala, K. (2003) 'What is the most important component of blood pressure: systolic, diastolic or pulse pressure?', *Current opinion in nephrology and hypertension*, Journal Article, Research Support, Non-U.S. Gov't, Review, England, 12(3), pp. 293–297.

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. and Dermitzakis, E. T. (2007) 'Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes', *Science*, 315(5813), pp. 848–853.

Stuart, P. E., Hüffmeier, U., Nair, R. P., Palla, R., Tejasvi, T., Schalkwijk, J., Elder, J. T., Reis, A. and Armour, J. A. L. (2012) 'Association of β-Defensin Copy Number and Psoriasis in Three Cohorts of European Origin', *Journal of Investigative Dermatology*, 132(10), pp. 2407–2413.

Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B. P., Baker, C., Nordenfelt, S., Bamshad, M., Jorde, L. B., Posukh, O. L., Sahakyan, H., Watkins, W. S., Yepiskoposyan, L., Abdullah, M. S., Bravi, C. M., Capelli, C., Hervig, T., Wee, J. T. S., Tyler-Smith, C., van Driem, G., Romero, I. G., Jha, A. R., Karachanak-Yankova, S., Toncheva, D., Comas, D., Henn, B., Kivisild, T., Ruiz-Linares, A., Sajantila, A., Metspalu, E., Parik, J., Villems, R., Starikovskaya, E. B., Ayodo, G., Beall, C. M., Di Rienzo, A., Hammer, M. F., Khusainova, R., Khusnutdinova, E., Klitz, W., Winkler, C., Labuda, D., Metspalu, M.,

Tishkoff, S. A., Dryomov, S., Sukernik, R., Patterson, N., Reich, D. and Eichler, E. E. (2015) 'Global diversity, population stratification, and selection of human copy-number variation', *Science*, 349(6253), p. aab3761-aab3761.

Sudmant, P., Kitzman, J., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J. and Eichler, E. (2010) 'Diversity of Human Copy Number Variation and Multicopy Genes', *Science*, 330(6004), pp. 641–646.

Sudmant, P., Rausch, T., Gardner, E., Handsaker, R., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M., Malhotra, A., Stütz, A., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J., Kong, Y., Lameijer, E., McCarthy, S., Flicek, P., Gibbs, R., Marth, G., Mason, C., Menelaou, A., Muzny, D., Nelson, B., Noor, A., Parrish, N., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A., Untergasser, A., Walker, J., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M., McCarroll, S., Mills, R., Gerstein, M., Bashir, A., Stegle, O., Devine, S., Lee, C., Eichler, E. and Korbel, J. (2015) 'An integrated map of structural variation in 2,504 human genomes', *Nature*, 526(7571), pp. 75–81.

Sugawara, H., Harada, N., Ida, T., Ishida, T., Ledbetter, D. H., Yoshiura, K., Ohta, T., Kishino, T., Niikawa, N. and Matsumoto, N. (2003) 'Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23', *Genomics*, 82(2), pp. 238–244.

Sun, L., Finnegan, C., Kish-Catalone, T., Blumenthal, R., Garzino-Demo, P., La Terra Maggiore, G., Berrone, S., Kleinman, C., Wu, Z., Abdelwahab, S., Lu, W. and Garzino-Demo, A. (2005) 'Human beta-defensins suppress human immunodeficiency virus infection: potential role in mucosal protection.', *Journal of virology*, 79(22), pp. 14318–14329.

Swiss HIV Cohort Study (2015) 'About SHCS', [online] Available from: http://www.shcs.ch/157-about-shcs.

Taudien, S., Gäbel, G., Kuss, O., Groth, M., Grützmann, R., Huse, K., Kluttig, A., Wolf, A., Nothnagel, M., Rosenstiel, P., Greiser, K. H., Werdan, K., Krawczak, M., Pilarsky, C. and Platzer, M. (2012) 'Association studies of the copy-number variable ß-defensin cluster on 8p23.1 in adenocarcinoma and chronic pancreatitis.', *BMC research notes*, 5, p. 629.

Taudien, S., Huse, K., Groth, M. and Platzer, M. (2014) 'Narrowing down the distal border of the copy number variable beta-defensin gene cluster on human 8p23.', *BMC research notes*, BMC Research Notes, 7(1), p. 93.

The International HapMap Consortium (2003) 'The International HapMap Project', *Nature*, 426, pp. 789–796.

The Lipid Reaserch Clinics Program Epidemiology Committee (1979) 'Plasma lipid distributions in selected North American populations: The Lipid Reaserch Clinics Program Prevalence Study', *Circulation*, 60, pp. 427–439.

Tobin, M. D., Sheehan, N. A., Scurrah, K. J. and Burton, P. R. (2005) 'Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure.', *Statistics in medicine*, Comparative Study, Journal Article, Research Support, Non-U.S. Gov't, England, 24(19), pp. 2911–2935.

Tollner, T. L., Venners, S. A., Hollox, E. J., Yudin, A. I., Liu, X., Tang, G., Xing, H., Kays, R. J., Lau, T., Overstreet, J. W., Xu, X., Bevins, C. L. and Cherr, G. N. (2011) 'A Common Mutation in the Defensin DEFB126 Causes Impaired Sperm Function and Subfertility', *Science Translational Medicine*, 3(92), p. 92ra65-92ra65.

Tollner, T. L., Yudin, A. I., Tarantal, A. F., Treece, C. A., Overstreet, J. W. and Cherr, G. N. (2008) 'Beta-defensin 126 on the surface of macaque sperm mediates attachment of sperm to oviductal epithelia.', *Biology of reproduction*, Journal Article, United States, 78(3), pp. 400–412.

Tomaszewski, M., Charchar, F. J., Barnes, T., Gawron-kiszka, M., Sedkowska, A., Podolecka, E., Kowalczyk, J., Rathbone, W., Grzeszczak, W., Goodall, A. H. and Samani, N. J. (2009) 'A common variant in low density lipoprotein receptor-related protein 6 gene (LRP6) is associated with LDL-cholesterol', *Arterioscler Thromb Vasc Biol*, 29(9), pp. 1–12.

Tomaszewski, M., Charchar, F. J., Maric, C., McClure, J., Crawford, L., Grzeszczak, W., Sattar, N., Zukowska-Szczechowska, E. and Dominiczak, a F. (2007) 'Glomerular hyperfiltration: A new marker of metabolic risk', *Kidney International*, 71(8), pp. 816–821.

Townson, J. R., Barcellos, L. F. and Nibbs, R. J. B. (2002) 'Gene copy number regulates the production of the human chemokine CCL3-L1.', *European journal of immunology*, 32(10), pp. 3016–26.

Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M. I., Kidd, K. K. and Lenz, J. (2001)

'Insertional polymorphisms of full-length endogenous retroviruses in humans', *Current Biology*, 11(19), pp. 1531–1535.

Ulm, H., Wilmes, M., Shai, Y. and Sahl, H.-G. (2012) 'Antimicrobial Host Defensins – Specific Antibiotic Activities and Innate Defense Modulation', *Frontiers in Immunology*, 3.

Valore, E. V, Park, C. H., Quayle, a J., Wiles, K. R., McCray, P. B. and Ganz, T. (1998) 'Human beta-defensin-1: an antimicrobial peptide of urogenital tissues.', *The Journal of clinical investigation*, 101(8), pp. 1633–1642.

Veal, C. D., Xu, H., Reekie, K., Free, R., Hardwick, R. J., McVey, D., Brookes, A., Hollox, E. J. and Talbot, C. J. (2013) 'Automated design of paralogue ratio test assays for the accurate and rapid typing of copy number variation', *Bioinformatics*, pp. 1–7.

Verschuren, W., Boerma, G. and Kromhout, D. (1994) 'Total and HDL-cholesterol in The Netherlands: 1987-1992. Levels and changes over time in relation to age, gender and educational level.', *International journal of epidemiology*, Journal Article, Research Support, Non-U.S. Gov't, ENGLAND, 23(5), pp. 948–956.

Vittori, A., Orth, M., Roos, R., Outeiro, T., Giorgini, F. and Hollox, E. (2013) 'Beta Defensin Genomic Copy Number Does Not Influence the Age of Onset in Huntington's Disease', *Journal of Huntington's Disease 2*, pp. 107–124.

Vogelstein, B. and Kinzler, K. W. (1999) 'Digital PCR.', *Proceedings of the National Academy of Sciences of the United States of America*, 96(16), pp. 9236–41.

Volik, S., Raphael, B. J., Huang, G., Stratton, M. R., Bignel, G., Murnane, J., Brebner, J. H., Bajsarowicz, K., Paris, P. L., Tao, Q., Kowbel, D., Lapuk, A., Shagin, D. A., Shagina, I. A., Gray, J. W., Cheng, J.-F., de Jong, P. J., Pevzner, P. and Collins, C. (2006) 'Decoding the fine-scale structure of a breast cancer genome and transcriptome.', *Genome research*, 16(3), pp. 394–404.

Wain, L. and Tobin, M. (2011) 'Copy Number Variation', In Teare, M. D. (ed.), *Genetic Epidemiology*, Methods in Molecular Biology, Totowa, NJ, Humana Press, pp. 167–183.

Wain, L. V, Armour, J. a L. and Tobin, M. D. (2009) 'Genomic copy number variation, human health, and disease.', *Lancet*, Elsevier Ltd, 374(9686), pp. 340–50.

Wain, L. V, Odenthal-Hesse, L., Abujaber, R., Sayers, I., Beardsmore, C., Gaillard, E. A., Chappell, S., Dogaru, C. M., McKeever, T., Guetta-Baranes, T., Kalsheker, N., Kuehni, C. E., Hall, I. P., Tobin, M. D. and Hollox, E. J. (2014) 'Copy number variation of the beta-defensin genes in

europeans: no supporting evidence for association with lung function, chronic obstructive pulmonary disease or asthma.', Ahuja, S. K. (ed.), *PloS one*, 9(1), p. e84192.

Walker, S., Janyakhantikul, S. and Armour, J. A. L. L. (2009) 'Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs', *Genomics*, Elsevier Inc., 93(1), pp. 98–103.

Wan, Y. I., Shrine, N. R., Soler Artigas, M., Wain, L. V, Blakey, J. D., Moffatt, M. F., Bush, A., Chung, K. F., Cookson, W. O., Strachan, D. P., Heaney, L., Al-Momani, B. A., Mansur, A. H., Manney, S., Thomson, N. C., Chaudhuri, R., Brightling, C. E., Bafadhel, M., Singapuri, A., Niven, R., Simpson, A., Holloway, J. W., Howarth, P. H., Hui, J., Musk, A. W., James, A. L., Australian Asthma Genetics, C., Brown, M. A., Baltic, S., Ferreira, M. A., Thompson, P. J., Tobin, M. D., Sayers, I. and Hall, I. P. (2012) 'Genome-wide association study to identify genetic determinants of severe asthma', *Thorax*, 67(9), pp. 762–768.

Wang, G. (2014) 'Human Antimicrobial Peptides and Proteins', *Pharmaceuticals*, 7(5), pp. 545–594.

Wat, M. J., Shchelochkov, O. a, Holder, A. M., Breman, A. M., Bacino, C., Scaglia, F., Zori, R. T., Cheung, S. W., Daryl, A. and Kang, S. L. (2009) 'Chromosome 8p23.1 Deletions as a Cause of Complex Congenital Heart Defects and Diaphragmatic Hernia', *Am J Med Genet A.*, 149A(8), pp. 1661–1677.

Weinberg, A., Jin, G., Sieg, S. and McCormick, T. S. (2012) 'The yin and yang of human Beta-defensins in health and disease.', *Frontiers in immunology*, 3(October), p. 294.

Weinberg, A., Quinones-Mateu, M. and Lederman, M. (2006) 'Role of Human -defensins in HIV Infection', *Advances in Dental Research*, 19(1), pp. 42–48.

Weiss, M. M., Hermsen, M. A., Meijer, G. A., van Grieken, N. C., Baak, J. P., Kuipers, E. J. and van Diest, P. J. (1999) 'Comparative genomic hybridisation.', *Molecular pathology : MP*, 52(5), pp. 243–51.

Wenger, N. (1999) 'Coronary risk reduction in the menopausal women', *Rev Port Cardiol*, 18(Suppl III), pp. 39–47.

Whittington, C. M., Papenfuss, A. T., Bansal, P., Torres, A. M., Wong, E. S. W., Deakin, J. E., Graves, T., Alsop, A., Schatzkamer, K., Kremitzki, C., Ponting, C. P., Temple-Smith, P., Warren, W. C., Kuchel, P. W. and Belov, K. (2008) 'Defensins and the convergent evolution of platypus and reptile venom genes', *Genome Research*, 18(6), pp. 986–994.

WHO (2013a) *A global brief on Hypertension: Silent killer, global public health crisis*,.

WHO (2013b) ‘Asthma’, *Media Centre Fact Sheets*, [online] Available from: http://www.who.int/mediacentre/factsheets/fs307/en/.

WHO (2015a) ‘Chronic obstructive pulmonary disease (COPD)’, *Media Centre Fact Sheets*, [online] Available from: http://www.who.int/mediacentre/factsheets/fs315/en/.

WHO (2015b) ‘HIV/AIDS’, *Media Centre Fact Sheets*, [online] Available from: http://www.who.int/mediacentre/factsheets/fs360/en/ (Accessed 4 August 2015).

Wilk, J. B., Chen, T., Gottlieb, D. J., Walter, R. E., Nagle, M. W., Brandler, B. J., Myers, R. H., Borecki, I. B., Silverman, E. K., Weiss, S. T. and O’Connor, G. T. (2009) ‘A Genome-Wide Association Study of Pulmonary Function Measures in the Framingham Heart Study’, McCarthy, M. I. (ed.), *PLoS Genetics*, 5(3), p. e1000429.

Wilk, J. B., Shrine, N. R. G., Loehr, L. R., Zhao, J. H., Manichaikul, A., Lopez, L. M., Smith, A. V., Heckbert, S. R., Smolonska, J., Tang, W., Loth, D. W., Curjuric, I., Hui, J., Cho, M. H., Latourelle, J. C., Henry, A. P., Aldrich, M., Bakke, P., Beaty, T. H., Bentley, A. R., Borecki, I. B., Brusselle, G. G., Burkart, K. M., Chen, T. H., Couper, D., Crapo, J. D., Davies, G., Dupuis, J., Franceschini, N., Gulsvik, A., Hancock, D. B., Harris, T. B., Hofman, A., Imboden, M., James, A. L., Khaw, K. T., Lahousse, L., Launer, L. J., Litonjua, A., Liu, Y., Lohman, K. K., Lomas, D. A., Lumley, T., Marciante, K. D., McArdle, W. L., Meibohm, B., Morrison, A. C., Musk, A. W., Myers, R. H., North, K. E., Postma, D. S., Psaty, B. M., Rich, S. S., Rivadeneira, F., Rochat, T., Rotter, J. I., Soler Artigas, M., Starr, J. M., Uitterlinden, A. G., Wareham, N. J., Wijmenga, C., Zanen, P., Province, M. A., Silverman, E. K., Deary, I. J., Palmer, L. J., Cassano, P. A., Gudnason, V., Barr, R. G., Loos, R. J. F., Strachan, D. P., London, S. J., Boezen, H. M., Probst-Hensch, N., Gharib, S. A., Hall, I. P., O’Connor, G. T., Tobin, M. D. and Stricker, B. H. (2012) ‘Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction’, *American Journal of Respiratory and Critical Care Medicine*, 186(7), pp. 622–632.

Willcocks, L., Lyons, P., Clatworthy, M., Robinson, J., Yang, W., Newland, S., Plagnol, V., McGovern, N., Condliffe, A., Chilvers, E., Adu, D., Jolly, E., Watts, R., Lau, Y., Morgan, A., Nash, G. and Smith, K. (2008) ‘Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake’, *Journal of Experimental Medicine*, 205(7), pp. 1573–1582.

Wu, L., Gerard, N. P., Wyatt, R., Choe, H., Parolin, C., Ruffing, N., Borsetti, A., Cardoso, A. A., Desjardin, E., Newman, W., Gerard, C. and Sodroski, J. (1996) ‘CD4-induced interaction of

primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5.', *Nature*, 384(6605), pp. 179–183.

Wu, Z., Cocchi, F., Gentles, D., Ericksen, B., Lubkowski, J., Devico, A., Lehrer, R. I. and Lu, W. (2005) 'Human neutrophil alpha-defensin 4 inhibits HIV-1 infection in vitro.', *FEBS letters*, 579(1), pp. 162–6.

Wu, Z., Hoover, D. M., Yang, D., Boulegue, C., Santamaria, F., Oppenheim, J. J., Lubkowski, J. and Lu, W. (2003) 'Engineering disulfide bridges to dissect antimicrobial and chemotactic activities of human beta-defensin 3.', *Proceedings of the National Academy of Sciences of the United States of America*, Journal Article, Research Support, Non-U.S. Gov't, Research Support, U.S. Gov't, P.H.S., United States, 100(15), pp. 8880–8885.

Yamaguchi, Y., Nagase, T., Makita, R., Fukuhara, S., Tomita, T., Tominaga, T., Kurihara, H. and Ouchi, Y. (2002) 'Identification of Multiple Novel Epididymis-Specific -Defensin Isoforms in Humans and Mice', *The Journal of Immunology*, 169(5), pp. 2516–2523.

Yanagi, S., Ashitani, J., Ishimoto, H., Date, Y., Mukae, H., Chino, N. and Nakazato, M. (2005) 'Isolation of human beta-defensin-4 in lung tissue and its increase in lower respiratory tract infection.', *Respiratory research*, 6, p. 130.

Yang, D., Chertov, O., Bykovskaia, S. N., Chen, Q., Buffo, M. J., Shogan, J., Anderson, M., Schröder, J. M., Wang, J. M., Howard, O. M. and Oppenheim, J. J. (1999) 'Beta-defensins: linking innate and adaptive immunity through dendritic and T cell CCR6.', *Science*, 286(5439), pp. 525–528.

Young, R. P., Hopkins, R. and Eaton, T. E. (2007) 'Forced expiratory volume in one second: not just a lung function test but a marker of premature death from all causes.', *The European respiratory journal*, Journal Article, Review, Switzerland, 30(4), pp. 616–622.

Yudin, A. I., Generao, S. E., Tollner, T. L., Treece, C. A., Overstreet, J. W. and Cherr, G. N. (2005) 'Beta-defensin 126 on the cell surface protects sperm from immunorecognition and binding of anti-sperm antibodies.', *Biology of reproduction*, Journal Article, United States, 73(6), pp. 1243–1252.

Yusuf, S., Hawken, S., Ounpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J. and Lisheng, L. (2004) 'Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study.', *Lancet*, Elsevier, 364(9438), pp. 937–52.

Yusuf, S., Rangarajan, S., Teo, K., Islam, S., Li, W., Liu, L., Bo, J., Lou, Q., Lu, F., Liu, T., Yu, L., Zhang, S., Mony, P., Swaminathan, S., Mohan, V., Gupta, R., Kumar, R., Vijayakumar, K., Lear, S., Anand, S., Wielgosz, A., Diaz, R., Avezum, A., Lopez-Jaramillo, P., Lanas, F., Yusoff, K., Ismail, N., Iqbal, R., Rahman, O., Rosengren, A., Yusufali, A., Kelishadi, R., Kruger, A., Puoane, T., Szuba, A., Chifamba, J., Oguz, A., McQueen, M., McKee, M. and Dagenais, G. (2014) 'Cardiovascular Risk and Events in 17 Low-, Middle-, and High-Income Countries', *New England Journal of Medicine*, 371(9), pp. 818–827.

Zeng, M., Smith, A. J., Wietgrefe, S. W., Southern, P. J., Schacker, T. W., Reilly, C. S., Estes, J. D., Burton, G. F., Silvestri, G., Lifson, J. D., Carlis, J. V. and Haase, A. T. (2011) 'Cumulative mechanisms of lymphoid tissue fibrosis and T cell depletion in HIV-1 and SIV infections', *Journal of Clinical Investigation*, 121(3), pp. 998–1008.

Zhang, F., Gu, W., Hurles, M. E. and Lupski, J. R. (2009) 'Copy number variation in human health, disease, and evolution.', *Annual review of genomics and human genetics*, 10, pp. 451–81.

Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D. and Lupski, J. R. (2009) 'The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans', *Nature Genetics*, Nature Publishing Group, 41(7), pp. 849–853.

Zhang, G., Wu, H., Shi, J., Ganz, T., Ross, C. R. and Blecha, F. (1998) 'Molecular cloning and tissue expression of porcine β-defensin-1', *FEBS Letters*, Elsevier, 424(1–2), pp. 37–40.

Zhang, L., Yu, W., He, T., Yu, J., Caffrey, R. E., Dalmasso, E. A., Fu, S., Pham, T., Mei, J., Ho, J. J., Zhang, W., Lopez, P. and Ho, D. D. (2002) 'Contribution of human alpha-defensin 1, 2, and 3 to the anti-HIV-1 activity of CD8 antiviral factor.', *Science*, 298(5595), pp. 995–1000.

Zhang, X., Müller, S., Möller, M., Huse, K., Taudien, S., Book, M., Stuber, F., Platzer, M. and Groth, M. (2014) '8p23 beta-defensin copy number determination by single-locus pseudogene-based paralog ratio tests risk bias due to low-frequency sequence variations', *BMC Genomics*, 15(1), p. 64.

Zhao, C., Nguyen, T., Liu, L., Sacco, R. E., Brogden, K. A. and Lehrer, R. I. (2001) 'Gallinacin-3, an inducible epithelial beta-defensin in the chicken.', *Infection and immunity*, Journal Article, United States, 69(4), pp. 2684–2691.

Zhao, J., Wang, Z., Chen, X., Wang, J. and Li, J. (2011) 'Effects of intravesical liposome-mediated human beta-defensin-2 gene transfection in a mouse urinary tract infection model',

Microbiology and Immunology, 55(4), pp. 217–223.

Zhou, X.-J., Cheng, F.-J., Lv, J.-C., Luo, H., Yu, F., Chen, M., Zhao, M.-H. and Zhang, H. (2012) 'Higher DEFB4 genomic copy number in SLE and ANCA-associated small vasculitis.', *Rheumatology*, 51(6), pp. 992–5.

Zhou, Y. S., Webb, S., Lettice, L., Tardif, S., Kilanowski, F., Tyrrell, C., MacPherson, H., Semple, F., Tennant, P., Baker, T., Hart, A., Devenney, P., Perry, P., Davey, T., Barran, P., Barratt, C. L. and Dorin, J. R. (2013) 'Partial Deletion of Chromosome 8 Î²-defensin Cluster Confers Sperm Dysfunction and Infertility in Male Mice', Barsh, G. S. (ed.), *PLoS Genetics*, 9(10), p. e1003826.

Zilbauer, M., Jenke, A., Wenzel, G., Postberg, J., Heusch, A., Phillips, A. D., Noble-Jamieson, G., Torrente, F., Salvestrini, C., Heuschkel, R. and Wirth, S. (2010) 'Expression of human beta-defensins in children with chronic inflammatory bowel disease.', *PloS one*, 5(10), p. e15389.

Zimmer, C. (2012) *A Planet of Viruses*, book, University of Chicago Press.

Zou, J., Mercier, C., Koussounadis, A. and Secombes, C. (2007) 'Discovery of multiple beta-defensin like homologues in teleost fish.', *Molecular immunology*, 44(4), pp. 638–47.