Evolution of copy number variation in the rhesus macaque β-defensin region

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

Barbara Ottolini
Department of Genetics
University of Leicester

September 2013

To my parents

Ai miei genitori

Carmen e Giuseppe

ABSTRACT

Beta defensins are multifunctional secreted short peptides rapidly evolving in mammals. They present 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 and an approximate repeat length of 240kb but the extent and nature of CNV in other mammals remains poorly known.

The rhesus macaque (*Macaca mulatta*) is the most widespread non-human primate, hence constituting a good model to study adaptation and its divergence time from the human lineage (~25MYA) presents enough sequence diversity to investigate mechanisms of copy number variation and evolution. Its genome has been sequenced, although there is poor assembly quality in repeated segments such as the β-defensin region.

This thesis studied the genomic architecture of the rhesus macaque β-defensin region using a variety of methods (aCGH, PCR-based methods, BAC library screening and cytogenetic approaches) with the aim of overcoming the limitations of the assembly and of determining the copy number distribution for this region.

Evidences are here provided that only the region containing *DEFB2L* gene (orthologue to human *DEFB4*) is CNV, with a diploid copy number between 2 and 11, with a repeat size of 20kb, while the rest of the cluster shows no variation. This could represent a case where the same area prone to copy number variation evolved to present a different copy number unit structure in two different lineages, still converging in the same copy number distribution for the possible effect of similar functional constraints. Also, evidences of non-synonymous variations are shown for the *DEFB2L* gene, suggestive of the different evolutionary pattern followed by the rhesus macaque β-defensin region.

ACKNOWLEDGEMENTS

I would like to thank Ed Hollox, my supervisor, for the enormous help during these years in the lab, for all the last minute corrections and for having survived so long with loud Italians around. Thanks also to all the members of G3 lab. Ange, Lee, Rob, Helen, Jen, Raheleh, Linda, Shamik, Razan... It has been great to work with you guys!

A big thanks goes to all my adoptive lab G2... To Mark for all the support, Chiara, Pille, Pierpaolo, Jon, Dan, Turi and Gurdeep. I also thank all the collaborators for this project: Charles Bevins, Jay Solnick and Mike Hornsby from UC Davis, and Trude Schwarzacher, for all the patience in teaching me FISH.

But really, I am grateful to everyone working in the Genetics Department, for four splendid years!

I would then like to thank Chris, for all the funny moments spent together here in Leicester: I miss you big time, mate! Giorgio, always Rock dude! Laura, for being a friend with a bottle always opened ☺ . Diego, we will play again Cards Against Humanity someday. All the Olive's bar staff: it has been a pleasure to work with you guys!

To all my friends from Italy and surroundings: Misia, Oriana, Dado (Mate!) and Kesia, Dario, Daniel, Cento, Chiara, Riccardino, Paolo, Asse, Gab, Lord Phobos, Barbara and Diego, Achille, Sara, Maria, Monika, Mariantonietta, Elisa and many others… I'm so grateful for knowing you all: you've always been there for me every single time I came back home; grazie ragazzi, mi mancate davvero!

A thought goes also to all the Sevengroup restaurants staff, for having supported me during my University years in Italy: if I arrived until here it's also thanks to you guys!

My family deserves a super special hug, as you needed to deal with this little troublemaker since a long time!

Mikaël, thank you for having replied to that housemate advert. All the other reasons for which I should thank you every day for being so special to me are needless to say in here.

And if I haven't just forgotten you…Yippee ki-yay! ☺.

ABBREVIATIONS

Acronim	Description
6-HEX	6-carboxy-2,4,4 ,5 ,7,7-hexachlorofluorescein succinimidyl ester
ACE2	Angiotensin I Converting Enzyme (peptidyl-dipeptidase A) 2
AIDS	Acquired Immunodeficiency Syndrome
AMY1	Amylase 1
ANK3	Ankyrin 3
APOL1	Apolipoprotein L-1
AQP7	AquaPorin-7
AZF	Azospermia Factor
BAC	Bacterial artificial chromosome
BAF	B allele frequency
BCIP	5-bromo-4-chloro-3-indolyl phosphate
BOLA2	BolA homolog 2
BRS3	Bombesin-like receptor 3
BSA	Bovine serum albumin
C4	Complement component 4
CACNA1C	Calcium Channel, voltage-dependent, L type, alpha 1C subunit
CCL1	Chemokine (C-C motif) ligand 1
CCR5	Chemokine (C-C motif) receptor 5

CCR6	Chemokine (C-C motif) receptor 6
CD	Cluster of differentiation
CES	Carboxylesterase
CF	Cystic fibrosis
CGH	Comparative genomic hybridisation
CHEF	Contour-clamped homogeneous electric field
CI	Confidence interval
CMT1	Charcot-Marie-Tooth neuropathy Type 1
CN	Copy number
CNV	Copy number variation
CT	Threshold cycle
CXCR4	Chemokine (C-X-C motif) receptor 4
Cy3-Cy5	Cyanin 3-Cyanin 5
CYP2D6	Cytochrome P450 2D6
DAPI	4',6-diamidino-2-phenylindole
DC	Dendritic cell
ddPCR	Droplet digital PCR
DMSO	Dimethyl sulfoxide
DSB	DNA double strand break
DTT	Dithiothreitol
EDTA	Ethylenediaminetetracetic acid
EGFR	Epidermal growth factor receptor

F8	Coagulation factor VIII
FAM	6-carboxyfluorescin
FBS	Fetal bovine serum
FCGR	FC gamma receptor
FISH	Fluorescence i*n situ* hybridisation
FITC	Fluorescein isothiocyanate
FLT3	Fms-related tyrosine kinase 3
FMS	Fibromyalgia syndrome
FoSTeS	Fork stalling and template switching
FOXP3	Forkhead box 3
FPEM	Fosmid paired-end mapping
FXII	Coagulation factor XII
GO	Gene ontology
GRA	Glucocorticoid-remediable aldosteronism
GWAS	Genome wide association study
HERV	Human endogenous retrovirus
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HNPP	Hereditary neuropathy with liability to pressure palsies
IFN	Interferon
IL	Interleukin
INT	2-[4-iodophenyl]-3-[4-nitrophenyl]-5-phenyltetrazolium chloride

IUCN	International Union for the Conservation of Nature
KIR3DX1	Killer cell Immunoglobulin-like receptor, three domains, X1
L1	Long-intersperse element 1
LARC	Liver activation regulated chemokine
LCR	Low-copy repeat
LILR	Leukocyte immunoglobulin-like receptors
LINE	Long interspersed element
LPS	Lipopolysaccharide
LRR	Log R ratio
LTA	Lipoteichoic acid
LWS	Long-wavelength-sensitive
MAPH	Multiplex amplifiable probe hybridisation
Mc1r	Melanocortin receptor 1
MDP	Muramyldipeptide
MGB	Dihydrocyclopyrroloindole tripeptide minor groove binder
MLPA	Multiplex ligation-dependent probe amplification
MRGPRX1	Mas-related G-protein coupled receptor member X1
MYA	Million years ago
NAHR	Non-allelic homologous recombination
NBF	Neuroblastoma breakpoint family
N_e	Effective population size
NFkB	Nuclear factor of kappa light polypeptide gene enhancer in B-cells

NGS	Next-generation gequencing
NHEJ	Non homologous end Joining
NK	Natural killer cell
NME4	Non-metastatic cell 4
NMR	Nuclear magnetic resonance
OMIM	Online mendelian inheritance in man
OPN1LW	Opsin 1 long-wave
OPN1MW	Opsin 1 medium-wave
OR	Olfactory receptor
OSS	Oligonucleotide stop solution
PABPC5	Poly(A) binding protein, cytoplasmic 5
PAR	Protease-activated receptor
PAX9	Paired box gene 9
PEM	Paired-end mapping
PFGE	Pulsed–field–gel-electrophoresis
PFKP	Phosphofructokinase platelet form
PMA	Phorbol-12-myristate
PMP22	Peripheral myelin protein 22
PRNP	Prion protein
PRT	Paralogue ratio test
PSV	Paralogous sequence variant
PTEN	Phosphatase and tensin homolog

PTX	Pertussis toxin
qPCR	Quantitative PCR
RBM8A	RNA-binding protein 8A
REPD	Repeat distal
REPP	Repeat proximal
RNASE1	Ribonuclease pancreatic enzyme
SAIDS	Simian acquired immunodeficiency syndrome
SAP (SH2D1A)	SH2 domain containing 1A
SD	Segmental duplication
SINE	Short interspersed element
SIV	Simian immunodeficiency virus
SLC6A2	Solute carrier family 6, member 2
SLE	Systemic lupus erythematosus
SMS	Smith-Magenis syndrome
SNP	Single nucleotide polymorphism
SOX9	Sex determining region Y-box 9
SUN	Singly unique nucleotide
SV	Structural variation
SWS1	Short-wavelength-sensitive 1
TAR	Thrombocytopenia with absent radii
TAZ	Tafazzin
TCR	T cell receptor

Th17	T helper 17 cell
TLR4	Toll-like receptor 4
TNF	Tumour necrosis factor
UGT2B33	UDP glucuronosyltransferase 2 family, polypeptide B33
UV	Ultraviolet
VEGF	Vascular endothelial growth factor
VNTR	Variable number of tandem repeats
WGA	Whole genome amplification
WGS	Whole genome shotgun
WTCCC	Wellcome Trust Case Control Consortium
α-MSH	α-melanocyte-stimulating hormone

CONTENTS

Contents

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

1.1 Aim of the study

This thesis presents an analysis of the complex genomic region of the β-defensins cluster in the rhesus macaque *Macaca mulatta*. In particular, it describes molecular and cytogenetic approaches for copy number variation detection and typing and strategies to overcome problems rising with poorly assembled genomic areas.

1.2 Definition of Copy Number Variation

The notion of gene dosage, where differences in gene number could quantitatively affect gene expression levels and phenotype has been known since 1930s, in the pioneer years of scientific research. The earliest known example is the discovery of the *Bar* locus in *Drosophila melanogaster* [1]. *Bar* is a dominant, homozygous viable, X-linked mutation that has the effect of reducing the number of facets of the *Drosophila* compound eyes; females heterozygous for *Bar* present half the number of facets compared to wild-type. The gene duplication of the *Bar* locus, generated by unequal crossing over, has a direct effect on the fly phenotype with reduced eye size and slit eye shape (**Figure 1.1**). Interestingly, cases of gene triplication give rise to an even more extreme phenotype, called Double-Bar, characterised by a greater reduction in the eye size, with their reciprocal products being wild-type revertants. This represents the first paradigmatic example of a quantitative trait determined by copy number with a clearly visible effect both at the microscopic level, with polytene chromosomes banding, and at the phenotypic level.

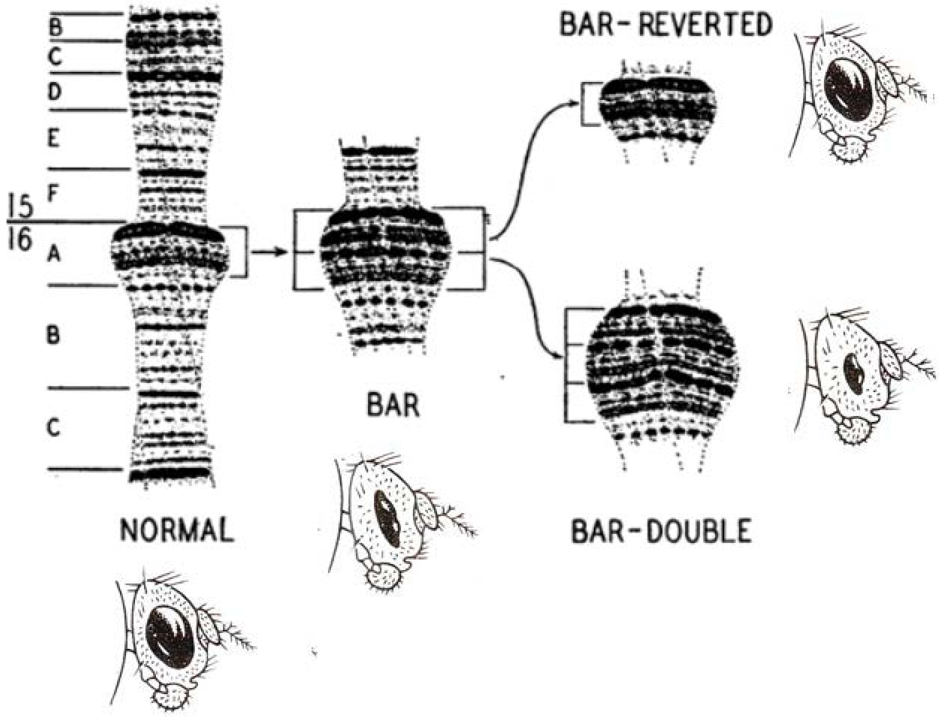


Figure 1.1: Diagram of the BAR locus size in *Drosophila* salivary glands polytene chromosomes. Adapted from Bridges 1936 [1].

Another early observation on the effects of quantitative variations was done on complement component C4[2], that was discovered to have an electrophoretic polymorphism. In particular, immunofixation electrophoresis revealed 3 clusters of bands: 4 fast-moving anodal bands, 4 slow-moving cathodal bands and a combination of the two. This was consistent with the presence of two loci that could both be present, simultaneously or not, in the general population. Further studies revealed the presence of two isoforms for the complement component C4, *C4A* (acidic) and *C4B* (basic), generating proteins different for hemolytic activity and extent of binding to carbohydrate antigens and immune complexes of soluble antigens [3]. Following activation and cleavage, C4 fragments attach to the erythrocytes membrane; this characteristic is of pivotal importance for blood transfusions, as blood types with different C4 isoforms are not compatible. Antibodies specific for the *C4B* isoform (Chido) and for the *C4A* isoform (Rodgers) are used to discriminate between the two variant in the Chido/Rodgers Blood Group System [4].

Although these quantitative variations were visible using Southern blotting approaches, they were not thought to be widespread along the genome. Conversely, the discovery of the single nucleotide polymorphisms (SNPs) extent during the last twenty years of scientific research opened a period of time in which variations at single nucleotide level were believed to encompass the major part of human genomic diversity.

Gradually, a new level of DNA complexity was unveiled, with the discovery of different landmarks of genomic instability, like variable number of tandem repeats (VNTRs) as minisatellites and microsatellites, or the presence of retrotransposable elements, as long interspersed elements 1 (L1) or *Alu* sequences [5 6].

Nevertheless, in recent years more powerful genome wide analysis tools made their appearance, such as array comparative genomic hybridisation (CGH) and next-generation sequencing (NGS), with “paired-end” methods, higher resolution at dramatically lower costs. These scientific advancements led to the discovery of extensive structural variations spread across the genome in 2004, below the resolution limit of the traditional chromosome banding techniques, which were called Copy Number Variations (CNV) [7 8]. Since then, the scientific community placed an increasing effort in the identification and characterisation of CNVs across the genome, in parallel with the development of more accurate and dedicated techniques for CNV detection, as illustrated in **Figure 1.2:** . In less than one decade, CNVs have become a key aspect of disease susceptibility studies, in particular with respect to clinical diagnostics, cancer research and genome-wide association studies (GWAS) data analysis.

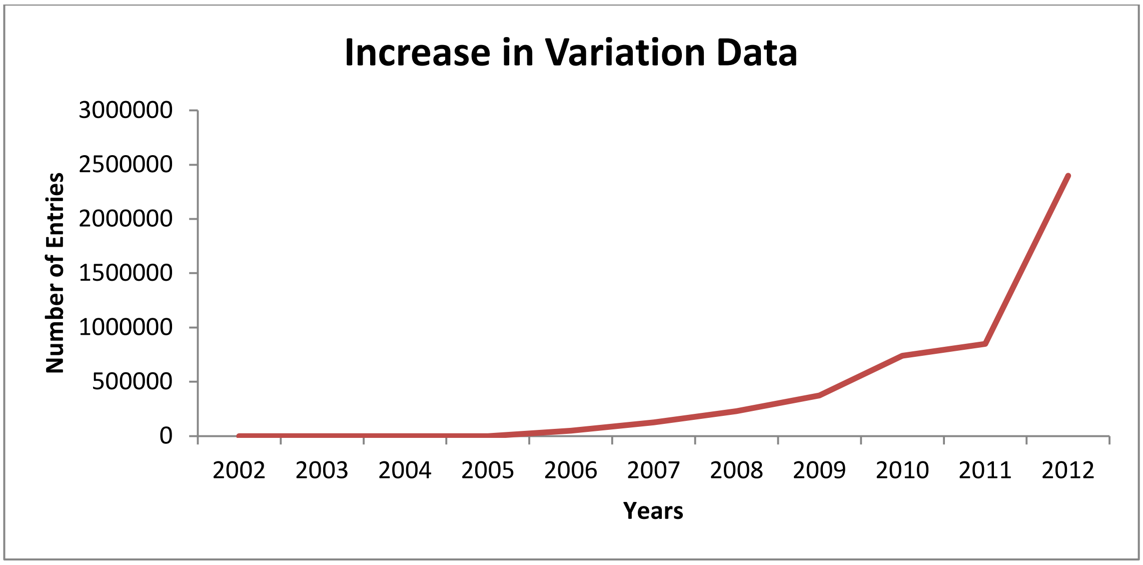
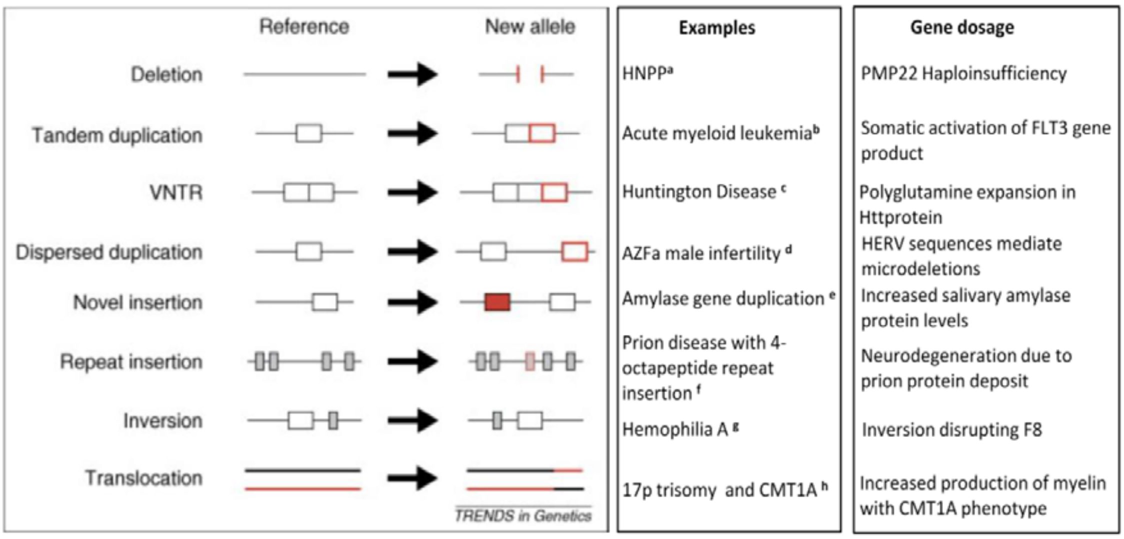


Figure 1.2: Graph showing the increase in published structural variation data that have been added to the Database of Genomic Variants (http://projects.tcag.ca/variation/) since its start in 2004; the numbers on the X axis reflect the year of publication. Adapted from http://dgv.tcag.ca/dgv/app/statistics?ref=NCBI36/hg18 accessed on 15/08/2013.

Formally, a CNV is defined as a segment of DNA ranging from one kilobase to several megabases in size that represents an imbalance between two genomes from one species. In particular, they present at variable copy number in comparison to a reference genome with the usual copy number of N = 2 [9]. Duplications, triplications, deletions, insertions or unbalanced translocations can give rise to CNV (**Figure 1.3**). The late characterisation of CNV, compared to other DNA variation, like SNPs, can be explainable by their size: they cannot be identified from a single sequence read and could hence be solved just with the advent of microarray and sequencing technologies. CNV could be defined as unbalanced structural variations (SV), considering it a broad term for genetic variants that alter the chromosomal structure. Genomic inversions and reciprocal translocations can determine balanced structural variations, which can still contribute to genomic instability. Both unbalanced and balanced variations require a break in the DNA phosphodiester backbone. Interestingly, most CNVs are flanked by segmental duplications, defined as contiguous portions of DNA interspersed or organised in tandem mapping to two or more genomic locations [10].



	Examples	Gene dosage
Deletion	HNPP[a]	PMP22 Haploinsufficiency
Tandem duplication	Acute myeloid leukemia[b]	Somatic activation of FLT3 gene product
VNTR	Huntington Disease [c]	Polyglutamine expansion in Httprotein
Dispersed duplication	AZFa male infertility [d]	HERV sequences mediate microdeletions
Novel insertion	Amylase gene duplication [e]	Increased salivary amylase protein levels
Repeat insertion	Prion disease with 4-octapeptide repeat insertion [f]	Neurodegeneration due to prion protein deposit
Inversion	Hemophilia A [g]	Inversion disrupting F8
Translocation	17p trisomy and CMT1A [h]	Increased production of myelin with CMT1A phenotype

Figure 1.3: Types of structural variant. Eight different types of structural variant are depicted, defined relatively to the reference genome sequence. Adapted from Hurles 2008[11].

1.3 Methods for CNV detection and typing

At present, there are different methods available for CNV screening; none of them can be considered as the gold standard for CNV detection and typing, and each has advantages and disadvantages as will be presented in this section.

1.3.1 Cytogenetic techniques

Historically, the first successful approach for large CNV detection (between 5 and 10 Mb) came from late 1960s chromosome banding, which allowed the identification of many rearrangements, often involved in congenital syndromes and cancer.

Later on, Fluorescence *In Situ* Hybridisation (FISH) improved the resolution limit of CNV detection: this technique uses multiple probes labelled in different colours on metaphasic chromosomes and is widely used in diagnostics to identify large copy number variations and structural variants in patients and carriers. Interestingly, a higher resolution version of FISH is now available, called fibre-FISH that allows the identification of closely spaced probes on interphasic nuclei, down to few kilobases, visualising them on stretched single DNA fibres. Currently, according to Perry GH *et al.* (2007), this seems the more precise method to determine the genomic structure of complex CNV, even if it requires high technical skills to perform it successfully and cannot be used to process large amounts of samples as it is not high-throughput [12] (**Figure 1.4**).

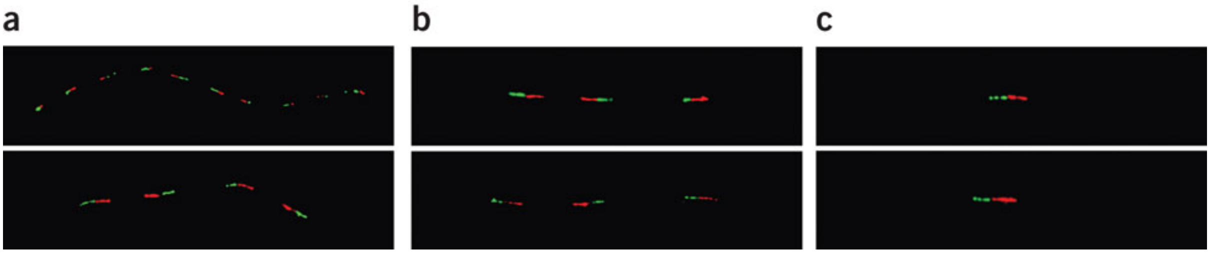


Figure 1.4: Example of high-resolution fibre FISH for the validation of the amylase gene AMY1 copy number estimates. The red signal is a 10kb probe encompassing the entire AMY1 gene and the green signal is given by a 8kb probe, specific for a retrotransposon directly upstream to the AMY1 gene. Panel a) shows a Japanese individual (GM18972) with 14 diploid AMY1 gene copies (one allele with 10 copies and the other allele with four copies. Panel b) shows a Biaka individual (GM10472) with 6 diploid AMY1 gene copies (3 copies for each allele). Panel c) shows the reference chimpanzee (Clint; S006006) with two diploid AMY1 gene copies [12]**.**

1.3.2 Genome wide CNV mapping

In order to map CNVs at the genome-wide level there are three principal methods available: CGH-based methods, SNP-based arrays and sequencing-based methods.

1.3.2.1 Comparative genomic hybridisation (CGH)

This method was developed based on the FISH concept, where DNA sequences in chromosome metaphase spreads can hybridise to complementary probes specifically designed. In particular, the test and the reference samples are labelled using two different fluorescent dyes (such as Cy3 and Cy5); assuming that the reference sample has a normal diploid copy number, it is then possible to detect CNVs as gains or losses of signal in the test sample[13] (**Figure 1.5**).

The major limit of this technique is its relatively low resolution (>5Mb) but this has been overcome through the development of array based-CGH, where different DNA sources, including BAC clones [14], cDNA clones, oligonucleotides and PCR genomic products could be hybridised onto a microarray platform. This leads to higher resolution power and allows a better identification of CNV regions. Nevertheless, cDNA microarrays in particular present two main problems: the bias towards gene-rich regions, with loss of information from areas with no genes not interrogated by the array, and the lack of introns, causing hybridisation problems and hence errors in the ratio calculation in case of mismatches between genomic DNA and cDNA.

BAC arrays allow more robust detections, with higher sensitivity and a lower CNV resolution limit of 20-30 kb. Nevertheless, their main problem remains the identification of the CNV boundaries, as BACs contain large DNA inserts of 150-250 kb; for this reason, studies using this approach report just the start and end coordinates of the BAC positive for a certain CNV variant, hence limiting resolution and overestimating the CNV size. Taking into account these caveats, BAC arrays find their best application in diagnostics, for the identification of large rearrangements [15].

Currently the oligonucleotide array platforms are the most reliable, with 60-mer oligonucleotides probes distributed across the genome with an average spacing of 35 kb (Figure 1.5). Interestingly, for region-based studies, also high density array CGH platforms are available, with tiling path probes, which has been used in this study.

It is noticeable that a 2010 industry report (UBS Investment Research Q-Series®) estimated that microarray-based molecular diagnostics was a market worthy of more than $100 million, mainly represented by DNA-based arrays; new array technology advancements are hence expectable over the next years for CNV detection in clinical diagnostics.

An interesting approach involving array CGH for CNV detection has been proposed by Park H *et al.* in 2010; they combined high resolution array CGH data with whole genome DNA sequencing data to obtain a comprehensive catalogue of common CNVs in Asian individuals. In this way it is possible to transform the relative copy number information obtained from array CGH experiments into absolute copy number values; this technique has the advantage of increasing the reliability of the CNV data obtained [16].

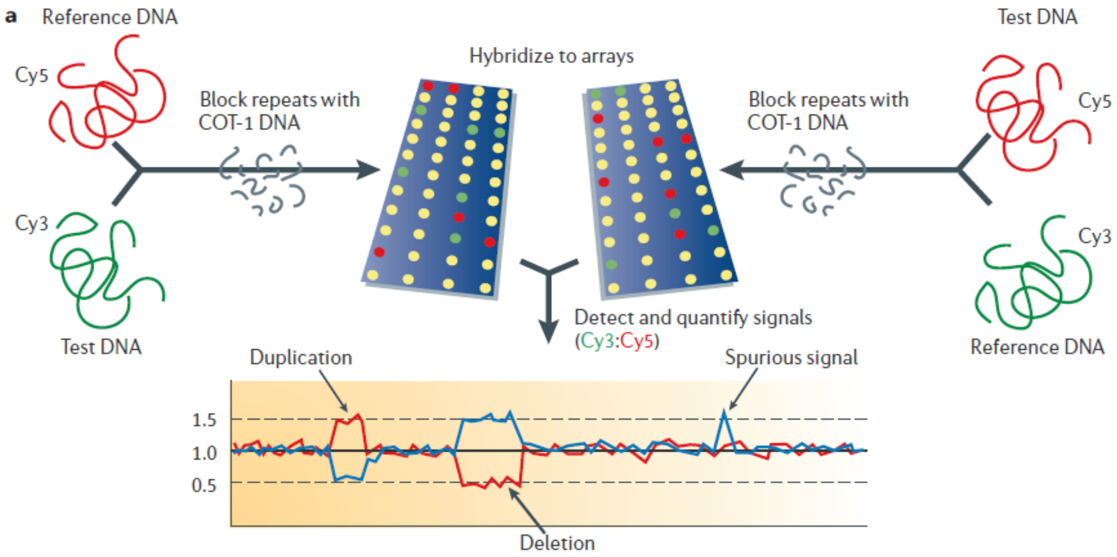


Figure 1.5: Array CGH scheme. The reference and the test DNA are labelled with two different fluorescent dyes (Cy5 and Cy3 respectively) and hybridised on a chip array, after repetitive-element binding is blocked using cot-1 DNA (DNA mainly composed of repetitive sequences. It is produced when short fragments of denatured genomic DNA are re-annealed). After hybridisation, the fluorescent ratio between the two dyes is determined, revealing copy number differences between the two DNA samples, as variations in signal intensity. This technique is typically carried out using a 'dye swap' method, in which the initial labelling of the reference and test DNA samples is reversed for a second hybridisation. This allows the detection of spurious signals for which the reciprocal ratio is not observed, as indicated in the graph, where the red line represents the first hybridisation, whereas the blue line represents the reciprocal, dye-swapped, hybridisation[9].

1.3.2.2 SNP arrays versus CGH platforms for CNV detection

In the present state of the art the most comprehensive assessment of different array platforms and of the algorithms used for copy number calling has been performed by Pinto D *et al.* in 2012 [17]. This study compared CNV detection among eleven microarray platforms, from different manufacturers, based on CGH technology, SNP array and combined SNP+CNV probes. In particular, three levels of analysis were performed: evaluation of the signal variability of the raw data, prior to CNV calling; data set analysis with different algorithms for CNV calling, assessing the reproducibility between duplicates and the copy number size distribution; comparison of the CNV calls obtained through the different platforms with a validated set of variants, to identify the rate of false-positives and falls negatives and the accuracy of the CNV boundaries.

The first interesting point raised by the study is the inverse correlation between probe length and variability in the fluorescence intensity, demonstrating how longer probes generate less variance. This characteristic determines also a better signal-to-noise ratio using CGH platforms versus SNP arrays, demonstrated in the detection of 2:1 intensity signals on X chromosomes between females and males samples. Moreover, the choice of the baseline reference library represents a crucial step, especially for SNP arrays that, besides presenting a sample-specific variability, can generate artificial variability estimates in sample to reference comparison.

The reproducibility of the CNV calls has been found to be <70% for most of the platforms analysed, with better performances obtained for newer and CNV-focused arrays; interestingly, inter-laboratory variability correlates with reproducibility. The biggest variables affecting reproducibility are the platform used and the algorithm selected for CNV calling. These evidences raise the attention on the necessity of careful platform choices accordingly to each study aims. Moreover, also the CNV size can affect reproducibility: bigger CNVs (>200kb) tend to be miscalled if they fall in low complexity regions encompassing segmental duplications or, in case of lower probe coverage, they can be erroneously fragmented into several copy number calls.

Another substantial difference reported is that SNP arrays tend to miss deletions, whereas CGH platforms perform better in this respect; this bias can be due to the differences in the type of reference: SNP chips utilise a population reference, where CGH arrays use a single sample reference. This difference makes CGH platform better performing also for the detection of small copy number differences.

A last important consideration regards the detection of copy number breakpoints: also in this case different algorithms used on the same raw data sets can generate variability and discrepancies. The accuracy of this parameter follows the same trend observed for reproducibility, with CGH arrays and CNV focused platforms performing better; nevertheless, all the platforms present a tendency towards breakpoints underestimation, as the results reported for each algorithm correspond to the last probes within a certain variant showing evidence of copy number differences.

1.3.2.3 Sequencing based methods

An alternative to overcome the bias problems arising with hybridisation-based experimental technologies is represented by sequencing-based methods. One of the first comprehensive study to detect structural variation with higher resolution than BAC arrays [7] was performed with fosmid paired-end mapping (FPEM) [18]. The authors developed a computational method for mapping paired-end sequence data from a human fosmid DNA genomic library, against the human reference sequence, built on a different source individual. Discrepancies in size and orientation of multiple paired-end sequences were used to call deletions, duplications and inversions, for a total of 297 structural variation sites, with a lower detection limit of 8 kb. Nevertheless, FPEM has the limitation of not calling smaller CNVs and difficulties in detecting duplication > 40 kb, which is the average size of the fosmid insert.

Shortly after, other studies followed the same concept, basing the detection of structural variation on paired-end read mapping (PEM); for example, Korbel J *et al.* in 2007 isolated and sequenced with 454 technology paired end fragments of 3 kb (**Figure 1.6**), from one African individual previously included in the study of Redon R 2006 [19] and from the same European individual analysed with FPEM [18]; the resulting sequences were then computationally mapped to the human reference assembly. Insertions and deletions were then identified as significant differences between mapped read pairs and the average insert size of the corresponding genomic library[20]. This approach can detect SV at 3 kb resolution, with an average resolution of breakpoint assignment of 664 bp. At this detection level, with an effective coverage of 2.1X and 4.3X, the authors could predict 761 SV for the European individual (versus 297 SV identified by Tuzun E *et al.* [18]) and 887 SV events for the African, relative to the reference genome. Disadvantages of PEM include a poor detection of structural variations within complex genomic areas rich in segmental duplications, difficulties in the identification of insertions larger than the average insert size of the library and underestimation of small CNVs [18].

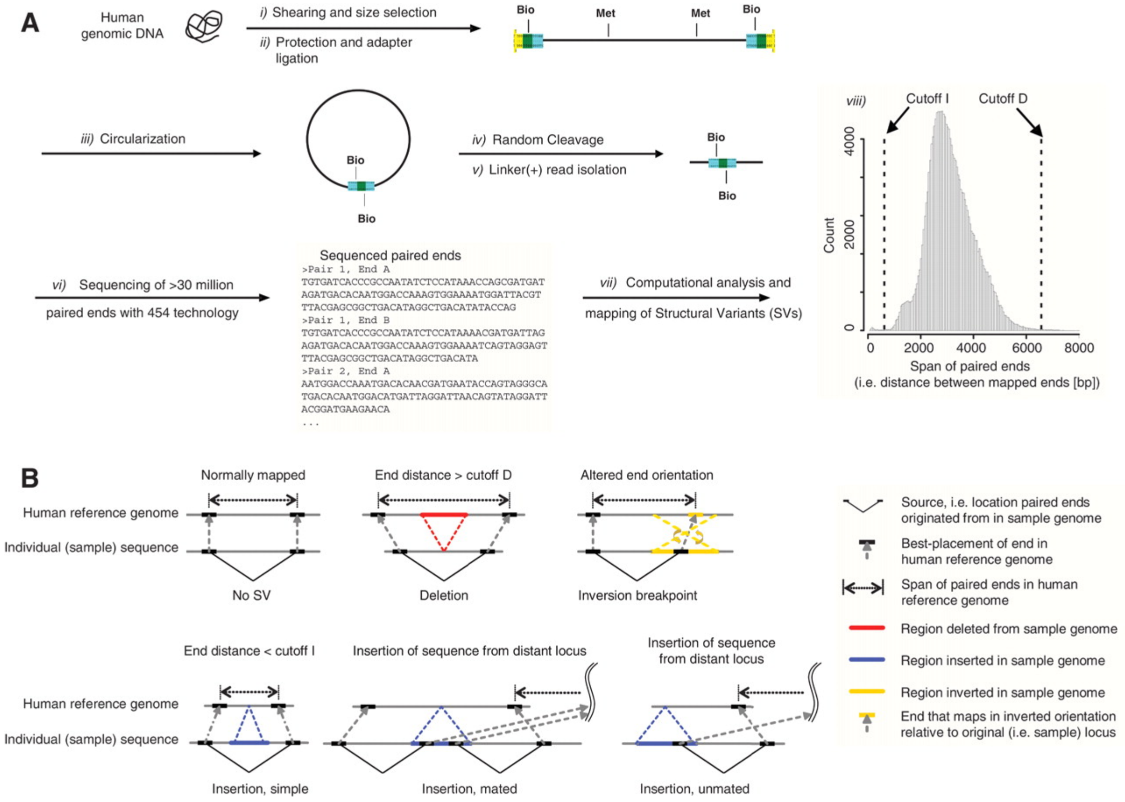

A
Human genomic DNA
i) Shearing and size selection
ii) Protection and adapter ligation
Bio
Met
Met
Bio
iii) Circularization
Bio
Bio
iv) Random Cleavage
v) Linker(+) read isolation
Bio
Bio
vi) Sequencing of >30 million paired ends with 454 technology
Sequenced paired ends
>Pair 1, End A
TGTGATCACCCGCCAATATCTCCATAAACCAGCGATGAT
AGATGACACAATGGACCAAAGTGGAAAATGGATTACGT
TTACGAGCGGCTGACATAGGCTGACATATACCAG
>Pair 1, End B
TGTGATCACCCGCCAATATCTCCATAAAACGATGATTAG
AGATGACACAATGGACCAAAGTGGAAAATCAGTAGGAGT
TTACGAGCGGCTGACATAGGCTGACATA
>Pair 2, End A
AATGGACCAAATGACACAACGATGAATACCAGTAGGGCA
TGACACAATGGACATGATTAGGATTAACAGTATAGGATT
ACGGATGAAGAACA
...
vii) Computational analysis and mapping of Structural Variants (SVs)
viii)
Cutoff I
Cutoff D
Count
0
2000
4000
0
2000
4000
6000
8000
Span of paired ends (i.e. distance between mapped ends [bp])
B
Normally mapped
End distance > cutoff D
Altered end orientation
Human reference genome
Individual (sample) sequence
No SV
Deletion
Inversion breakpoint
End distance < cutoff I
Insertion of sequence from distant locus
Insertion of sequence from distant locus
Human reference genome
Individual (sample) sequence
Insertion, simple
Insertion, mated
Insertion, unmated
Source, i.e. location paired ends originated from in sample genome
Best-placement of end in human reference genome
Span of paired ends in human reference genome
Region deleted from sample genome
Region inserted in sample genome
Region inverted in sample genome
End that maps in inverted orientation relative to original (i.e. sample) locus

Figure 1.6: (A) Flow chart illustrating PEM. (i) Genomic DNA is sheared to yield DNA fragments ~3 kb; (ii) biotinylated hairpin adapters are ligated to the fragment ends; (iii) circularisation of the fragments (iv) and random sharing; (v) isolation of linker (+) fragments; (vi) 454 sequencing of the library. (vii) Computational analysis of paired ends to determine (viii) the distribution of "paired-end spans" (shown for a single 454 sequencing pool). (B) Types of SVs. Deletions predicted from paired-end spans larger than a specified cut-off D; simple insertions have a span <cut-off I; inversions are seen when ends map to the genome at different relative orientations; other types of insertions are detected with evidence of sequence integration from a distal locus[20].

1.3.3 CNV typing

After the identification of CNV regions in a particular genome, the following challenge is the accurate typing of copy number differences among different individuals, to get more information about copy number distribution of the CNV region of interest. Despite the role that hybridisation-based platforms and sequencing approaches cover, as key resource for the generation of accurate CNV maps, their cost-effectiveness and their time of execution and analysis remains high. In order to investigate the copy number distribution of a certain locus, be it for diagnostic purposes, population genetics analysis or association studies, the main parameters to be considered are the reproducibility, the cost and the time of execution and their accuracy for the detection of integer copy number. Diverse approaches have been developed over the years, with good improvements in the parameters mentioned, as presented in the following paragraphs.

1.3.3.1 Southern blot

Southern blot has been one of the first approaches for CNVs detection, and it was widely used to detect duplications and deletions in the past. The presence of rearrangements in the region of interest generates restriction fragment size alterations that appear as novel bands on the blot, assuming to find a rare cutter enzyme suitable for the area flanking the CNV of interest.

The workload involved represents the main problem associated with this technique (DNA digestion, electrophoresis, blotting, hybridization and exposure), together with the large amount of high-quality DNA required (5-10µg). Moreover, it is a low throughput technique, for the time of execution and the limited number of samples on which it can be performed. Another issue raising from Southern blot analysis regards the detection of CNVs spanning segmental duplication, whose signal can be altered or loss in case coating DNA is used to prevent non-specific probe hybridisation.

An improvement in this sense came with the combination of the Southern blot approach with Pulsed–Field–Gel-Electrophoresis (PFGE). Normal gel electrophoresis poorly resolves DNA molecules longer than 25 kb, whereas in PFGE the direction of the electric field in the gel is periodically changed, allowing the resolution of molecules from ~10 to 2000 kb. To resolve molecules beyond the range of field inversion, it is also possible to use field-angle alternation electrophoresis, as CHEF (contour-clamped homogeneous electric field) which allows resolution up to 6 Mb. In this way it becomes possible to generate restriction profiles with rare cutters generating bigger bands, for the detection of larger CNVs. Also, restriction patterns different from a known reference sample can be helpful for the detection of deletions and duplications (**Figure 1.7**) [21]. Being a semi-quantitative method, it is possible to detect duplications as increases in the band intensity associated with the CNV compared to a non-duplicated reference.

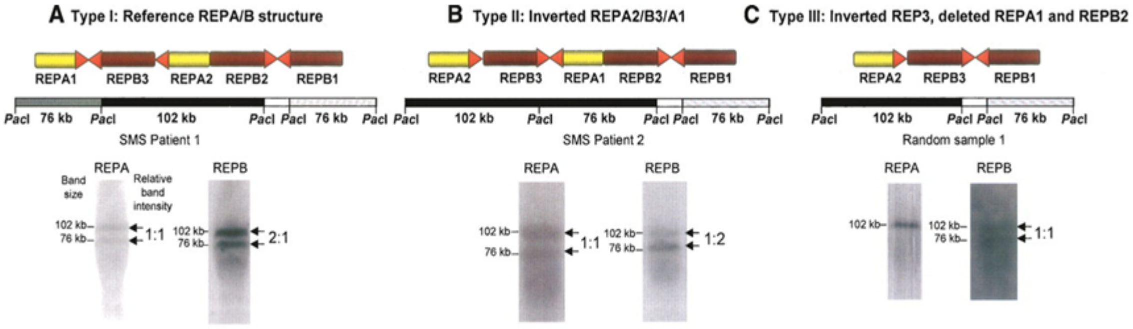

A Type I: Reference REPA/B structure
REPA1
REPB3
REPA2
REPB2
REPB1
PacI 76 kb PacI 102 kb PacI PacI 76 kb PacI
SMS Patient 1
REPA
REPB
Band size
Relative band intensity
102 kb
76 kb
1:1
2:1
B Type II: Inverted REPA2/B3/A1
REPA2
REPB3
REPA1
REPB2
REPB1
PacI 102 kb PacI 76 kb PacI PacI 76 kb PacI
SMS Patient 2
1:1
1:2
C Type III: Inverted REP3, deleted REPA1 and REPB2
REPA2
REPB3
REPB1
PacI 102 kb PacI PacI 76 kb PacI
Random sample 1
1:1

Figure 1.7: PFGE detection of REPA/REPB low copy repeat regions, involved in genomic rearrangements leading to Smith-Magenis syndrome. Differences in the hybridisation profiles obtained with *PacI* digestion are indicative of structural rearrangements, as indicated in the three panels. Band sizes are shown at the left side of the hybridised membrane, whereas relative band intensities are shown on the right [21].

1.3.3.2 Quantitative PCR (qPCR)

Real Time PCR can be used for direct CNV detection. This method is based on the simultaneous amplification of a DNA fragment known to be inside a CNV region, conjugated with a fluorescent dye, and of an internal amplification control that is non copy number variable, and labelled with a different dye [22]. The main feature of qPCR is that the concentration of amplified DNA is detected while the reaction progresses, in real time. The ratio of the threshold cycles (ct, defined as the fractional cycle where a threshold amount of amplified cDNA is produced during the exponential phase of the PCR reaction, detectable from the baseline noise) between the target gene and the non variable reference sequence generates ΔCt values which are used for CNV calculation. This technique presents a large dynamic range, being able to quantify products of 7-8 magnitudes differences, and high sensitivity, being able to amplify from 5-10 molecules of template. Hence its potentialities for quantitative data analysis have been widely explored in molecular medicine, biotechnology, microbiology and diagnostics.

However, the number or targets that can be interrogated in a reaction is limited by the number of fluorophores available. Moreover, the Ct method for qPCR quantification assumes the samples compared to have similar PCR efficiencies; being based on an exponential reaction, differences of 5% in the PCR efficiency among samples will generate 3-fold differences in the real DNA concentration after 25 cycles. Also the differences in base pair composition between the test and the internal control can bias the PCR reaction kinetics towards one product or the other, hence altering the CNV ratio. For this reason, different studies addressed the importance of performing accurate estimates of the PCR efficiency prior to reliable qPCR data analysis. One of the first solutions came from the use of dilution series of the sample to be tested, obtaining the PCR efficiency as slopes from plots of Ct versus the logarithm of the sample concentrations (**Figure 1.8**). Deviation from linearity indicates efficiency loss [23]. Another parameter considered by the same authors was the relative sensitivity constant for the test and the reference PCR assays, determined using cDNA samples with known concentrations. Another method (C_{y0}), not requiring an equal PCR efficiency, fits the Richard equation to qPCR data by nonlinear regression in order to obtain the best fit estimators of reaction parameters [24].

Another factor influencing real-time PCR results is the presence of biological inhibitors co-purified from the biological material the DNA was extracted from (bile salts, urea, immunoglobulin G) or present among the reagents used in the procedure of DNA extraction. The presence of inhibitors can generate strongly inaccurate quantitative results, and, in case of high inhibition degree, false-negative results can be obtained [25].

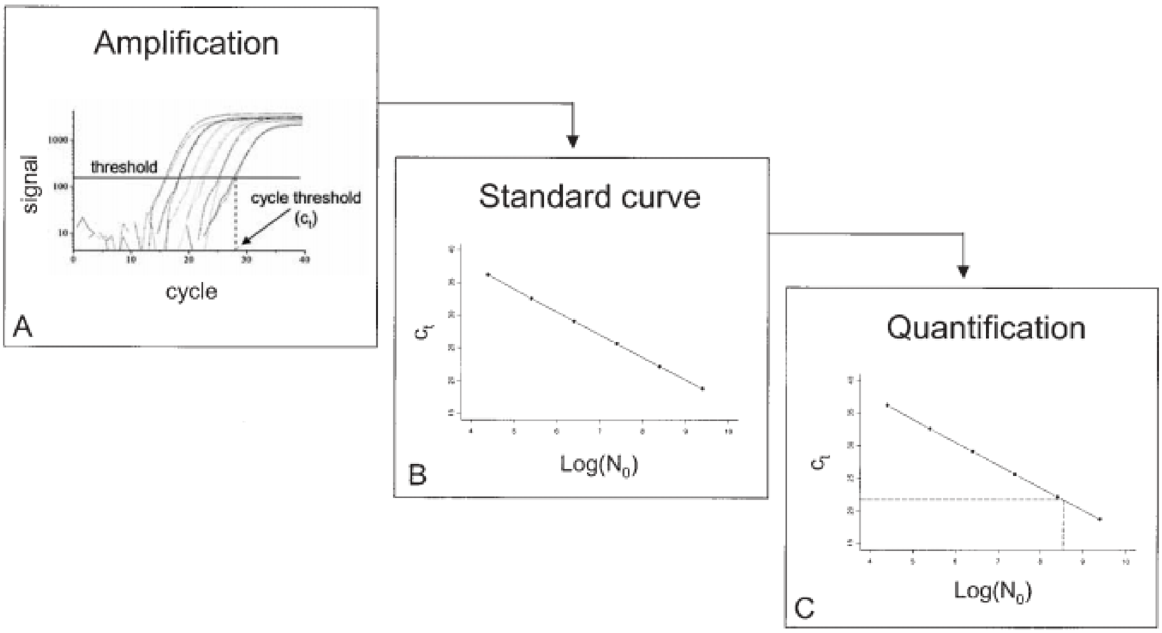


Figure 1.8: Flowchart of typical real-time PCR optimisation steps: panel (A) shows amplifications of standard dilutions. Panel (B) indicates the generation of the standard curve, plotting Ct values against the logarithm of the known template concentrations. Panel (C) shows the interpolation of the target quantification in an unknown sample [26].

Taken together, all these variables make qPCR optimisation a laborious process that does not easily provide the highest throughput required for large association studies, especially in case of DNA sample of different qualities and concentrations. A paradigmatic example of quantitative PCR used to determine CNVs is provided by a study by Chen Q *et al.* of 2006, where the authors used this method to measure beta defensin copy number [27]. To overcome the problems raised by different efficiency rates for the defensin gene to test and the reference (human serum albumin), the authors prepared a dedicated calibrator recombining one copy of the target gene and one copy of the reference in a plasmid. After correcting the PCR amplification efficiency, which differed between test and reference, and normalization by the calibrator, they determined the ratio of the copy number of the target gene to that of the reference gene in an unknown sample. This normalised ratio directly related to the gene copy number. Nevertheless, the use of a plasmid as calibrator does not solve possible problems arising from low quality DNA, presence of interfering secondary structures or the effect of co-purified inhibitors in the template.

An interesting study addressing the problem in qPCR accuracy for the calling of small CNV differences (for example the correct clustering of 4-copies and 5-copies individuals, that generates just a 1.25-fold difference in relative quantity) was performed by Weaver S *et al.* in 2010 [22]. The authors calculated an equation model to relate the number of replicates per assay per sample to a 95% Confidence Interval (CI). Considering that the standard deviation σ of the average value can depend on sample preparation, assay performance and instrument platform, they calculated the number of replicates needed for a reliable copy number call in presence of different system σ. In case of σ=0.16, 18 duplicates would be needed to discriminate 1.25 fold differences in relative quantities and, with σ values increasing to 0.25, the number of replicates needed would raise to 40. To discriminate higher copy numbers, the number of replicates increases dramatically, and to distinguish reliably 10 copies from 11 with a σ=0.16, at least 87 replicates should be run per each sample in analysis. The results of this study can help to explain the high variability observed among studies based on qPCR technology: an insufficient number of replicates can deeply affect copy number calls, with consequent inconsistent and fluctuating results.

Another critical study on the efficacy of qPCR in CNV detection utilised as calibrator a sample with known diploid copy number of 2 for *DEFB4*, *DEFB103A* and *DEFB104A*, as ascertained with Multiplex Amplifiable Probe Hybridisation (MAPH), a technique described in the next paragraph [28]. Moreover, as it will be presented later in this text (paragraph 1.8.3), the three genes analysed present a concordant number of copies, as demonstrated through MAPH and semi-quantitative fluorescence *in situ* hybridisation by Hollox EJ 2003 [29]. The study performed, demonstrated the same concordance just in 27 samples out of 60 when using qPCR, with concordance failure more frequent for samples with more than 4 copies. These results highlight the limit of real-time PCR for CNV quantification at high-copy number. Interestingly, the study of Chen *et al.* 2006 previously described collaterally confirm these limitations, as they found only loose correlation of beta-defensin copy numbers within the same sample with *DEFB4* always showing a higher copy number than *DEFB104*.

1.3.3.3 Multiplex amplification and probe hybridization (MAPH)

In the MAPH technique (**Figure 1.9**) 1 µg of genomic DNA is fixed to a membrane and hybridised with a set of probes corresponding to the target sequence to be detected [30]. The probes are generated by cloning the target sequences into a plasmid vector and then amplifying the cloned sequence using primers directed to the vector, with the result that all probes are then flanked with the same sequence. Probes must be designed with enough size difference to be resolved by capillary electrophoresis. After washing away any unbound probes, the specifically bound ones will be present in proportion to their target copy number. The probes are then stripped from the membrane and amplified simultaneously with universal primer pairs. The resulting products are separated by electrophoresis, and the comparison between the heights of the products and of the internal control probes allows the identification of copy number reductions or increases. This technique provides a good level of sensitivity and accuracy; however, it is complex and long to perform and requires a considerable DNA amount for each experiment. An interesting example of MAPH application is provided by a work of Hollox EJ *et al.* of 2005, where this technique was applied to test for correlation between beta-defensin cluster genomic copy number and lung disease associated with cystic fibrosis (CF). No association with CF was found, although the data obtained through MAPH present a good level of accuracy: the 95% confidence intervals for each copy number reading in the full cohort analysed (355 samples tested, 205 in duplicate) ranged from 0.05 to 1.2 with a median value of 0.24, confirming the reliability of this technique [31].

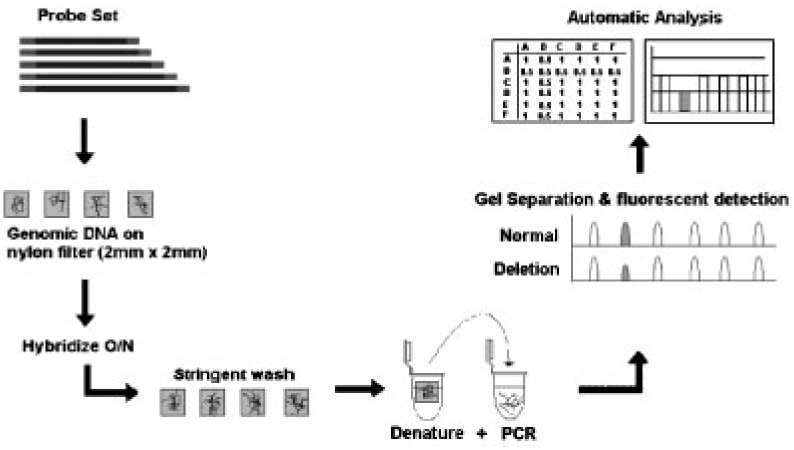


Figure 1.9: Scheme of MAPH protocol [32].

1.3.3.4 Multiplex ligation-dependent probe amplification (MLPA)

Multiplex ligation-dependent probe amplification was first presented by Schouten *et al.* in 2002. In this technique 100-200 ng of genomic DNA are hybridised in solution to probe sets, each of which is formed by two distinct halves. One half consists of a target specific sequence of 20-30 nucleotides flanked by a universal primer sequence; the other has a target specific sequence of 25-43 nucleotides at one end and a universal primer sequence at the other, but with a variable length random fragment in between of 19-130 nucleotides. The random fragment allows the size-based resolution of the probes through capillary electrophoresis (**Figure 1.10**) [33].

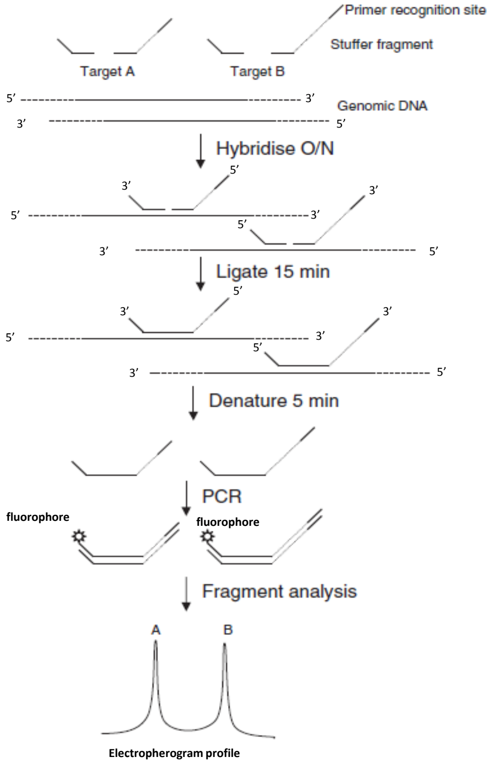


Figure 1.10 : MLPA scheme[32].

This particular probe design allows the binding of the target specific sequences adjacent to the target DNA which can then be joined using a ligase enzyme. This generates a contiguous probe flanked by universal primer binding sites amplifiable by PCR; no washing step is needed, since the unbound probe halves cannot be amplified. The amount of ligated probe produced will be proportional to the target copy number and can be quantified analysing the peak height after running the PCR products on capillary electrophoresis.

Disadvantages of this technique include a difficult probe design, which required cloning into M13 derived vectors and reduced hybridisation efficiency in case of polymorphisms or single base mutations, because of the short length of the specific probe region. On the other hand, MLPA requires much less DNA than MAPH, and as it is in liquid phase, it is more suitable for automation and multiplexing.

An example of MLPA used for defensins CNV typing is provided by a work of Groth M *et al.* of 2008 [28]: the authors designed 20 probes spanning the β-defensin cluster, 16 probes for the α-defensin cluster and 26 reference probes designed on bona fide single copy loci flanking the defensin clusters. The peak intensities from the areas in analysis were normalized against the summed peak areas of the ‘‘five nearest neighbour’’ (5nn) reference probes, for each individual sample (for a total of 42 samples analysed). Probes annealing to the same CNV were giving concordant peak intensities and defined as "relative locus doses". When relative locus dose values were plotted in increasing magnitude, plateaus indicated discrete copy number value (**Figure 1.11**).

Figure 1.11 shows another limitation of the MLPA method: the variable values obtained for *DEFA3* are false positives, as the probe used amplified also the paralogue *DEFA1* as a by-product. Moreover, it is possible to observe an increased fluctuation of the probes signals around an integer copy number for high-copy number values, where the discriminatory power of the technique decreases.

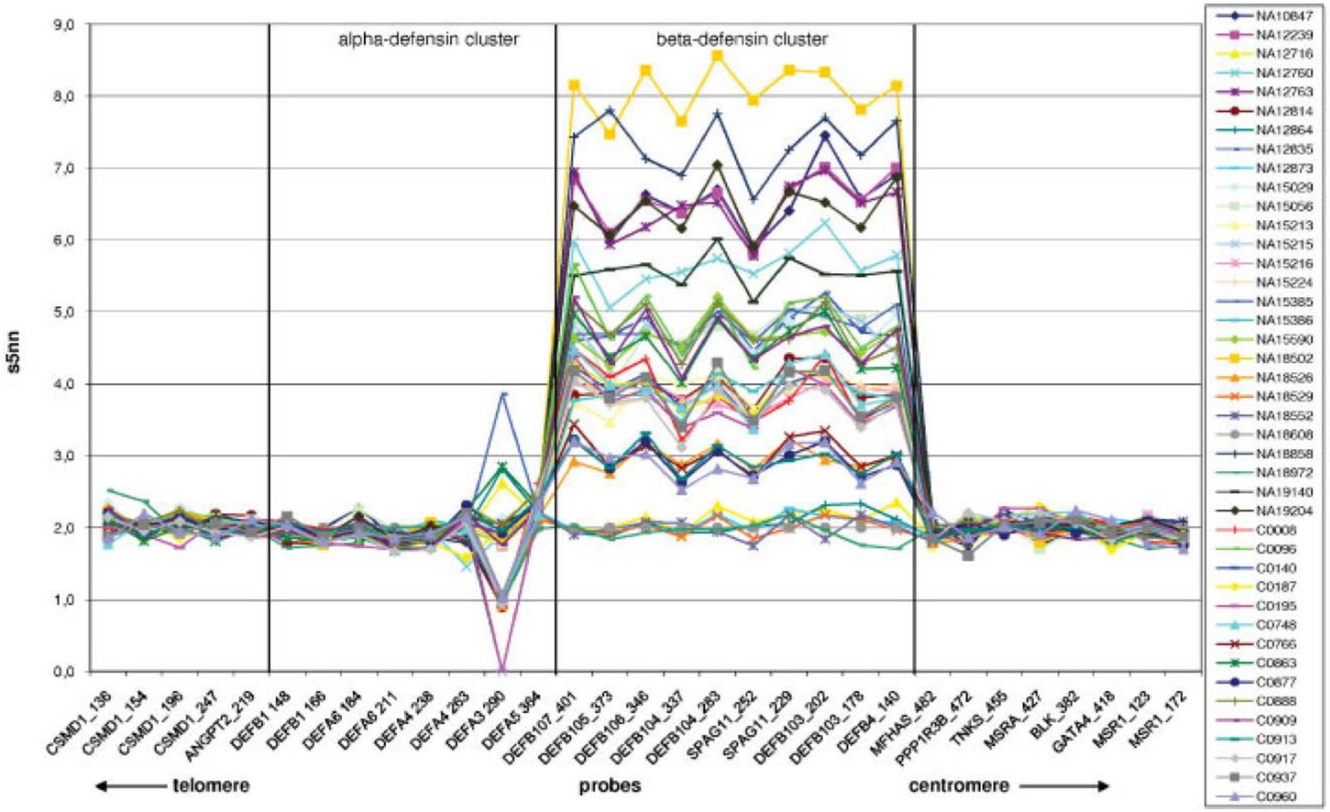


Figure 1.11: Scaled "five nearest neighbours" (5nn) values (y-axis) of chromosome 8 probes (x-axis). Each line represents the probe values for each of the 42 individuals analysed with MLPA by Groth M *et al.* 2008

1.3.3.5 Paralogue Ratio Test (PRT)

The Paralogue Ratio Test, also known as paralogous sequence quantification [34] is a PCR-based method designed in 2007 by Armour and colleagues [35]. They adapted the quantitative multiplex PCR approach designing primers specific for repeated DNA elements. In particular, the primers are accurately designed in order to amplify from a copy of one element within the variable repeat unit, plus exactly one other unlinked reference locus with similar size falling in a non CNV region (**Figure 1.12**). Through capillary electrophoresis is possible the size-based discrimination of the two PCR products to calculate the ratio of the area under the peaks between the test and the reference. This allows a precise detection of the diploid copy number of the region of interest for each sample analysed.

This technique requires only small amounts of genomic DNA (10-20 ng) and its accuracy is comparable to MAPH, MLPA and array-CGH [35]. Moreover, it is less expensive and time consuming, and thanks to its high throughput it can be used for CNV typing of large cohort of samples. The main limitation of this technique is that it is not always possible to find a single primer pair specific for a test and a reference locus, especially in case of small CNVs. Despite this, most CNVs encompass segmental duplications as they often mediate the events of CNVs formation; an accurate primer design focussed on divergent repeated elements can increase the chance of obtaining suitable PRT candidates.

For all the advantages previously described, we selected the PRT as the most appropriate method for CNV typing during this study.

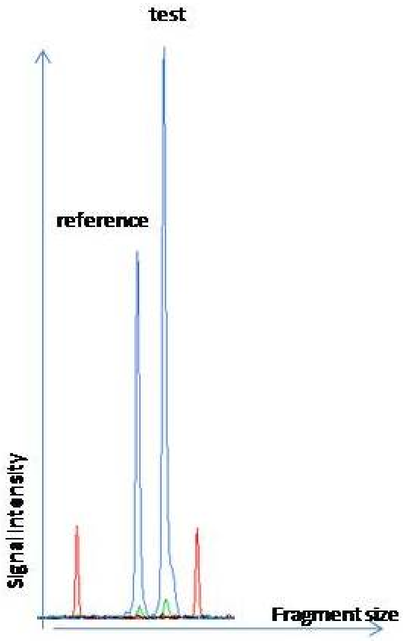


Figure 1.12: Scheme of a PRT electropherogram. Co-amplified PCR products of different size spanning a putative CNV locus (test) and a known non-CNV region (reference) are run on capillary electrophoresis to generate two peaks (in blue) of different signal intensity. The red peaks correspond to ROX 500 internal lane size standards.

1.3.3.6 Droplet digital PCR (ddPCR)

Droplet digital PCR is a newly commercialised technique that represents the latest advancement for DNA quantification, gene expression and CNV detection. The combination of limiting dilutions, end-point PCR, and Poisson statistics for absolute DNA quantification has been proposed in the early 90s [36] and constituted the basis for a new technology development.

ddPCR is based on the generation of discrete droplet partitions containing randomly distributed DNA or cDNA template, to the point where some contain no template and other contain one or more copies of nucleic acid. The generation of droplets is described in **Figure 1.13**.

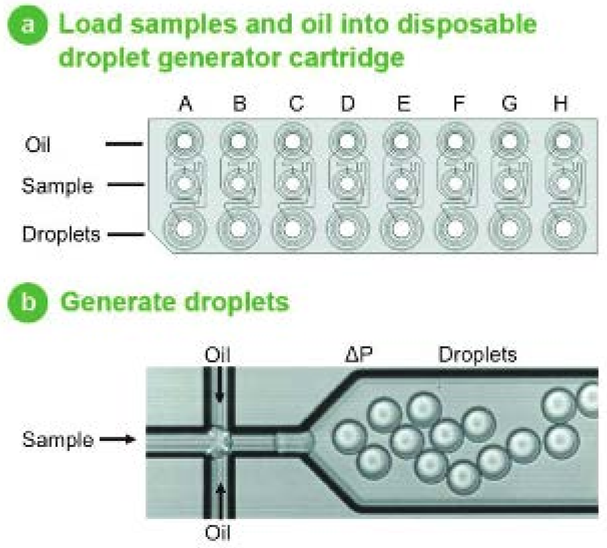


Figure 1.13: Panel (A) shows the eight-channel droplet generation cartridge used for the generation of discrete partitions. Samples and droplet generation oil are loaded in the dedicated wells. Panel (B) shows the principle of droplets generation: vacuum is applied to the droplet well, using a dedicated vacuum generator. It draws sample and oil through a flow-focussing nozzle that helps the formation of monodisperse 1nL droplet. In less than 2 minutes eight samples can be converted into eight sets of approximately 20000 droplets. Adapted from Hindson BJ *et al.* 2011 [37].

The droplet emulsion is then thermally cycled to end-point. Each PCR reaction is duplexed to contain primers for the amplification of the test and the reference locus together with Taqman® probes conjugated to two different fluorophores specific for each of the amplified products. The partitions are then read using a droplet reader

(**Figure 1.14**) which sips droplets from each well streaming them in single-file past a two-colour detector. Each droplet is assigned as positive or negative based on their fluorescence amplitude.

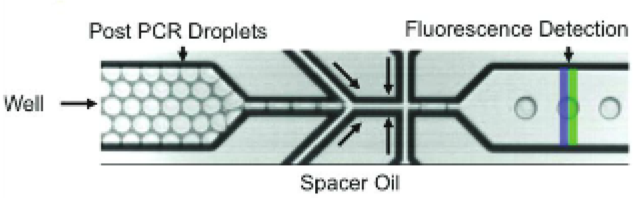


Figure 1.14: Scheme showing the functioning of a droplet reader: the droplets generated per each sample are aspirated into a single line column and the fluorescence amplitude of each drop is detected by a two-colour reader. Adapted from Hindson BJ *et al.* 2011 [37].

Following detection it is possible to determine the fraction of positive partitions and to calculate the concentration of the nucleic acid in analysis, applying the following equation: λ= - ln(1-*p*) where λ is the average number of target DNA molecules per replicate reaction and *p* is the fraction of positive end-point reactions.

The difference in fluorescence amplitude divides the droplet population into four discrete clusters: no target (-/-), one of the target (+/-, -/+) or both targets (+/+). A fluorescence threshold is set per each detection channel and following Poisson modelling it is possible to calculate the copy number as the ratio between the number of droplets positive for the target DNA and the number of droplets positive for the reference DNA (**Figure 1.15**).

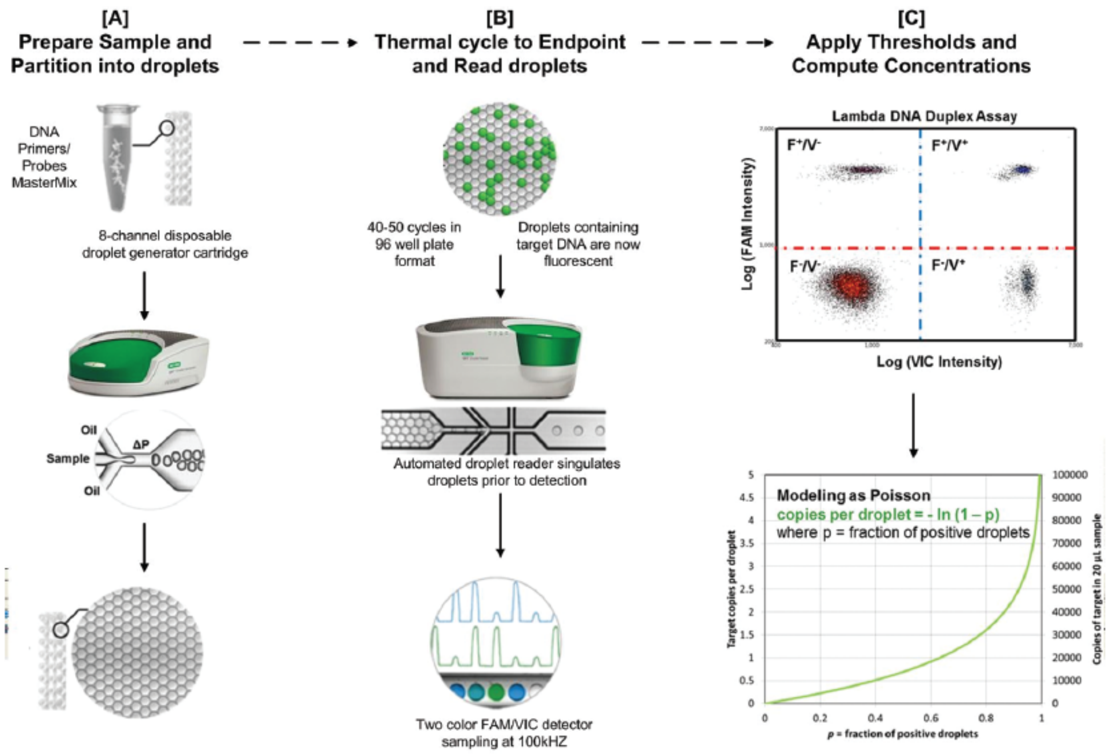

[A]
Prepare Sample and Partition into droplets
[B]
Thermal cycle to Endpoint and Read droplets
[C]
Apply Thresholds and Compute Concentrations
DNA
Primers/
Probes
MasterMix
8-channel disposable droplet generator cartridge
Oil
Sample
Oil
ΔP
40-50 cycles in 96 well plate format
Droplets containing target DNA are now fluorescent
Automated droplet reader singulates droplets prior to detection
Two color FAM/VIC detector sampling at 100kHZ
Lambda DNA Duplex Assay
F+/V-
F+/V+
F-/V-
F-/V+
Log (FAM Intensity)
Log (VIC Intensity)
Modeling as Poisson
copies per droplet = - ln (1 – p)
where p = fraction of positive droplets
Target copies per droplet
Copies of target in 20 µL sample
p = fraction of positive droplets
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
0
0.2
0.4
0.6
0.8
1

Figure 1.15: Droplet digital PCR general overflow [38].

The first studies that used this technology for copy number calling or viral load detection have been quite promising [37] [38] [39]: Hindson BJ *et al.* in 2011 retyped 7 HapMap samples with known copy number for three genes of interest, with copy number ranging between 1 and 6, obtaining a really precise estimate with just the droplets generated from a single well per sample, demonstrating the advantage of not requiring replicates. Thirteen samples have been retyped by the same authors for the gene *CCL3L1*, characterised by a wider copy number range: in this case the Poisson 95% Confidence Intervals for each copy number determination are wider as the copy number estimate increases (**Figure 1.16**).

ddPCR shows a great potential as new technology for CNV typing and its format is suitable for large scale typing, for the low amount of template required (10ng per reaction), the accuracy, with no need for replicates, and the short time of execution compared to other more laborious techniques previously discussed. The only drawbacks are constituted by the initial cost of the dedicated equipment and by the importance of accurate droplet generation for reliable CNV calling. For all these reasons, ddPCR is used in this project as confirmatory method to PRT in copy number estimates.

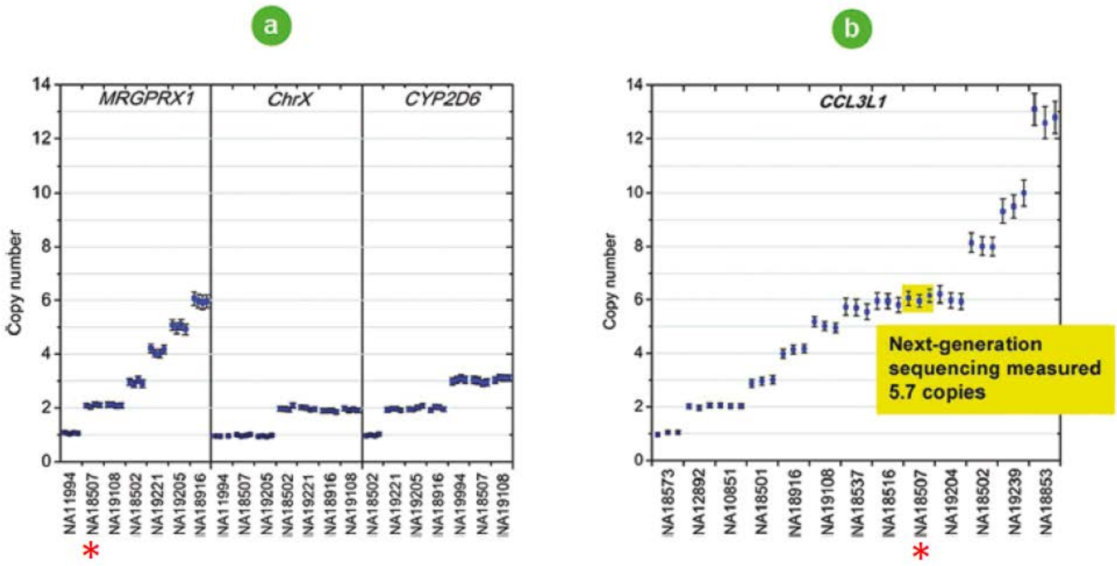


Figure 1.16: Panel (A) shows the copy number estimates for 7 HapMap samples retyped for the gene MRGPRX1, for chromosome X and for the gene CYP2D6. Panel (B) presents the copy number estimates for the gene CCL3L1, implicated in HIV-1/AIDS susceptibility. The sample NA18507 (red asterisk), when typed with next generation sequencing (Illumina WGS), gave a less precise copy number estimate of 5.7 [40]. The error bars indicate the Poisson 95% Confidence Intervals for each copy number determination.

1.3.3.7 Advantages and disadvantages of copy number typing methods

All the methods previously described for copy number typing present advantages and disadvantages that should be carefully considered in the selection process of the most appropriate technique to answer a specific biological question. **Table 1.1** presents a summary of the characteristics of each technique hereby presented.

As previously described in this chapter, fibre-FISH represents the most accurate method for CNV typing, being based on absolute CN count of closely spaced probes hybridised to stretched chromosome fibres. Nevertheless, the applicability of this technique is hampered by its difficulty of execution, as signals coming from broken or overlapping DNA fibres might be misleading, and by the necessity of establishing cell lines for each sample in analysis. Despite these caveats, fibre-FISH is a valuable approach to generate and validate control samples of known copy number to be used as internal calibrator for more high-throughput approaches. Similarly, a Southern blot approach can be used as a complementary validation technique, especially in combination with PFGE, as it is very low-throughput, for the workload involved and the high amount of DNA required for each experiment.

The qPCR technology represented a great innovation in terms of throughput, DNA amount required and time of execution but, as presented in paragraph 1.3.3.2, its use for copy number typing is challenged by low accuracy and reproducibility, especially for high copy number calling.

MAPH and MLPA are two hybridisation-based techniques that provide good levels of sensitivity and accuracy; disadvantages of the former include time of execution and amount of DNA required, whereas for the latter the major issue is associated with a very difficult probe design. PRT and digital droplet PCR are, at the present state of the art, the best methods available for CNV typing, in terms of accuracy, precision, quantity of DNA required, throughput and time to results. In particular, PRT is more time consuming in terms of primers design, as it is not always possible to generate valid assays for each given CNV area of interest, but it is counterbalanced by a slightly higher accuracy when compared to ddPCR.

In terms of CNV mapping, array CGH platforms perform better than SNP arrays, with better signal-to-noise ratios and greater accuracy for the detection of deletions and small copy number differences. Next generation sequencing approaches present the advantage of calling absolute copy number, despite the short-length reads produced can generate issues for an accurate mapping, especially within complex genomic areas rich in segmental duplications.

Technique	Detection	Sample	Loci	Through put	Minimum resolution	Cost per sample	Time to result	Labour requirement
Fibre-FISH	Absolute CN	Cells	Single	Low	>1 kb	Low	>48 h	High
Southern blot	Inferred absolute CN/change from diploid dosage	2-5µg DNA	Single	Low	>1 kb	Low	2-3 days	High
PFGE	Inferred absolute CN	2-5µg DNA	Single	Low	0.5-1kb	Low	2-3 days	High
qPCR	Change from diploid dosage	5-10ng DNA	Single	High	100bp	Low	4 hours	Low
MAPH	Change from diploid dosage	0.5-1µg DNA	>40	High	100bp	Low	>24 h	Low
MLPA	Change from diploid dosage	100-200ng DNA	>40	High	100bp	Low	>24 h	Low
PRT	Change from diploid dosage	10-20ng DNA	Single	High	100bp	Low	4 hours	Low
ddPCR	Change from diploid dosage	5-10 ng DNA	Single	High	100bp	Low	4 hours	Low
SNP array	Change from diploid dosage	0.5-1µg DNA	>2million	High	5-10kb	Moderate	>24 h	Moderate
Array CGH	Change from diploid dosage	0.5-1µg DNA	>2million	High	5-10kb	Moderate	>24 h	Moderate
NGS	Absolute CN	1-2µg DNA	Genome-wide	Low-moderate	>1kb	High	2-3 days	High

Table 1.1: Summary of the main characteristics of the main methods for CNV detection. Adapted from Cantsilieris S *et al.* 2012 [41]**.**

1.4 Extent of CNVs in non-human primates

The discovery of copy number variations in humans and the increasing interest of the scientific community on their distribution and roles in human evolution and pathology, naturally brought the necessity of a better understanding of the CNV pattern in non-human primates. It became clear that sequence gain and loss events could alter the gene complement and expression of an organism, bringing to phenotypic variations subjected to selective pressures, as altered environmental conditions or infectious diseases. Conceptually, the presence of 414 living species of primates [42] is an extraordinary resource to understand how evolution shaped the genome architecture in different species. With the advent of high-throughput techniques, as microarray platforms first and next generation sequencing later on, the questions on the role of CNVs in evolution could start to be addressed.

1.4.1 Genome-wide studies for CNV discovery in primates

The first study aimed at the identification of large-structural rearrangements in primates has been performed in 2003, using BAC-array comparative genomic hybridisation. The authors used 2460 large-insert human clones, as reference DNA, with an average spacing of 1.4 Mb, to detect fixed inter-specific copy number differences between the genomes of humans and great apes. In particular they analysed common chimpanzee, bonobo (*Pan* genus diverged around 4-7 MYA), gorilla (*Gorilla* genus that diverged 6-9 MYA) and orangutan (*Pongo* genus 12-16 MYA). They found 63 variant sites, mostly lineage-specific, ranging in size from 40 kb to 175 kb; despite the low resolution of the study, it was already possible to observe a trend correlating the divergence time of the species in analysis with the number of rearrangements observed, with orangutan showing the greatest number of ratio differences and chimpanzee species the lowest number, when compared to the human genome [43]. The authors also realised that, contrarily to what had previously been reported [44] [45], structural rearrangements were not as strongly biased towards pericentromeric and heterochromatic regions, as they were also present in gene-rich euchromatic areas. Another novelty was the report of segmental duplications in most of the variable regions identified, an aspect that was then extensively reported and studied over the following years, for their role in copy number formation mechanisms.

Confirmations of this last aspect came shortly after from a study by Perry GH *et al.*, in 2006, aimed at the identification of CNV differences among 20 wild-born western chimpanzees and humans, using a human BAC library (2632 clones) on BAC-CGH platforms, as the mean sequence divergence between the two species in analysis is <2%. Also in this case it was possible to observe an enrichment of segmental duplications in the BAC clones spanning or flanking the CNV areas. In particular, a 20 fold increase in segmental duplications was reported for CNV regions shared between humans and chimpanzees. These data are consistent with the hypothesis of CNV hotspots determined by non-allelic homologous recombination (NAHR, presented in paragraph 1.6.1) between ancient segmental duplications (**Figure 1.17**). It is noticeable that CNVs detected in orthologous regions are observed considerably more often than would be expected by chance and likely reflects recurrent CNV formation instead of an "identity-by-descent" transmission. This theory is supported by the fact that the coalescence time for human genomic loci is less than two million years [46] while human and chimpanzee lines diverged approximately 4-7 MYA.

Authors identified 355 CNVs, in contrast with just 255 loci when the same array was used for CNV detection on 55 non-related human samples of diverse ancestries [7]; this suggested a higher genetic diversity in the chimpanzee population than previously believed, that could be due to species-specific differences in selective pressures on CNV formation and maintenance and/or higher rates of duplication and deletion in chimpanzees. Nevertheless, the authors were aware of the clear underestimation performed on the CNV numbers identified, and further studies were then conducted to support the idea of common unstable genomic areas more prone to drive evolutionary processes [47].

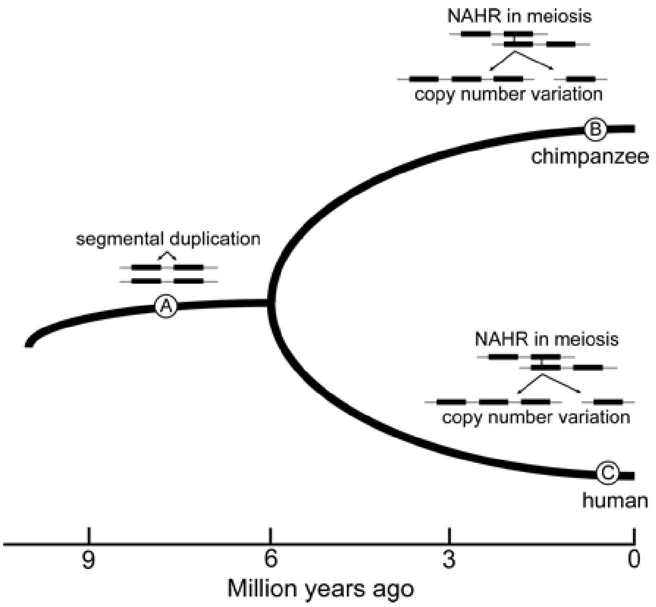


Figure 1.17: Model for evolution of CNV hotspots. Certain segmental duplications that arose in a human–chimpanzee common ancestor (A) may facilitate separate non-allelic homologous recombination (NAHR) in both chimpanzees (B) and humans (C), leading to the genesis of CNVs in both species. If NAHR in these regions occurs frequently, it may be expected to lead to the maintenance of common CNVs by way of recurrent duplications and deletions [47].

Perry GH and colleagues further investigated the evolution of CN differences between humans and chimpanzee with greater detail [48], using human tiling-path array CGH platforms that allowed an increase in study resolution to ~1kb, compared to previous resolution power of ~250kb obtained in their previous work with BAC arrays. The analysis of 30 chimpanzees and 30 humans from different populations confirmed the previously reported data of a copy number enrichment in orthologous regions between human and chimp, more frequently than what is expected by chance, and strongly associated with the presence of highly homologous segmental duplications. The resolution of the study allowed then to add an extra level of analysis, performing tests of neutrality with the identification of functional classes of copy number variable genes that likely evolved under purifying or positive selection. The stronger signals came from genes with inflammatory response functions; interestingly, genes such as *APOL1* (Apolipoprotein L-1, involved in trypanosome parasites resistance, causing sleeping sickness) and members of the interleukin-1 family, were completely deleted from the chimpanzee genome, suggestive of purifying selection due to a different pathogenic pressure, with the development of different pathways to regulate inflammation and infections in the chimpanzee lineage.

The natural following step towards a better understanding of the CNV distribution and role in primates came from a more targeted analysis to identify copy number variable genes in different primate species. For the identification of genome-wide gene duplications and losses, the best method available was a cDNA based array-CGH platform. This was performed by Dumas L *et al.*, in 2007 [49] on 10 primate species (human, bonobo, chimpanzee, gorilla, orangutan, gibbon, macaque, baboon, marmoset and lemur), using a human cDNA array covering 24473 human genes. The authors predicted 4159 genes to have lineage-specific changes in the species analysed, more specifically 84 in human, 79 shared between *Pan* lineages, 102 in gorilla, 117 in orangutan, 549 in gibbon, 369 in Old World monkeys (macaque and baboon), 543 in marmoset, and 1209 in lemur. Also, gene duplications outnumbered gene losses, with 23 copy number increases present just in the Great Apes lineage and absent in humans. Despite the presence of possible problems of sequence divergence affecting the hybridisation efficiency in human-based cDNA arrays (**Figure 1.18**), especially for the species with the most distant last recent common ancestor with humans (marmoset and lemur), it was still possible to observe a strong pattern of gene duplications impacting the evolution of the primates genome. Nevertheless, from this study design it is not possible to evince conclusions on the functional status of the variable genes identified.

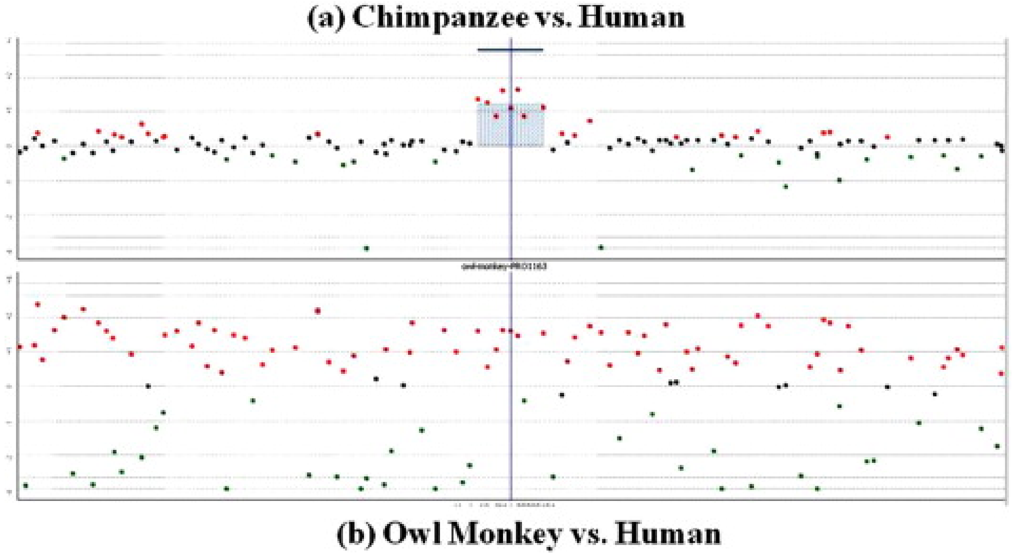


Figure 1.18: Examples of CNV calls on the same 105K Agilent "human" array against the same human reference individual. The snapshot shows differences in $\log_2$ ratios distribution within a 2.65 Mb region on chromosome 2 (based on human genome) with (a) a chimpanzee sample and (b) an owl monkey sample. Sequence mismatches can influence the signal intensity and hence the observed $\log_2$ ratios [50].

Interestingly, the comparison of the variable regions identified with the available genomic assemblies of the species in the analysis revealed a good level of concordance in mapping to gene areas for human and chimpanzee (84% and 81%) and a low concordance for rhesus macaque (33%). This controversial data could be ascribed to the presence of large gaps in the rhesus macaque assembly, due to the technical difficulties of assembling copy number variable regions.

The discovery of extensive gene gains and losses in the primates genome raised the question of quantifying these variation rates; this aspect was even more intriguing by the fact that the nucleotide substitution rates in primates are slower compared to other mammalian species, as rodents [51], with the molecular clock being even slower in hominoids (humans and chimpanzees) compared to other primates [52]. The result of this phenomenon is the high percentage of identity at orthologous nucleotides between humans and chimpanzees, unable to explain, alone, their great inter-species diversity. Structural variations have been hence studied as alternative molecular mechanism driving evolution and diversity. This analysis has been performed by Hahn M *et al.* in 2007 [53], applying a maximum likelihood approach on 9990 gene families sizes in macaque, human, chimpanzee, rat, mouse, and dog, retrieving the data from their respective published genomes, to estimate the rate of gene turnover. As shown in **Figure 1.19**, the best fitting model presents the same accelerated gene gain and loss rate for human and chimpanzee, an intermediate rate for rhesus macaque and great apes ancestor and a slower rate for the other mammalian species in the analysis. A possible mechanism explaining these structural changes is the great expansion of transposable elements that dates back to approximately 35 MYA, that could have driven an increased number of non-homologous allelic recombination events [54], together with a faster fixation of duplicated genes in species with smaller effective population sizes [55] .

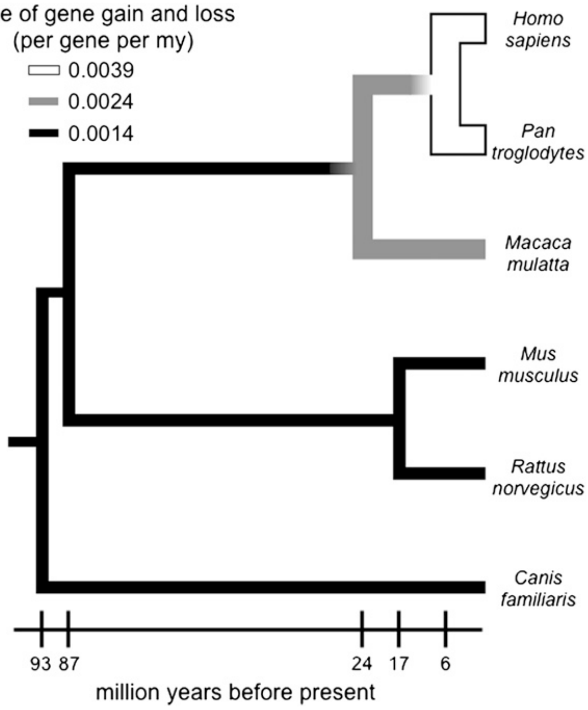


Figure 1.19: Rates of gene gain and loss across different mammalian species. The species tree of the six mammalian genomes considered in Hahn *et al.* study is shown, shaded according to the estimated rates of gene gain and loss [53].

Also, using the same maximum likelihood approach, the authors realised that each primate species in the analysis presented different species-specific rates of gene gain and loss for single gene families, as presented in **Figure 1.20**. 180 gene families showed important expansions or contractions, unlikely to be caused by random gene gain or loss events (p<0.0001); this genomic organisation can be ascribed to adaptive selection acting on specific gene families and promoting structural changes. Nevertheless, the computational model applied utilised a lower threshold of 20 kb for the duplication and deletion events calling, hence excluding all possible smaller CNVs, whose role in evolution remained to be identified.

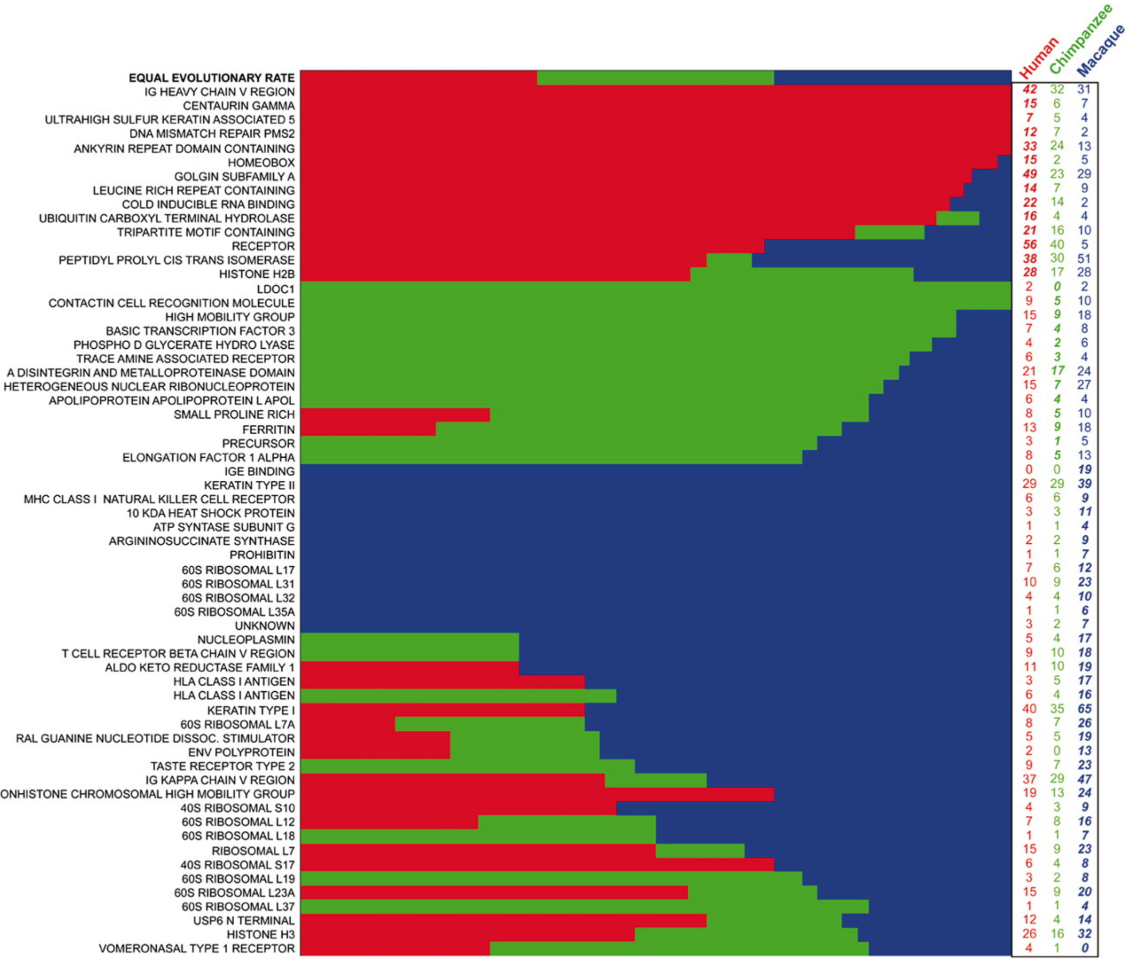


Figure 1.20: Rapidly evolving gene families. Individual families showing significantly accelerated rates of evolution along the human, chimpanzee, and macaque lineages are shown. Each row is a single gene family, with the relative rate of evolution along the human (red), chimpanzee (green), and macaque (blue) lineages given by the width of the coloured bars. The numbers on the right indicate the size of the family in each of the three species. From Hahn M *et al.* 2007 [53].

With the advent of high-throughput next generation sequencing technique in the last years, it became possible to overcome the limitations arising with the use of CGH platforms, as the limited resolution, the bias due to the use of human probes for the detection of CNVs in other primates and inaccurate copy number detection in complex areas due to genomic assembly problems of the species in analysis.

Sudmant and colleagues performed a key study in 2013 exploring the advantages of the Illumina HiSeq 2000 next-generation sequencing platform [56]; they constructed accurate maps of deletions and segmental duplications on 97 great apes individuals (Bornean and Sumatran orangutans, the four chimpanzee subspecies, bonobos, Eastern and Western gorillas, 10 humans from different populations and a high-coverage archaic Denisovan individual), with a median coverage of 25X. Firstly, they confirmed that fixed, lineage specific, CNVs are non-randomly distributed and >20% map within 5kb of shared ancestral duplications. On the contrary, deletions followed a random distribution pattern in the Great Ape genomes. Nevertheless, the chimpanzee lineage showed a heavier burden for fixed deletions that could be due to a recent bottleneck and to a dramatic decline in their effective population size. More generally, this may suggest that fluctuations in the effective population size could play a significant role in the rates of genomic deletions and duplications of a certain lineage [57]. With this approach, the authors identified 11836 fixed duplicated loci, 5528 fixed deletions (seen as homozygous events in most individuals) and 6406 private (observed only once) and polymorphic CNVs. Among those, it was possible to identify 407 lineage-specific gene duplications and 340 gene deletions with complete or partial exon loss. The gene families more involved in deletion events appeared to be olfaction genes, immunity related genes, drug detoxification and sperm surface membrane genes. Of the duplicated genes, 33 were found to be duplicated just in the human lineage, after the divergence from chimpanzee; these include *BOLA2*, which deletion has been associated to developmental delays and features of autism [58]. Interestingly, in disagreement with the study of Hahn M *et al.* of 2007 previously described [53], humans and chimpanzee did not show an increased copy number polymorphisms burden compared to the other Great Apes analysed; this possibly suggests that the increased CNVs detection power allowed by high coverage next generation sequencing can help to limit false positive signals.

1.4.2 Examples of evolving CNVs in primates

Several studies have indicated the presence of copy number differences in multiple chromosomal regions among humans, chimpanzees, great apes and Old World monkeys[59];[60];[61].

Part of these variations are lineage-specific and, according to different studies [12] [62], the last 5-6 million years saw a strong expansion of CNV regions which highly diversified humans from other primate species (**Figure 1.19**). For example it is estimated [53] that about 6% of human genes do not have a one to one counterpart in the chimpanzee genome. This phenomenon could help to explain the physical and behavioural differences of human species, given the positive selection evidenced on its nucleotide sequence when compared to other primates. Thus, selection could have played an important role in shaping copy number differences among primates. Nevertheless, these results could be an artefact because they are based on the reliability of the mammalian genome assemblies considered. As mentioned before, the quality of the human genome assembly is far better than the WGS assemblies available for other primate species; it is therefore possible that the higher quality of the human CNVs annotation could have biased the results of these studies.

One interesting example of gene duplication deeply impacting on the evolution of apes and Old World Monkeys (identified together as Catarrhini clade) is their development of trichromatic vision. The ability to distinguish green and red derived from a duplication of the long-wavelength-sensitive (LWS) gene on the X chromosome, that, in most extant catarrhines, gave rise to two functional copies of the ancestral gene, distinguishable by three polymorphic amino acid substitutions (*OPN1LW* and *OPN1MW*) [63]. This event took place at the start of the catarrhini radiation, 30-40 MYA, as New World Monkeys possess just one copy of the ancestral gene. Interestingly, howler monkey (*Alouatta* genus, from the Platyrrhini clade) is the only New World Monkey that evolved trichromatic vision as well, as result of an independent duplication event of the same ancestral gene, for convergent evolution. Other New World Monkeys have a striking intermediate opsins set, with polymorphic variants of the LWS gene, the same observed in the catarrhini duplicated copies (**Figure 1.21**). As a consequence of random X inactivation, females polymorphic for these three substitutions will be mosaic for green and red opsins; on the contrary, males will always be dichromatic, because of hemizygosity. Almost all the other Eutherian mammals possess just two photopigments, produced by the LWS gene (red-green) and by the autosomal short-wavelength-sensitive SWS1 (blue). Trichromatic vision may have been an advantageous trait for species adapting to diurnal activities, conferring increased abilities in discriminating colours of fruits and plants.

In parallel with the development of trichromatic vision, it is possible to observe a dramatic increase in the number of pseudogenised olfactory receptors (OR). Noticeably, OR genes represent the largest gene superfamily in mammalian genomes, they are organised in different clusters in extremely dynamic genomic areas, prone to events of gene duplications and gene conversions [64]. The study of Gilad Y *et al.* 2007 investigated the proportion of OR pseudogenes in 19 primate species, highlighting how this proportion decreases from 60% in human, to 30% in non-human primates to 20% in other mammals as dog or mouse. Their most striking finding was that the proportion of OR pseudogenes was not high just in Old World Monkeys and Apes, that could have been consistent with the hypothesis of a catarrhini-clade-specific functional loss, but also in the howler monkey, the only New World Monkey to have a fully trichromatic vision for an independent duplication event (**Figure 1.21**). Hence, the deterioration of the olfactory repertoire occurred in concomitance with the evolution of full trichromatic vision in two separate primate lineages, suggesting an exchange in the importance of olfaction and sight in the evolution of primates.

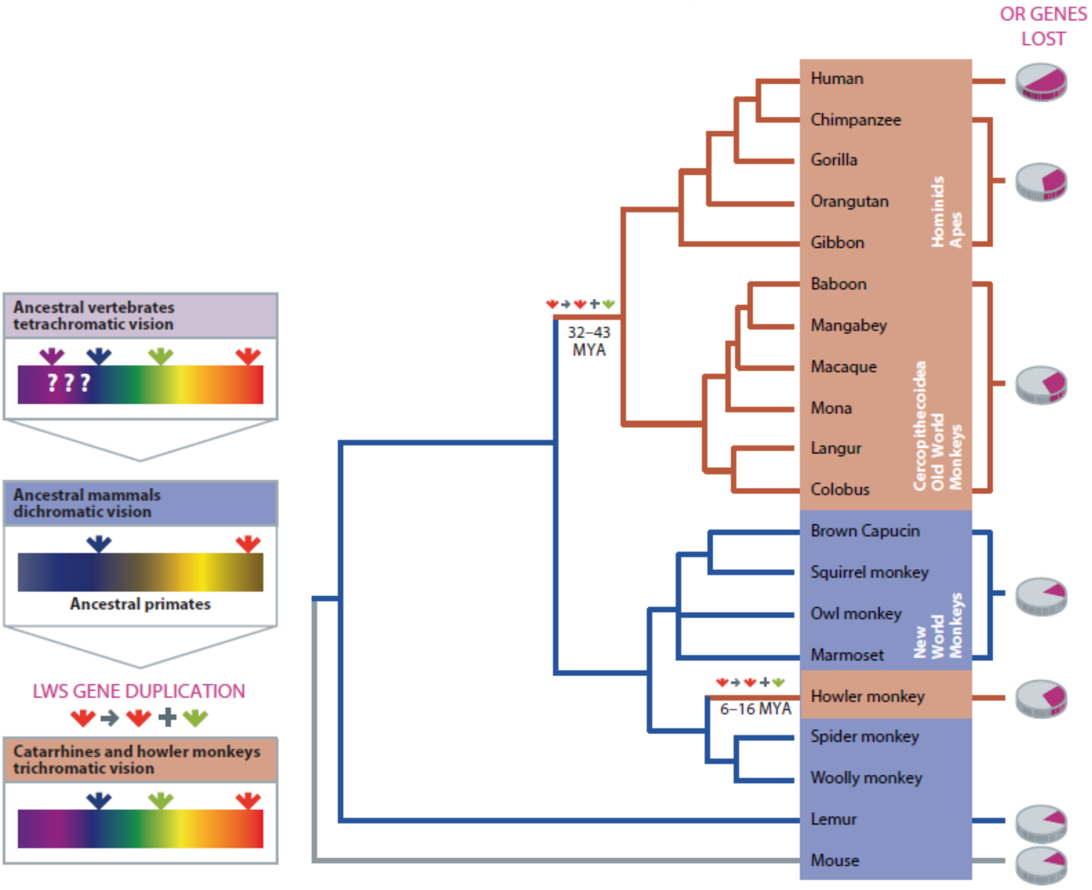


Figure 1.21: Origin of trichromatic vision in catarrhines. On the left it is shown the ancestral tetrachromatic vision of vertebrates, able to distinguish three colours of the visible light spectrum and one colour in the UV spectrum. The adaptation to nocturnal vision, brought the ancestors of modern mammals to retain just the short (blue) and long (green-red) wavelength colours. 30-40 MYA the ancestor of the catarrhini clade acquired the ability to distinguish red and green, firstly arisen as polymorphism of the LWS gene. The polymorphism has been then fixed independently via gene duplication in the catarrhini clade and in the New World howler monkeys. On the right, the pie charts represent the percentage of olfactory receptors (OR) pseudogenised in each primate group (estimates from [65]). Figure taken from Human Evolutionary Genetics second edition 2013, chapter 7.

Another important example of the role of gene duplication in organisms adaptation and evolution is given by the pancreatic ribonuclease gene (*RNASE1*) that codes for a ribonuclease enzyme capable of degrading bacterial RNA. All mammals possess one copy of this gene, with the exception of ruminants and the leaf-eating colobine monkey [66]. Interestingly, this subfamily of Old World Monkeys are mostly exclusive herbivores using leaves as primary food source, through fermentation by symbiotic bacteria in the foregut. The colobines are able to recover nutrients through degradation and digestion of bacteria with different enzymes, as RNASE1, secreted from the pancreas and transported to the small intestine, where it exerts its degrading functions. Noticeably, the pancreas of foregut fermenting animals, such as colobines and ruminants, produces greater amount of ribonuclease, compared to other mammals [67]. It is thought that these higher concentrations are due to the fact that rapidly growing bacteria have the highest ratio of RNA-nitrogen to total nitrogen of all cells, so higher concentrations of RNase are needed to degrade bacterial RNAs for efficient nitrogen recycling [68]. Ribonucleases are also present in other mammals, where they play non-digestive related roles as host-defence barrier against pathogenic viruses [69]. This observation of higher ribonuclease levels in colobines brought in 2002 to the discovery of a recent gene duplication of the ribonuclease gene, dating to 4 MYA, that gave rise to its duplicated copy RNASE1 B [66]. Interestingly, the *RNASE1* copy did not change after duplication, whereas *RNASE1 B* accumulated many substitutions. Also, *RNASE1 B* showed signs of positive selection, with a high rate of charge-changing substitutions; these reduce the net charge of RNASE1B protein from 8.8 to 0.8 (at pH 7) and the isoelectric point from 9.1 to 7.3. Given the negative charge of RNA, net changes in ribonuclease charge influence its catalytic activity and its optimal pH. The optimal pH for humans, primates and colobines RNASE1 is 7.4, compatible with a pH range of 7.4-8 measured in human and primates small intestine [70]; on the contrary, because of the foregut fermentation and the consequent changes in digestive physiology, colobines small intestine pH lowers to 6-7 and strikingly, the optimal pH for RNASE1 B was measured to be 6.3, level at which it is six times more active in degrading RNA compared to RNASE1. It is hence plausible that in colobines the RNAse digestive activity is performed just by RNASE1B, leaving its ancestral paralogue RNASE1 to a more specific antiviral activity. These findings demonstrated that the duplication and the rapid amino acid substitutions observed in *RNASE1 B* were driven by selection, enhancing its activity in environments with lower pH and favouring a functional specialisation of the newly duplicated genes.

RNASE1 genes followed a parallel evolution in ruminants, which posses three ribonuclease paralogues, as consequence of a double round of duplication events that took place ~40 MYA. Interestingly, ruminant paralogues present the same functional diversification observed in colobines but accumulated substitutions at different sites. This demonstrates how evolutionary adaptation can have different solutions for the effect of the stochastic mutation and drift processes and of the genetic background of the organism considered [71].

Another remarkable example of gene duplication in response to adaptation is constituted by the salivary amylase gene *AMY1*. Salivary amylase is responsible for starch hydrolysis [72] and *AMY1* shows extensive multi-allelic copy number variation in human [73] that positively correlates with the secreted protein levels [12]. Impressively, human populations characterised by high-starch consumption present on average more *AMY1* copies than lower-starch consumption ones; moreover, these data do not present a geographic correlation, that would be consistent with the hypothesis of an ancestral gene duplication followed by migratory flows, but diet was the strongest predictor of copy numbers. For instance, populations not geographically distant as Yakut, a pastoralist, fishing society from North-East Asia (low-starch consumption), and Japanese (agricultural, high-starch consumption) showed a significantly different copy number distribution, with a higher *AMY1* copy number observed for the Japanese population. *AMY1* hence shows strong signs of recent positive selection in high-starch consumption populations, possibly because a higher amylase activity can improve starch hydrolysis and glucose release starting from the mastication process, through stomach and intestines. Also, a higher oral digestion of starch seems to be critically importance for energy absorption in case of intestinal diseases, hence increasing the fitness level of the population [72]. qPCR and fibre-FISH data performed in the Perry *et al.* study [12] revealed a fixed diploid *AMY1* copy number in the chimpanzee lineage. Interestingly, despite their sequence-based analysis detected an increased *AMY1* copy number in bonobo, they realised that all the bonobo copies presented a disrupted coding sequence and may hence all be non functional. On the other hand, New World monkeys, having a null-starch diet, do not express salivary amylase, consistently with the evolutionary adaptation proposed. Nevertheless, an old study of 1982 reported high salivary amylase expression, even higher than human, for the Old World Monkeys cercopithenices subfamily, rhesus macaque and mangabeys. One possible explanation for these increased expression levels could be that cercopithenices are the only primates to have cheek pouches, where they could start the digestion of starch-rich food, as the seeds of unripe fruits [74].

An interesting example of expanded human-specific CNV genes are DUF1220-encoding genes, members of the neuroblastoma breakpoint family (NBPF) [75]; DUF1220 domains are highly expressed in brain regions associated with higher cognitive function [76] and, remarkably, alterations in their copy number are associated with mental retardation [19] and autism spectrum disorders [77]. It is then possible that this mechanism of gene expansion, also involving other genomic regions, could have played a central role in the improvement of human cognitive functions during evolution, compared with other primate species.

Another paradigmatic case of how CNVs could have influenced divergence and selection of the human species is the lineage specific expansion of the aquaporin 7 genes (*AQP7*). Differently from other primates, humans had to adapt to living in open savannahs developing skills in endurance running [78] and the aquaporin gene family is important for two main activities: water transport across membranes, helping in heat dissipation through sweating and glycerol transport and to mobilise energy stored in adipocytes for prolonged periods of time.

1.4.3 CNVs in rhesus macaque

As discussed before, the progressive discovery of the CNVs extent and importance in human raised the interest towards a better understanding of the CNV landscape in non-human primates. In February 2006 the first assembly of the *Macaca mulatta* genome was released. Differently from the first human assembly, which was generated through sequencing and assembling of a human BAC library, the rhesus macaque assembly has been created using a whole-genome shotgun approach. This allowed the starting of a new phase of whole-genome studies on rhesus macaque species, without the constraints arising from the use of a less specific human reference. Also, from an evolutionary point of view, the comparison between intra-specific copy number differences in the rhesus macaque genome and the patterns observed in human can give information on the active evolutionary forces in the two species. The first genome-wide study to evaluate intra-specific copy number extent in the rhesus macaque genome was performed in 2008 by Lee *et al.*, who designed a customised array-CGH, based on the *Macaca mulatta* genomic assembly, with 380000 isothermic oligonucleotide probes to interrogate the genome with a resolution of ~6.5kb. The effective resolution of the study was ~40 kb, as a threshold of 6 consecutive probes was set to call CNV. They analysed 9 unrelated individuals against a rhesus macaque reference, identifying 132 genomic gains and 82 loss, corresponding to 123 distinct CNVs [79]. Functional gene families identified as CNV included: immune related genes (HLA family, β-defensins) and metabolic genes as cytochrome P450 family 2, subfamily A. Nevertheless, the authors report a known underestimation of smaller CNVs (<40kb) that could account for a considerable proportion of genome diversity. In addition, they confirmed that, also in the macaque genome, CNVs often encompass segmental duplications (30% of the CNVs reported). It should be noted that at the time of the study, data coming from the first rhesus macaque genomic assembly [80] were suggesting that segmental duplication content of the macaque genome was less than half that of the human genome; also this value could have been underestimated, because of the presence of gaps in the rhesus assembly and for the difficulties in performing an accurate building of high-complexity genomic areas. The main outcome of this technical problem could have been the loss of detection power for CNVs spanning repeated elements. Despite these caveats, it was still possible to observe a clear pattern of duplication-mediated non-allelic homologous recombination driving CNVs formation at certain genomic hot-spots. Authors reported 101 out of 123 orthologous CNV shared between human [19] and rhesus macaque, present in multiple individuals. Interestingly, given the divergence time of the two species (>25 MYA) it is likely that these orthologous CNVs may have arisen as consequence of independent events. Another significant aspect is that the segmental duplications flanking shared CNVs do not present the same level of nucleotide diversity that would be expected by chance (~7%); they appeared to be more conserved, constituting an ideal substrate for NAHR events (requiring a percentage of identity >95%). One possible explanation for this finding could be a series of gene conversion events, with the long term maintenance of recombination hotspots in the primate genomes.

Three years later, Gokcumen O and colleagues generated a new high-resolution CNV map for the rhesus macaque genome, increasing dramatically the CNV detection power [81]. For this aim, they designed a rhesus macaque-specific array CGH platform containing 950,843 unique 60-mer oligonucleotide probes, on 17 non-related individuals. This approach provided a 15 kb effective resolution, and, setting a threshold to 5 minimum consecutive probes with a log intensity ratio significantly different from 0, it was possible to identify 1160 CNVs. Of those, 95% were not previously identified, being in large part small CNVs not detectable with lower-resolution techniques. Interestingly, 74% of the CNVs discovered overlapped with *Ensembl* gene prediction, with the percentage increasing to 90% for multiallelic CNVs. These data were compared with a human non-redundant map of 12146 CNVs [82] [16] and a chimpanzee datasets with 438 merged CNV regions [48] to identify overlaps among the three species. It was possible to confirm that human CNVs overlap with non-human primate CNVs more than expected by chance alone; 1387 distinct human CNVs overlapped with 556 chimpanzee CNVs, 467 human CNVs overlapped with 385 rhesus macaque CNVs and 170 human CNVs overlapped with both chimpanzee and rhesus macaque CNVs. Interestingly, the analysis of CNVs shared among primates confirmed that primate hotspots of CNV formation overlap with regions of recurrent human CNV formation significantly more than expected. Also, if these CNV hotspots overlapping functional elements had evolved under neutral condition, it would have been possible to observe a depletion of functional loci; on the contrary, as the majority of shared CNVs fall into functional genomic areas in the three species in analysis, it was possible to deduce that these hotspots evolved under balancing or positive selection. Using empirical measures of selection (based on the ratios of non synonymous on synonymous nucleotide substitutions within species), it was possible to establish that most shared CNV regions evolve under species-specific positive selection. Nevertheless, immunity related CNV gene families (as HLA and LILR families) showed signs of balancing selection, preventing the fixation of their copy numbers among the three species.

1.4.4 Role of segmental duplication in primate CNV formation

The case studies hereby presented on copy number detection on primates highlighted the important role of segmental duplications (SD) in CNV formation and evolution. Experimental and computational analysis performed on Eutherian mammals (rat [83], cow [84], dog [85]) as progressively the genomic assemblies of different mammalian species were released, showed that segmental duplications were ancestrally organised in tandem repeats. For instance, despite human and mice are comparable for the amount of sequence mapping to high-identity duplications, in the mouse lineage segmental duplications are almost entirely organised in tandem repeats, whereas in human >59% of the annotated duplication are interspersed [86]. Also, human SD show a high percentage of identity, suggestive of recent formation and/or gene conversion, and studies on non-human primates confirmed a burst of segmental duplications in the primate lineage [49] [53] . Interestingly, this duplications expansion happened in parallel with the slowing down of other mutational processes that drove the evolution of mammals, as point mutations and retrotransposon activity [87] [88].

In 2009, T. Marquet-Bonet and E.E. Eichler [89] identified the ancestral origin of 4692 human duplication loci and analysed the organisation of 437 duplication blocks in the human genome, previously identified in the same research group by Jiang Z *et al.* 2007 [90]. Data suggested the clustering of these loci in 24 discrete groups, with two different origins: 10 blocks presented mainly pericentromeric and subtelomeric duplications, originated from non-homologous chromosomes gene flow; 14 blocks corresponded to intrachromosomal burst of segmental duplications. This hierarchical clustering provided enough evidence to formulate the 'core duplicon hypothesis', proposing that these duplications were formed around a core of ancestral duplicons, rich in genic areas and splice expressed sequence tags. Interestingly, several of the genes falling in core duplicons do not present orthologous outside the primate lineage, with a functional enrichment for genes related to brain development and neural functions. Nevertheless, the high plasticity of these areas, prone to CNVs and structural rearrangements, on one side favours greatly the evolution of the genome with the expansion of different gene families, but on the other increases dramatically the risk of pathologic rearrangements. Indeed, 30 of the duplicated loci reported, known to be copy number polymorphic in human, have been associated with neurocognitive and neurobehavioural disabilities [91] [92] as developmental delays, autism, schizophrenia and Smith-Magenis Syndrome. Strikingly, in a recent paper of 2013 [56], Sudmant PH and colleagues presented the first case of genomic disorder in the chimpanzee lineage, with an individual manifesting a Smith-Magenis-like phenotype. Further analysis highlighted a microdeletion on chromosome 17p11.2 (the same recurrent rearrangement observed in human for Smith-Magenis Syndrome [93]) but characterised by different breakpoints and different duplication blocks compared to human, suggesting a lineage-specific evolution of one of the core duplicon hotspots.

1.4.5 Examples of CNV gene families in rhesus macaque

The study of Hahn *et al.* 2007 evidenced noticeable expansions in multiple gene families in macaque (**Figure 1.20**) [53]. The HLA family showed the largest expansions, with at least 22 lineage-specific gene duplicates. This result is further supported by array CGH data identifying the same large HLA expansion along the macaque lineage [80]. The work of Lee AS *et al.* 2008 confirmed this data, reporting the same gene family expansion [79]. Other immunity-related gene families present expansions in macaque, including immunoglobulin k chain variable regions, T cell receptors, and killer cell immunoglobulins. It is noticeable how these duplications in immune-related genes are accompanied by macaque-specific expansions in several nuclear-encoded proteins of viral origin, as envelope and gag polyproteins. This may suggest a co-evolutionary competition between viral pathogens and the macaque host immune system.

Data derived from the study of Gibbs RA *et al.* 2007 aimed at the identification of differences in gene expansion between human and rhesus macaque, show a strong macaque-specific gene duplication signal for the phosphofructokinase platelet form (*PFKP*) gene, concentrated at the telomere of chromosome 9 but also scattered all across the macaque genome. *PFKP* is important in fructose metabolism and, given the fruit-rich diet of rhesus macaque, multiple duplication rounds of this gene could have been under strong positive selection [80].

1.5 Rhesus macaque importance as biomedical model

Rhesus macaques are the most widespread non-human primates. They belong to the family of Cercopithecidae, which last recent common ancestor with human can be placed ~25 MYA [94]. Their geographic distribution ranges from Afghanistan and India all across Asia to the Chinese shores of the Pacific Ocean; because of its wide distribution and large population size, rhesus macaque is the only non-human primate to be listed as 'Least Concern' in the IUCN (International Union for the Conservation of Nature) Red List or Threatened Species. This species shows a great adaptation to different environments and diets, often living in close proximity with humans, hence facing with them the same environmental challenges.

Historically, rhesus macaque has been the most studied primate model in applied biomedical research, as it exhibits greater similarity to human physiology, neurobiology, and susceptibility to infectious and metabolic diseases, compared to lower organisms such as yeast or fruit flies, or rodent models.

Noticeable scientific progresses made using rhesus macaque as an animal model include the discovery of the 'rhesus factor' blood groups and the development and testing of different vaccines, such as rabies, polio and smallpox. Even more intriguingly, because of their geographic distribution, rhesus monkeys are a naïve species for the Simian Immunodeficiency Virus (SIV), widespread in the African continent. A phylogeographic analysis of the SIV virus revealed that this pathogen may be present in Africa since 32000 years [95]; as a consequence, the resistance of non-human African primates to the virus is the result of a long-term host-pathogen co-evolution, and, in natural hosts, the infection is typically non pathogenic and characterised by a sustained preservation of peripheral CD4+ T cells (trait shared with human HIV-infected *long term non progressors*). Conversely, upon SIV infection rhesus monkeys share an extraordinary similarity in the characteristics of disease progression with humans, rapidly undergoing peripheral CD4+ T cell depletion and showing manifest signs of Simian Acquired Immunodeficiency Syndrome (SAIDS, analogous to human AIDS)[96]. Clearly, this striking analogy makes rhesus macaque the favourite model for studies on the early phases of SIV infection, on the deregulation mechanisms involved and for the development and testing of antiretroviral approaches for HIV/AIDS management and remission.

Also, after the completion of the first draft of the chimpanzee genomic assembly in 2003, many studies where addressed to compare the levels of identity and diversity with the human data. As chimpanzee and human diverged relatively recently (~6 MYA), the nucleotide diversity between the two genomes was found to be in the order of 1-2%, posing problems in distinguishing conserved elements from the overall high background level of conservation. In this respect, the addition of genomic information coming from a more divergent genome, as for the rhesus macaque, that shares ~93% of identity with humans, can help to overcome this problem. Using a 'three-way approach', comparing the sequence information from the three genomes, it becomes possible to understand, for example, whether a gene family showing expansions in humans and not in chimpanzee represents a species-specific expansion or, if present in rhesus macaque, is a case of chimpanzee-specific gene loss. Also, this concept opens the way to deeper analysis on the mutational mechanisms and the selection forces that shaped the evolution of the Primate lineage in the last 25 million years. In **Figure 1.22** is presented an example of this sort of comparative analysis, that brought to the discovery of 67 genes showing evidence of selection in Primates [80].

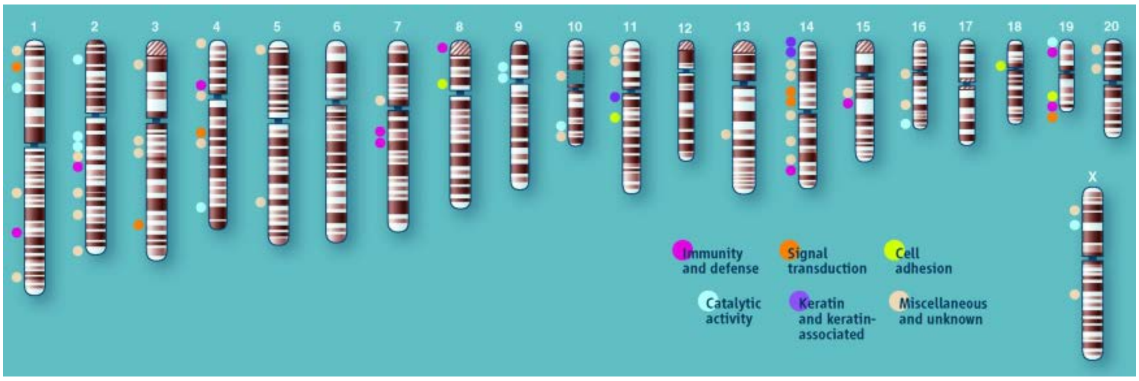


Figure 1.22: Cytogenetic map illustrating the 21 rhesus macaque chromosomes. The dots, colour coded accordingly to Gene Ontology categories, highlight 67 genes showing evidence for positive selection in primates, after comparative analysis of rhesus macaque, chimpanzee and human genomes. Picture taken from [80] interactive online (www.sciencemag.org/ sciext/macaqueposter/)

1.6 Mechanisms of CNVs formation

At present, there are several mechanisms known to generate genomic rearrangements and hence copy number variations.

1.6.1 Non allelic homologous recombination

Non-allelic homologous recombination (NAHR, also known as ectopic recombination), between two low-copy repeats (LCRs) is responsible for most recurrent genomic rearrangements observable in different individuals [97] [98]. In particular low copy repeats are region–specific DNA segments with a size that usually range between 10 and 300 kb, sharing 95% to 97% similarity to each other; LCRs cover

5% to 10% of the human genome, and are thought to have arisen during primate evolution, conferring plasticity to the genome and thus facilitating gene divergence [99] [100]. Because of their high similarity, non allelic LCRs can misalign during DNA replication and the subsequent crossover between them determines genomic rearrangements in the progeny (**Figure 1.23**).

NAHR can occur in both meiosis and mitosis. When it happens in meiosis, it results in unequal crossing over, leading to constitutional genomic rearrangements; these alterations can be benign polymorphisms or manifest sporadic, if *de novo*, or inherited genomic disorders [101]. NAHR in mitosis results in mosaic populations of somatic cells carrying different copy numbers [102].

This mechanism does not occur with the same frequency in each LCR, but is concentrated in certain hotspots [103] [104] [105]; interestingly, this is correlated with the presence of DNA structures more prone to undergo double strand breaks (as palindromes, non-B conformation DNA, G quadruplexes, minisatellites and DNA transposons). For this reason, CNVs generated through NAHR mechanisms tend to be clustered and to show common breakpoints.

1.6.2. Non homologous end joining

Non homologous end joining (NHEJ) is one of the two major pathways for repair of double strand DNA breaks (DSBs) (**Figure 1.23**); the other pathway is homologous recombination, which is restricted to late S or G(2) phase, while NEHJ can occur throughout the cell cycle and is the major repair mechanism in human cells and in multicellular eukaryotes. In particular, in humans it directs physiological V(D)J recombination, generating a diverse repertoire of immunoglobulins and T-cell receptors (TCRs), and is also used to repair pathological DSBs caused by ionizing radiation or reactive oxygen species [106].

NHEJ activity is essential in the organism, and has been linked to human pathology in different ways: inherited defects in this mechanism are responsible for 15% of human severe combined immunodeficiency cases [107], it is also considered to be the major mechanism rejoining translocated chromosomes in cancer and it might represent one potential contributor to biological ageing [108].

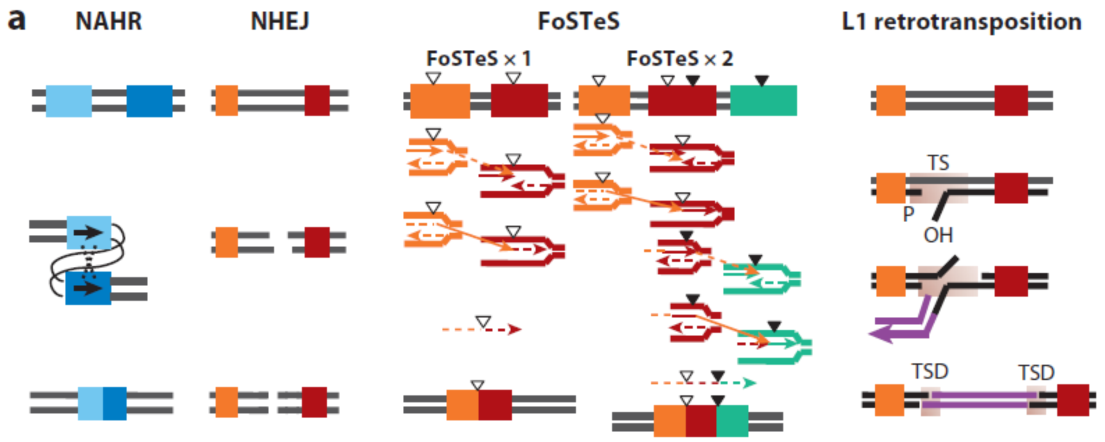
The NEHJ mechanism has two peculiar characteristics: first, unlike NAHR, it does not require substrates with extended homology, like LCRs; and secondly, it leaves an "information scar" at the join point, because the pre-rejoining editing of the ends includes cleavage or addition of several nucleotides from or to the ends [109].

1.6.3 Fork stalling and template switching

Recently, the advent of the new array CGH technology has allowed a deeper analysis of DNA rearrangements and common breakpoints, unravelling new complex details which could not be explained completely with the previously described models of recombination. In 2007 Lee JA and colleagues [110] proposed a further model to explain complex human genomic rearrangements, which was called "Fork stalling and template switching" (FoSTeS). This event can happen during DNA replication, where the DNA replication fork can stall at one position during the new DNA strand synthesis. Because of this, the lagging strand can be disengaged from the original template, switching and annealing to another replication fork; this event is mediated by 3'end micro-homology between the switched template and the new replication fork. After the annealing step, the transferred strand is able to prime its own template-driven extension at the new replication fork (**Figure 1.23**). Interestingly, this mechanism does not necessarily require the new template strand to be adjacent to the original replication fork in primary sequence, it is just enough that both of them are in three-dimensional physical proximity.

1.8.4 L1 retrotransposition

Retrotransposons, mainly LINEs, SINEs, and endogenous retroviruses, make up roughly 40% of the mammalian genome and have played an important role in genome evolution. In particular L1 elements are the only currently active autonomous transposons in the human genome [111]. Also L1 elements can mediate CNVs formation: L1 transposition [112] occurs via an RNA intermediate, probably transcribed by RNA polymerase II. The reverse transcription and integration are thought to occur in a coupled process named target primed reverse transcription, where the resultant insertion is flanked by duplicated target sites (**Figure 1.23**) [6].



b	NAHR	NHEJ	FoSTeS	Retrotransposition
Structural variation type	dup, del, inv	dup, del	dup, del, inv, complex	ins
Homology flanking breakpoint (before rearrangement)?	Yes (LCR/SD, *Alu*, L1, or pseudogene)	No	No	No
Breakpoint	Inside homology	Addition or deletion of basepairs, or microhomology	Microhomology	No specification
Sequence undergoing SV	Any	Any	Any	Transcribed sequences

Figure 1.23: Comparisons and characteristics of the four major mechanisms underlying human genomic rearrangements and CNV formation. (*a*) Models for Non-Allelic Homologous Recombination (NAHR) between repeat sequences (LCRs/SDs, *Alu*, or L1 elements); Non-Homologous End-Joining (NHEJ), recombination repair of double strand break; Fork Stalling and Template Switching (FoSTeS), multiple FoSTeS events (×2 or more) resulting in complex rearrangement and single FoSTeS event (×1) causing simple rearrangement; and retrotransposition. TS, target site; TSD, duplicated target site. Thick bars of different colours indicate different genomic fragments; completely different colours (as *orange* and *red* or *orange*/*red*/*green* in FoSTeS×2) symbolise that no homology between the two fragments is required. The two bars in two similar shades of blue indicate that the two fragments involved in NAHR should have extensive homology with each other. The triangles symbolise short sequences sharing micro-homologies. Each group of triangles (either *filled* or *empty*) indicates one group of sequences sharing the same micro-homology with each other. (*b*) Characteristic features for each rearrangement mechanism. Specific features of certain mechanisms are shown in red. Abbreviations: dup, duplication; del, deletion; inv, inversion; ins, insertion. [113].

1.7 Impact of CNVs on gene expression and phenotype

The discovery of the vast extent of genic CNVs in different species brought the attention on the multiple ways in which CNVs can impact gene expression and phenotypic effects.

For instance, the extensive and recurrent copy number changes characterising the olfactory receptor family have been a key driver in the adaptation and evolution of extremely diverse organisms, from insects to primates. The relaxation of selective pressure on olfaction, accompanied by the development of trichromatic vision in the catarrhini lineage (see chapter 1.4.2), brought to the accumulation of OR pseudogenes in the human lineage, shifting from an initial expansion under positive selection to genomic drift and stochastic changes in copy number [114]; hence, in human, the most duplicated gene superfamily evolves in neutral conditions, still affecting the sensitivity in the perception of different smells among individuals [114] [115].

When not silent, CNVs can have different effects, positive or negative. Frequently, copy number changes can positively correlate with gene expression, as in case of the salivary amylase gene *AMY1* [12] or human β-defensin 2 [29], with direct effects on the corresponding protein levels. In some other cases, it is possible to observe a more complex correlation between copy number and gene expression; one example of this phenomenon is given by the onco-suppressor gene *PTEN* (Phosphatase and tensin homolog): duplications of *PTEN* give rise to an expressed pseudogene that acts as a sponge for microRNAs. Deletions of the pseudogene bring to increased microRNAs levels that negatively regulate the expression of the parental gene [116].

Also, expression levels of some genes can be altered as a consequence of variation in copy number of another gene sharing the same promoter region; for example, reduction of the α-globin copy number dramatically increase the expression of the *NME4* gene (Non-Metastatic Cell 4), located 300 kb away from the α-globin cluster. Given the high-frequency of α-thalassemia in the human population, it is hence expectable to see related fluctuations in NME4 expression levels, unexpectedly due to α-globin CNVs [117].

An increasing number of diseases and human complex traits are being linked with copy number variations, with a consequent improvement of our knowledge about their molecular mechanisms of pathology. This correlation is evident in reports that more than 300 disease-causing genes overlapped with CNVs [8] and that 14.5% of CNVs overlapped with the OMIM Morbidity Map, that presents the cytogenetic map location of disease genes described in OMIM [19].

Rearrangements determining copy number variations of certain genomic regions can have phenotypic effects with different mechanisms, presented in **Table 1.2** and **Figure 1.24**.

Interestingly, recent studies have revealed an increasingly important role of CNVs in complex trait diseases, like neurodegenerative disorders [118], autism spectrum disorders [119], HIV susceptibility [120] and psoriasis [121]; [122]. Nevertheless, the attribution of risk factors for complex trait diseases is a difficult process, often leading different studies to call discordant genomic regions of association. For example, a different hypothesis on risk factors for bipolar disorder and schizophrenia does not contemplate a role for CNVs, bringing the attention on mutations affecting voltage-dependent calcium channels (on the genes *CACNA1C* and *ANK3*) and hence considering these disorders to be ion channelopathies [123] [124] .

Mechanism	**Example**
Gene dosage	Variation in copy number of dosage-sensitive genes. Duplications can lead to gene overexpression, whereas deletions can cause haploinsufficiency. The *PMP22* gene (peripheral myelin protein 22) provides an example of this mechanism; duplications of *PMP22* cause Charcot–Marie–Tooth disease (CMT, OMIM:118200), whereas deletions are associated with the Hereditary Neuropathy with liability to Pressure Palsies (HNPP, OMIM: 162500) [101].
Gene interruption	Deletion or duplications can affect a functional gene. it is often possible to observe a loss of function. Events of non-allelic homologous recombination at the opsins cluster can generate defective opsin products, with the deletion of the red or the green opsin cluster, causing colour-blindness [125].
Gene fusion	Fusion of two different genes, or of a gene with its regulatory elements. It is one of the possible effects of somatic chromosomal translocations; gain of function mutations of this kind can generate cancers. One example of disease associated with gene fusion is the glucocorticoid-remediable aldosteronism (GRA, OMIM: 103900), where two genes sharing the 95% of similarity fuse together as a consequence of NAHR[126].
Position effect	A CNV can remove or alter a regulatory sequence, a nearby gene can be affected. Mutations in *SOX9* (sex determining region Y-box 9) are known to cause campomelic dysplasia (OMIM: 114290), but also two balanced translocations mapping 900 kb upstream and 1.3 Mb downstream of *SOX9* can generate the same disease, in the absence of *SOX9* mutations [127].
Unmasking of recessive alleles or functional polymorphism	Deletion removing one allele may unmask another recessive allele or functional polymorphism. Patients with Sotos syndrome (OMIM: 117450) can present variable activities of the plasma coagulation factor 12 (FXII), due to the type of functional polymorphism present on the remaining copy of FXII [128]. TAR syndrome (Thrombocytopenia with absent radii) is caused by the compound inheritance of a rare null allele at chr1q21.1 and one of two low-frequency non coding SNPs in the regulatory sequence of the gene *RBM8A* [129]
Transvection	Epigenetic mechanism where one allele interacts with the allele present on the homologous chromosome, influencing its regulation. The presence of CNVs can alter this phenomenon, as in case of Smith–Magenis Syndrome (SMS, OMIM: 182290), where the penetrance of craniofacial anomalies typical of this disorder are related to CNVs of the genomic area surrounding the deletion of *Rai1* gene[130].

Table 1.2: Examples of phenotypic effects due to different copy number formation mechanisms. Adapted from [113].

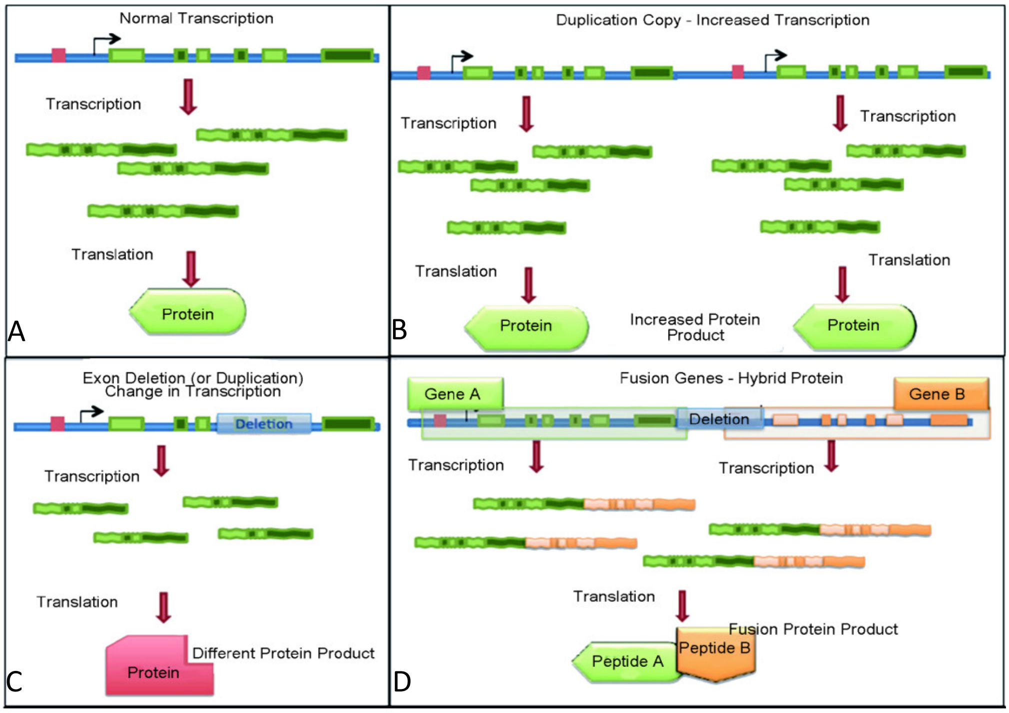


Figure 1.24: Impact of CNVs on gene regulation. There are multiple ways in which CNVs can impact transcription by overlapping coding regions of the genes [81]. Panel A shows a normal pattern of transcription. Panel B shows the effect of CNV duplications positively correlating with gene expression: it is possible to observe an increased level of protein produced, proportional with the number of gene copies. Panel C presents possible effects deriving from exonic deletions or duplications, with the synthesis of a different protein product. Panel D presents a case of gene fusion, with the creation of a fused protein product. Adapted from Gokcumen O *et al.* 2011 [81].

1.8 Defensins

Among the genes reported to be copy number variable in humans and other mammalian clades, this work focuses on beta defensin genes for several different reasons presented in this chapter. Defensins are multifunctional secreted small peptides that play a crucial role in the innate immune system. They possess a wide antimicrobial spectrum against both Gram positive and Gram negative bacteria, fungi and enveloped viruses [131-133]. Interestingly defensin-like genes have also been discovered in plants [134]. Their size varies between 2 and 6 kDa and are formed by 18 to 45 amino acids and have a high degree of degeneracy in amino acid sequence. Defensins are rich in cysteine and, in particular, vertebrate defensins show a characteristic pattern of 6 conserved cysteine residues. On the base of the spacing pattern and pairing of the six cysteine residues they can be further classified into three families in primates, α- β- and θ-defensins [135] (**Figure 1.25** and **Figure 1.26**).

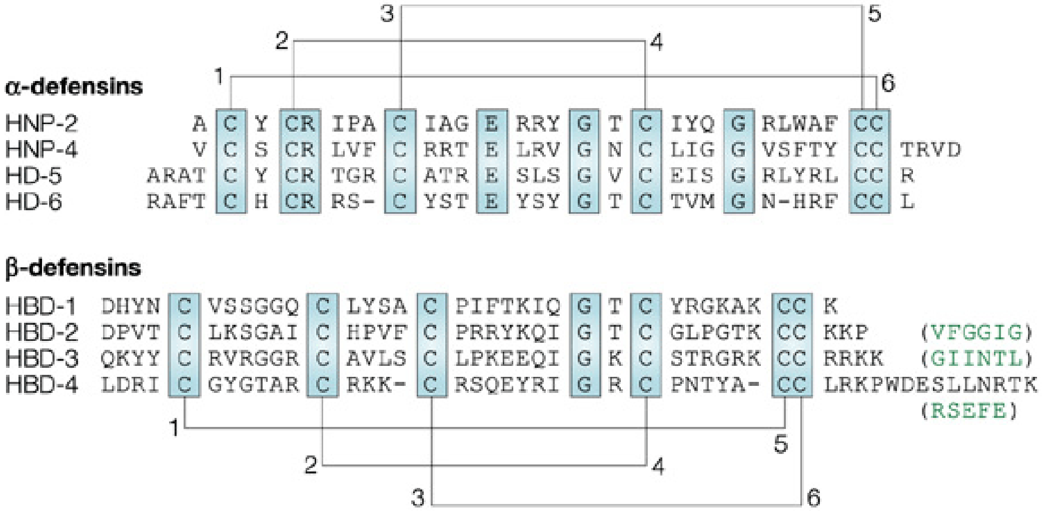

α-defensins
1
2
3
4
5
6
HNP-2 A C Y CR IPA C IAG E RRY G T C IYQ G RLWAF CC
HNP-4 V C S CR LVF C RRT E LRV G N C LIG G VSFTY CC TRVD
HD-5 ARAT C Y CR TGR C ATR E SLS G V C EIS G RLYRL CC R
HD-6 RAFT C H CR RS- C YST E YSY G T C TVM G N-HRF CC L
β-defensins
HBD-1 DHYN C VSSGGQ C LYSA C PIFTKIQ G T C YRGKAK CC K
HBD-2 DPVT C LKSGAI C HPVF C PRRYKQI G T C GLPGTK CC KKP (VFGGIG)
HBD-3 QKYY C RVRGGR C AVLS C LPKEEQI G K C STRGRK CC RRKK (GIINTL)
HBD-4 LDRI C GYGTAR C RKK- C RSQEYRI G R C PNTYA- CC LRKPWDESLLNRTK (RSEFE)
1
5
2
4
3
6

Figure 1.25: HNP-2 and HNP-4 are expressed in leukocytes, HD-5 and -6 are expressed mainly in Paneth cells. Note that sequence conservation is limited to the 6 cysteines (disulphide bonds form between cysteines 1 and 6, 2 and 4 and 3 and 5), an arginine residue and a glutamic acid residue that are known to form a salt bridge in HNP-3, and two glycine residues. In the four human β-defensins, only the cysteine residues and one glycine are invariant. In β-defensins, there are disulphide bonds between cysteines 1 and 5, 2 and 4 and 3 and 6. The residues shown in green are amino terminal to those aligned in the diagram and may or may not appear in the fully processed peptide[135].

1.8.1 Evolution of the defensin genes

Although the evolutionary relationship between vertebrate and non-vertebrate defensins is still unclear, phylogenetic analysis suggests that a primordial β-defensin is the common ancestor for all vertebrate defensins and this gene family expanded throughout vertebrate evolution [136]. This hypothesis is supported by the discovery of β-defensin genes in phylogenetically distant vertebrates, as reptiles, birds [137] and teleost fishes; in particular for the latter, they present 5 aa (DB-2 and 3) or 6 aa (DB-1) residues between the first two cysteines, indicating they are equivalent to mammalian β-defensins rather than α-defensins; moreover, zebrafish-defensins maintain the peptide fold of the mammalian counterparts and the disulphide bond pattern 1–5, 2–4,3–6[138]. α-defensins are mammalian specific genes, and in human α-defensin genes and different β-defensin genes are present on adjacent loci on chromosome 8p22-p23.

The organisation of this cluster is consistent with a model of successive duplication rounds and consequent divergence under positive selection from a common ancestral gene that produced a diversified cluster of paralogous genes [139 140].This expansion has been temporally placed before the human-baboon divergence (23-63 MYA)[141 142]. In particular, β-defensins probably evolved before the divergence of mammals from birds generating α-defensins in glires and primates after their divergence from other mammalian species [143].

θ–defensins further diverged from α-defensins after separation of primates from other mammals and are now actually expressed just in Old World monkeys, gibbons and orangutans. In human, chimpanzee and gorillas θ-defensins turned into pseudogenes, as their *DEFT* gene acquired a mutation that introduced a stop codon into the signal sequence. This divergence also reflects differences in defensin localization: α-defensins are mostly produced in the granules of either leukocytes or intestinal Paneth cells; β-defensins are primarily secreted from nongranular mucosal epithelial cells lining the respiratory, gastrointestinal, and genitourinary tracts [144]; θ–defensins are produced in monocytes and neutrophils granules and are lectins with broad-spectrum antiviral efficacy.

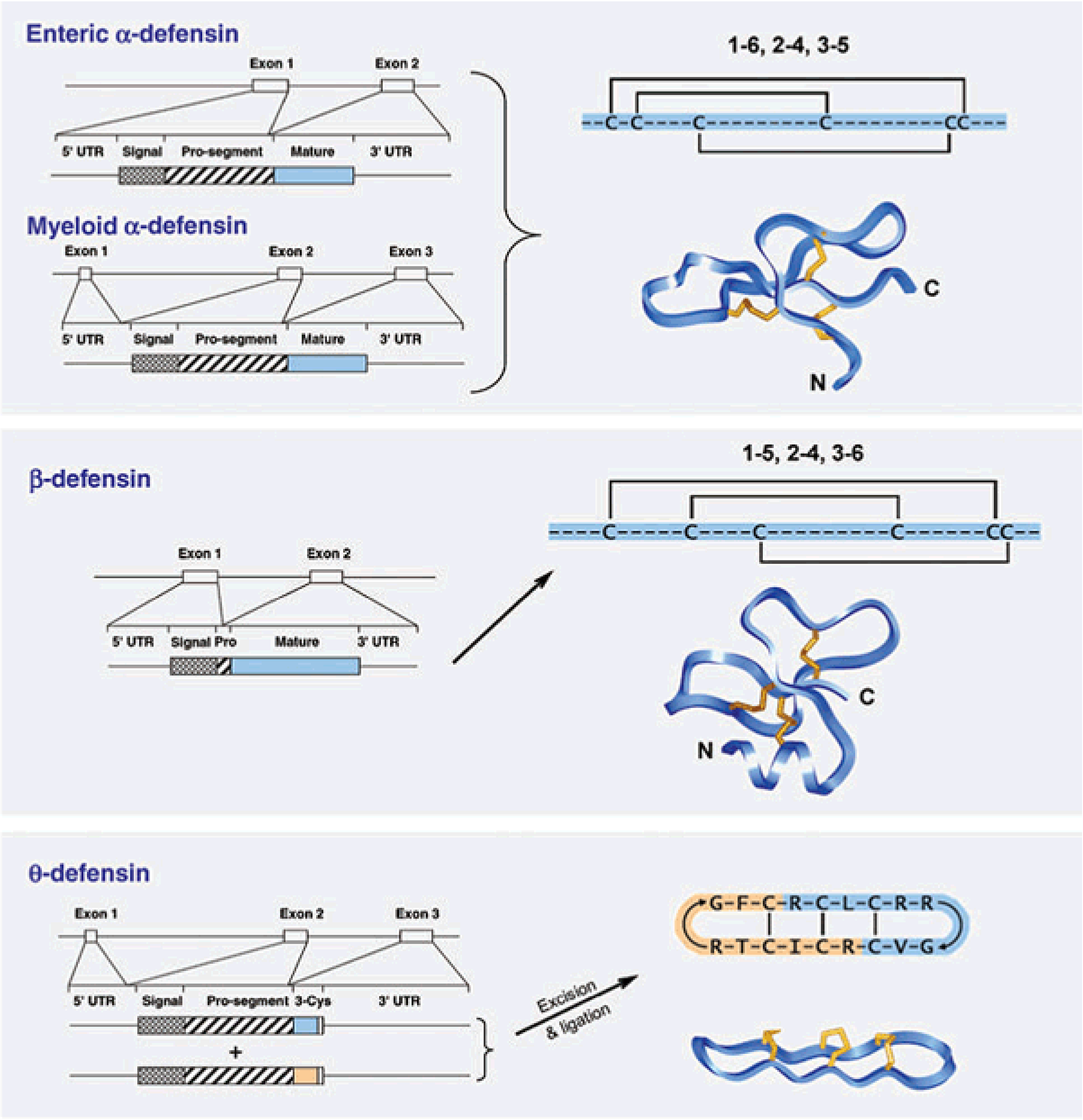


Figure 1.26: Defensin genes and peptides. Left, alignment of α-defensin, β-defensin and θ-defensin genes. Crosshatching, signal peptides (Signal) and pro-pieces (Pro-segment; Pro); blue, residues present in the mature defensin. Right, three different disulfide 'schemes'. Numbers of the above diagrams indicate the disulfide connections in each. The three-dimensional structures are of rabbit α-defensin RK-1 (top), human β-defensin-1 (middle) and θ-defensin RTD-1 (bottom)[145].

1.8.2 β-defensins

In human, β-defensins genes are organised into three main clusters: 8p23.1, 20p13 and 20q11.1, with another probable small cluster on chromosome 6p12 [131].

β-defensins present two exons, coding for a pre-pro peptide. In particular exon 1 codes for a signal sequence and exon 2 codes for a short pro piece and a mature carboxy-terminal peptide liberated by proteolytic cleavage [146]. Interestingly, evolutionary analysis with pairwise comparison of d_N and d_S showed how the DNA encoding for the mature peptide is under rapid divergence, with a significant excess of non synonymous substitutions over synonymous substitutions, suggestive of adaptive evolution in the second exon [140]. Conversely, exon 1, encoding for the signal peptides, and the nearby non-coding DNA show relative stasis [141]. Therefore, duplication of β-defensin genes brings to a relaxation of purifying selection and acts complementary to positive selection on the mature peptide, leading to functional divergence. In this way defensins can tolerate high rate of sequence change without losing completely their antimicrobial activity, and that some changes can be advantageous to confer pathogen specificity. In particular, the sites under positive selection tend to determine changes in the net charge and in the overall hydropathicity of the protein, conferring different properties and specificity in antimicrobial activity [147].

1.8.3 Genome organisation and structural variation of the β-defensin cluster in human

In human the β-defensin cluster maps to chromosome 8p23.1 and include *DEFB4*, *DEFB103*, *DEFB104*, *DEFB105*, *DEFB106*, *DEFB107* and *SPAG11*, lying on the same copy number variable unit, with an approximate size of 250kb, although its precise dimension and breakpoints remain to be identified. *SPAG11* is a gene subject to alternative splicing, formed by a head-to-tail fusion of two ancestral β-defensins; it codes for a protein with antimicrobial function expressed on spermatozoa [148] [149]. Other two β-defensins lie in the same genomic area, outside the copy number variable unit: *DEFB1* and *DEFB109. DEFB1* is not copy number variable and shows constitutive expression in response to pathogenic stimuli. *DEFB109* is annotated as a pseudogenised β-defensin, despite reports of *DEFB109* ubiquitous mRNA expression, with relatively high expression levels in heart, brain, lung epithelium [150], liver, kidney, pancreas, testis and ovary [151]. Also, constitutive expression of hBD9 protein has been detected at the ocular surface of healthy controls and patients with infectious keratitis [152], raising questions on the actual activity status of *DEFB109* gene.

The large genomic repeat unit containing the β-defensin genes hereby listed is flanked by more complex CNV areas containing retroviral elements and olfactory receptors, known as 'REPD' (for repeat distal) [153] and a smaller 'REPP' (standing for repeat proximal), which shares a high level of homology with REPD and is located 5Mb proximal on chr8p23.1. The presence of these repeated areas predispose to large scale imbalances of chromosome 8p, through NAHR, determining recurrent genomic disorders with mild to severe phenotypes. For example, clinical signs of the 8p23.1 deletion syndrome include developmental delays, behavioural problems and congenital heart disease [154]; its reciprocal syndrome, with the duplication of the entire area, is associated with mild dysmorphism and developmental and speech delay [155 156].

The total diploid copy number of the β-defensin cluster ranges from 1 to 12, more commonly between 2 and 7 copies [29] in the general population (**Figure 1.27**). The same authors reported a significant correlation between the levels of *DEFB4* transcripts and β-defensin cluster copy number. The null allele has been observed with an extremely low frequency, ~0.2% and it is possible that, because of the importance of β-defensin genes in immune response, the null homozygote would be extremely deleterious, bringing to a fast removal of null alleles from the population, under strong purifying selection. On the other tail of the β-defensin copy number distribution, it is possible that an extremely high copy number would favour an inappropriate immune response. Hence the different paralogues observed in the β-defensin cluster may be subjected to two different selective pressures: from one side the tendency towards sequence divergence and new function gains, on the other the tendency towards gene dosage equilibrium, through balancing selection and gene conversion between paralogues.

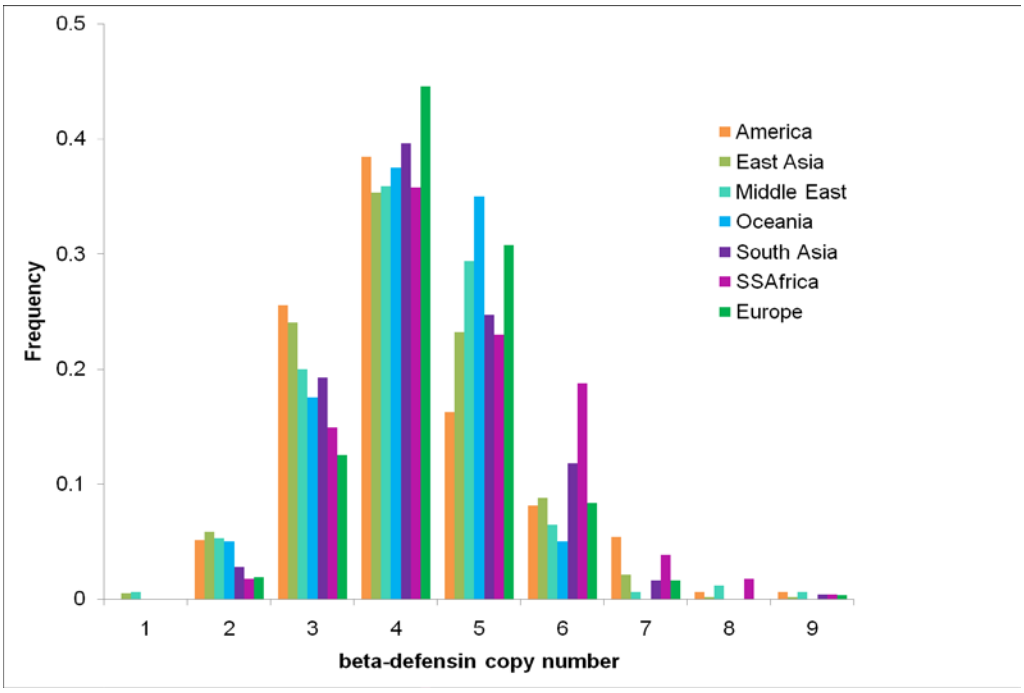


Figure 1.27: Frequency distribution of worldwide β-defensin copy number, detected on 1056 individuals from 67 different populations [157].

1.8.4 β-defensins copy number variation in primates

The natural consecutive step following the identification of a copy number variable cluster for β-defensin, has been the characterisation of the orthologous region in non-human primates, in order to get new insight on the evolution of the cluster.

Hollox EJ *et al.* in 2003 reported a copy number of 4 for the β-defensin cluster in one chimpanzee sample analysed, compared to a diploid copy number of 2 observed in one gorilla individual, using a MAPH approach [29]. A further study conducted on 17 non-related chimpanzee individuals using a previously validated PRT test [35] revealed copy numbers ranging between 4 and 6, confirming CNV in the chimpanzee β-defensin region; nevertheless, the extent of the copy number variable block and its breakpoint are yet to be identified [157].

The β-defensin area has shown signals of variability also through cross-species genome-wide array-CGH studies, using human as reference, against different primate species (**Figure 1.28**) [48 79 81 158]. Nevertheless, the low probe coverage in genomic regions containing segmental duplications, the poor quality of the genomic assemblies for non-human primates in paralogous regions and the increasing level of sequence divergence for primates with a longer divergence time from human, posed notable problems for an accurate comparative analysis of the genomic organisation of the β-defensin cluster in different primates.

The first study reporting β-defensins to be copy number variable in rhesus macaque monkeys has been performed by Lee AS *et al.*, in 2008 [79]. As described in paragraph 1.5.1, the authors used a dedicated genome-wide array-CGH platform for the identification of CNVs areas in the rhesus macaque genome. Among their findings, a strong signal was detected in the β-defensins cluster in 6 individuals out of 9 screened, with 4 gains and 2 losses compared with the reference sample. Interestingly, they noticed that the endpoint coordinates for the β-defensins locus were at least 5 times bigger in the human reference assembly compared to the macaque reference genome (Jan 2006 MGSG Merged 1.0/rheMac2 assembly, signal detected in the coordinates: chr8:8043509-8707577), and they hypothesised that this discrepancy could be due to the presence of large segmental duplication and to a big gap in the human and in the rhesus macaque reference sequences. Despite this, it was still possible to align the nucleotide sequence of the macaque CNV area to the human β-defensin locus, encompassing *DEFB4*.

Gokcumen *et al.* in 2011 performed a second-generation high-resolution CNV map for rhesus macaque, and, among the over 2000 CNVs overlapping with orthologous chimpanzee or macaque CNVs, also the β-defensin cluster on chromosome 8p23.1 has been reported, with large boundaries: chr8:7923933 -8906039 (~982 kb). Clearly, given the complexity of the area in analysis and the presence of olfactory receptors flanking the β-defensin cluster, the signal detected is a probable overestimation of the actual boundaries of the CNV cluster.

These authors recognised the importance of further studies in this direction and, given the bactericidal roles of β-defensins [159] and the association of *DEFB4* copy number with autoimmune diseases [121], rhesus macaque could constitute an optimal model to investigate the correlation between β-defensins gene dosage and microbial load or to optimise targeted antimicrobial drugs accordingly to the genomic β-defensins copy number.

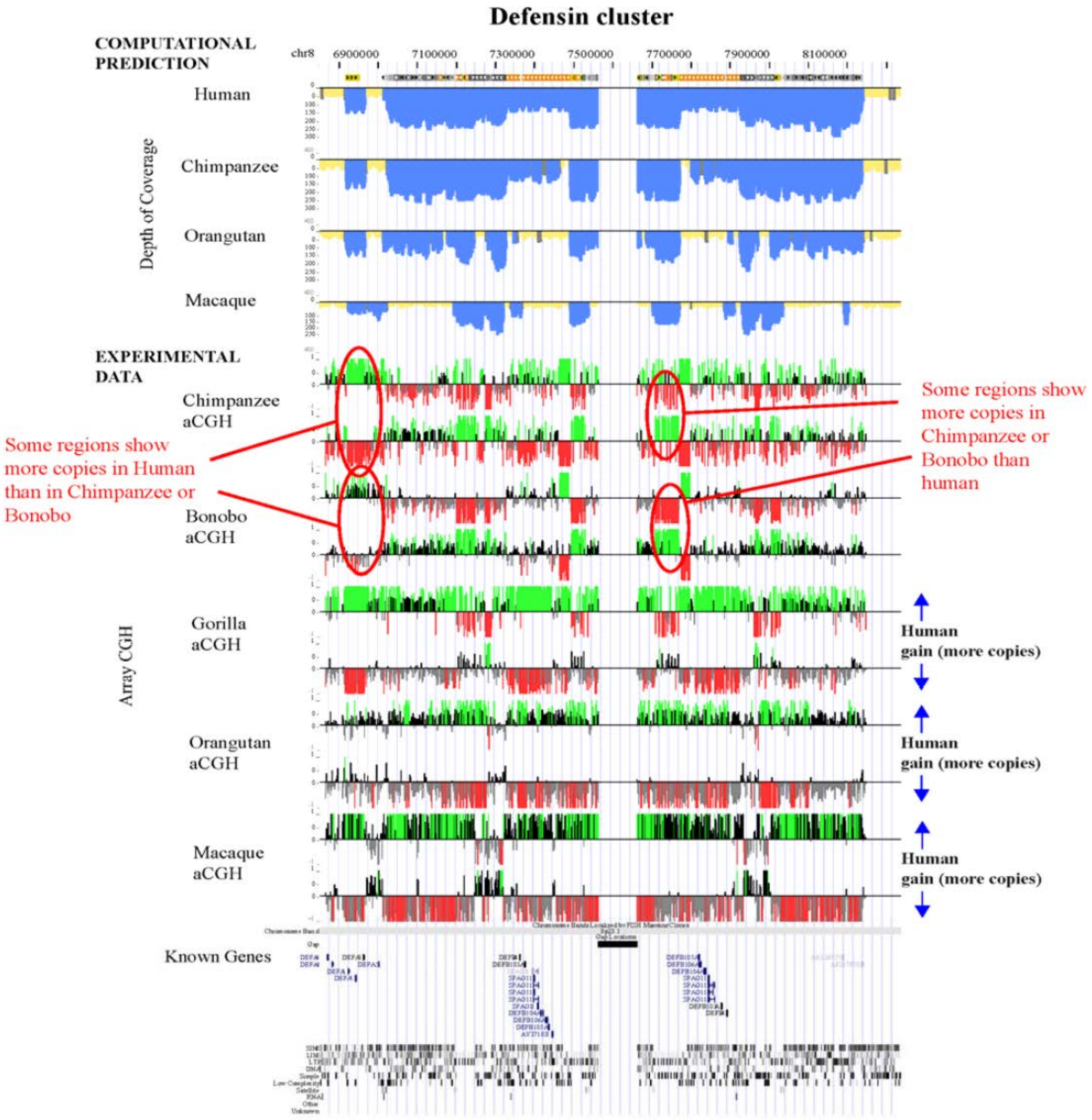


Figure 1.28: Comparative data for the β-defensin cluster, generated through array-CGH and sequence read depth approaches. Computational predictions show a decrease in coverage depth proportional to the increasing level of sequence diversity from human to rhesus macaque. The Experimental data show a complex situation, with copy number calls extremely difficult to disentangle between species [158].

1.7.4 β-defensins structure

The low sequence similarity among the members of the β-defensin family suggests that their antimicrobial activity is largely independent of their primary structure. The only nuclear magnetic resonance (NMR) spectroscopy data available focussed on the three-dimensional structures of human HBD1, HBD2 and HBD3 (PDB accession number: 1E4S, 1E4Q, 1KJ6 respectively), for which it was possible to isolate the native peptides [160 161]. These data confirm a remarkable similarity in their tertiary structure, despite the high degree of divergence in the amino acidic sequence. The mature peptides consist of three β-strands arranged in an antiparallel sheet that represents the major element of their secondary structure. The strands are held together by the three intramolecular disulfide bonds, formed between the 6 cysteins, in the following order: C^{I}-C^{V}, C^{II}-C^{IV} and C^{III}-C^{VI}. The amino-terminal region contains a short α-helix loop (absent in α-defensins), whose orientation is stabilised to the β-sheet by the disulfide bond C^{I}-C^{V}. As α-helical structures are quite common for protein regions within membranes, it has been proposed that this part of the β-defensin protein may be involved in anchoring to bacteria cell walls [162]. This hypothesis is strengthened by the presence of two sites under positive selection located in the N-terminal part that may contribute to β-defensin functional diversity [163].

All the defensins investigated do not present a distinct hydrophobic core and a common pattern of charged or hydrophobic residues on the protein surface; these characteristics may imply that the peptide folding is driven and stabilised just by the disulfide bonds formation. Moreover, the characteristic β-defensin three-dimensional structure can be preserved also accommodating amino acids with different properties at most positions.

Interestingly, the first 5 amino acids of the mature peptide sequence seem to be crucial for a correct protein folding in oxidative conditions, favouring the formation of the correct disulfide bonded pattern through the creation of a key intermediate [147]

1.8.5 β-defensins roles

The most studied function for β-defensins is their direct antimicrobial activity, through permeabilisation of the pathogen membrane. *In vitro* they have been shown to be active against both Gram negative and positive bacteria, unicellular parasites, yeast and viruses. More specifically, cationic peptides like β-defensins tend to be attracted to the net negative charges generated on the outer envelope of Gram negative bacteria by phospholipids and phosphate groups on lipopolysaccharides and to the teichoic acid present on the surface of Gram positive bacteria.

Their exact mechanism of action is yet to be completely understood and two different models have been proposed. The former is a carpet model, where several antimicrobial peptides opsonise the pathogen surface bringing about necrosis, possibly disrupting the electrostatic charge across the membrane [164]. The latter is a pore model, with several peptides oligomerising and forming pore-like membrane defects that allow efflux of essential ions and nutrients [165].

Different β-defensins are present in different epithelial and mucosal tissues and can be constitutively expressed or induced in response to various stimuli. Interestingly their body distribution clearly follows their affinity for different pathogens and they will be more abundant in body districts more prone to the microbial infections they are more specific for. For instance HBD2 is highly expressed in lung [166]; HBD4 is expressed most highly in testes and stomach, and HBD3 in skin and tonsil [165 167]. The respiratory tract sees the presence of HBD1-HBD4, with HBD1 constitutively expressed [168] and HBD2-HBD4 inducible in response to inflammation or infections [169] (**Table 1.3**).

In keratinocytes it is possible to observe a constitutive mRNA expression of HBD1; conversely hBD2 expression is induced by lipopolysaccharides (LPS) or other bacterial epitopes in combination with interleukin-1β, released by resident monocyte-derived cells. HBD3 and HBD4 can be inducible by stimulation with tumour necrosis factor (TNF), Toll-like receptor ligands, interferon (IFN)-γ or phorbolmyristate acetates [145]. hBD3 can be also induced in response to local release of surface-bound EGFR (epidermal growth factor receptor) ligands via activation of metalloproteinases [170 171].

In case of chronic inflammation states, like in psoriasis, HBD2 and HBD3 are strongly induced, with a protective effect from bacterial infections of the psoriatic lesions [172]. Conversely, in patients with atopic dermatitis, a substantial downregulation of HBD2 and HBD3 has been observed, due to an upregulation of T helper type 2 cytokines (IL-13 and IL-4), with the consequential effect of an increased susceptibility to skin infections [173].

In physiological conditions, the presence of additional cell types seems necessary to elicit a proper microbe-induced epithelial production of β-defensins; in particular monocyte/macrophage-like cells assist the production of HBD2, whereas lymphocytes induce expression of hBD1-hBD3. As these different lineages respond differently to bacterial antigens, it is possible to hypothesise that some microbe-derived molecules may induce preferentially a lymphocyte-mediated response, with other eliciting a monocyte/macrophage/dendritic cell-mediated response.

Moreover, β-defensins can recruit immature dendritic and memory T cells to the site of infection and/or inflammation thus linking innate and adaptive immunity [174]. The first indication of this aspect has been provided by D. Yang *et al.*, in 1999. They demonstrated that natural, synthetic and recombinant HBD2 are able to attract chemotactically *in vitro* human immature dendritic cells (DC, CD34+) and memory T cells (CD4+/CD45RO+) in a dose-dependent manner. Interestingly, previous evidences highlighted the role of pertussis toxin (PTX) as a DC-migration inhibitor. As the PTX mechanism of action involves the inhibition of G-protein coupled chemokine receptors, via ADP-ribosylation of $G_i$proteins, it was hypothesised that HBD2 would interact directly with one or more chemokine receptors. Consistently with the hypothesis, PTX was demonstrated to inhibit the β-defensins driven DC migration; anti-CCR6 was the only antibody able to produce a similar inhibitory result, in presence of hBD2, confirming the CCR6 specific binding of HBD2. Moreover, the natural ligand for CCR6, LARC, and hBD2 present a competitive binding kinetics, hence giving further support on HBD2 mechanism of action [175] (**Figure 1.29**).

Interestingly, also T helper 17 cells (Th17) express CCR6 receptor and are hence subjected to β-defensins chemoattractant action. Th17 are a lineage of CD4+ T helper cells producing interleukin-17 (IL-17), a pro-inflammatory cytokine that stimulates the production of further inflammatory mediators, ultimately leading to the recruitment of neutrophils and other leukocytes to the sites of infection [176]. Increased Th17 levels have been reported in different autoimmune diseases, such as multiple sclerosis [177], rheumatoid arthritis [178] and psoriasis [179], suggestive of a further link between β-defensin roles and autoimmunity. Moreover, upon SIV infection, natural hosts show preservation of Th17 levels in the chronic phase, resulting in non-progressive infection, in sharp contrast with Th17 depletion observed in non natural hosts, as rhesus macaque monkeys [180], and human, upon HIV infection [181].

Apart from signalling through CCR6 chemokine receptor, HBD2 has been shown to be a natural ligand for the Toll-like-receptor-4 (TLR-4), present on immature DCs, up-regulating co-stimulatory molecules and leading to DC maturation, and on CD4+ T helper cells, possibly stimulating their proliferation and survival [182].

The role of EGFR in HBD3 induction is another particularly appealing aspect to consider in the spectrum of β-defensins activities, as it raises the question of their role in mechanisms of growth and repair, as pregnancy or wound healing.

A remarkable study performed on human extra-placental membranes stimulated with *E. coli* demonstrated the critical role played by β-defensins in controlling/neutralising intrauterine infections [102]. In particular, HBD1 seems to be constitutively expressed and has been previously reported in the amnion, decidua and choriontrophoblast layer; it has been characterised as the major source of endogenous antimicrobial molecules during human pregnancy [183]. Following *E. coli* challenging, it is possible to observe a marked increase of HBD2 and HBD3 expression on the choriodecidua region of the extra-placental membrane; this has a neat pathophysiological meaning, being the first area to be infected in case of an ascendant infection. This is representing another case in which a differential expression pattern of antimicrobial peptides can help the immune system to obtain the most rapid and specific response following infections.

Regarding wound healing, recent studies highlighted the dose-dependent chemotactic role of HBD2 on human endothelial cells, with an action similar to the vascular endothelial growth factor (VEGF)[184]. HBD2 seems to double the speed of wound closure, in absence of any growth factor; moreover, it promotes endothelial cell proliferation, yet at a lesser extent when compared to VEGF. HBD2 is also capable of promoting capillary-like tube formation of human endothelial cells, and this pro-angiogenic activity is not compromised by the absence of VEGF.

Another remarkable role of β-defensins regards their possible role in tumour progression. As response of the inflammation process triggered by the tumour presence, β-defensins recruit dendritic cell precursors, signalling through CCR6, into the tumoral site. Here the presence of VEGF-A induces the differentiation of DC precursors into endothelial-like cells that in turn promote vasculogenesis and the tumour progression [185]. This discovery introduces a new variable in the equilibrium between DCs, that can break tolerance to tumoral antigens helping the tumoral regression [186], and VEFG-A, able to inhibit the differentiation and maturation of DC precursors [187], mediating tumour immune evasion [188], and to promote the tumoral angiogenesis [189].

β-defensins possess also antiviral activity, interacting directly with the virus and indirectly with its target cells. Noticeably, in mammals β–defensins are also produced by the oral mucosa and they are active against HIV-1 virus: in particular hBD1 is constitutively expressed whereas the presence of a low HIV-1 viral load can stimulate the expression of HBD2 and HBD3 β–defensins gene products through direct interaction with the virus. More specifically, HBD2 has been shown to down-regulate the HIV transcription of early reverse-transcribed DNA products [190] and HBD2 and hBD3 can mediate CXCR4 down-regulation (but not CCR5) and internalisation in immuno-stimulated peripheral blood mononuclear cells [191]. This mechanism diminishes the chances of infection [192]. This mechanism, together with other salivary glands components, could help to explain the oral mucosa natural resistance to HIV infection.

HBD3 also possesses an inhibitory effect on the influenza virus blocking the fusion of the viral membrane with the endosome of the host cell, through cross linking of the viral glycoproteins [193].

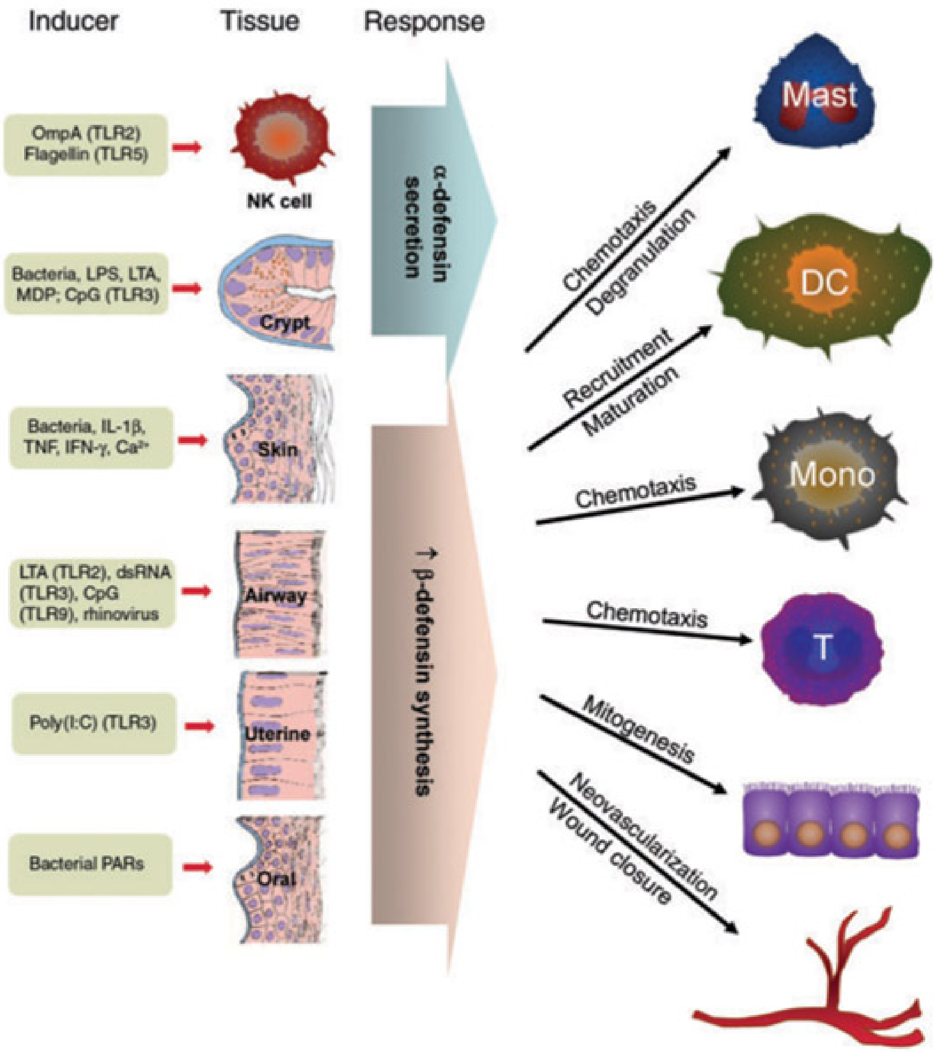


Figure 1.29: Immune functions of defensins may be induced by various physiological stimuli to mobilise pre-formed α-defensins or to upregulate β-defensin expression in various tissues. Released peptides interact with many target cells and tissues to promote secondary responses that may be critical for regulating acute inflammation, the recruitment of adaptive immune cells, angiogenesis and wound healing. NK, natural killer; LTA, lipoteichoic acid; MDP, muramyldipeptide; IL-1β, interleukin-1β; TNF, tumor necrosis factor; IFN-γ, interferon-γ; dsRNA, double-stranded RNA; PAR, protease-activated receptor; Mast, mast cell; DC, dendritic cell; Mono, mononuclear cell; T, T cell[145].

Gene	**Protein alias**	**Peptide**	**Tissue distribution**	**Synthesis and regulation**
DEFB1	MGC51822	Human β-defensin 1	Oral and nasal mucosa, lungs, plasma, salivary glands, small and large bowel, stomach, skin, eyes, mammary glands, urogenital tract and kidneys (HBD3 only)	lipopolysaccharide and peptidoglycan
DEFB4	DHBD2, SAP1, DEFB2	Human β-defensin 2		Inducible in response to viruses, bacteria, lypopolysaccharide, peptidoglycan, lypoproteins, cytokines (IL-1β,TNF) and growth factors. HBD3 constitutive expression on ocular surface
DEFB103	HBD3, HBP3, DEFB3, BD3	Human β-defensin 3		
DEFB104	HBD4, DEFB4, BD4	Human β-defensin 4	Gastric antrum and testis	Constitutive or inducible in response to PMA (phorbol 12-myristate) and bacteria
DEFB105	HBD5, BD5	Human β-defensin 5	testis	*in vitro* antimicrobial activity against *E.coli* but not *S.aureus [194]*. Constitutive mRNA expression in testis [141]
DEFB106	HBD6, BD6	Human β-defensin 6		
DEFB107	HBD7, BD7	Human β-defensin 7	Oral mucosa, testis	Constitutive mRNA expression in gengival keratinocytes. Constitutive mRNA expression in testis [141]
DEFB109	HBD9, BD9	Human β-defensin 9	Oral mucosa, ocular surface [152]	Constitutive mRNA expression in gengival keratinocytes. Constitutive expression on ocular surface

Table 1.3: β-defensins present on human chromosome 8p23.1 and correspondent activities. Adapted from Klotman ME *et al.* 2006 [195].

Another important aspect is the association between high copy number of β-defensins and the inflammatory skin disease psoriasis in humans [121]. In particular, this disease association study revealed that individuals with more than 5 β-defensin copies present a 5 fold increased risk of developing psoriasis when compared to 2 copies individuals and that there is a positive linear correlation between number of copies and relative risk increase.

Interestingly, some members of the β-defensin family have also an important role in mammalian reproduction. For example, there are five human defensin genes (*DEFB125*→*DEFB129*) clustered on chromosome 20, which are highly expressed in the male genital tract, in particular in the epithelial cell layer of the epididymal duct, which secretes factors responsible for sperm maturation [196]. Moreover, human *DEFB118* was shown to be a potent antimicrobial peptide able to bind to sperm, probably providing protection from microorganisms present in the sperm ducts [197]. It is noticeable how in long tailed macaque (*Macaca fascicularis)* and in rhesus macaque (*Macaca mulatta*) there is a similar β–defensin, called *DEFB126*, which is the principal protein that coats sperm [198]; this coating must be lost in the oviduct in order to allow fertilisation.

Beside their role in immune system response and fertility, recent studies have suggested that some of the β-defensin gene products, HBD-1 and HBD-3, can interact with a family of melanocortin receptors [199], modulating pigment expression in dogs and possibly in humans. Usually there are two genes that control pigment type switching: the melanocortin receptor 1 (Mc1r) and Agouti, encoding a ligand for the Mc1r which inhibits Mc1r signalling. Mc1r activation determines production of the dark pigment eumelanin exclusively, whereas Mc1r inhibition causes production of the lighter pigment pheomelanin. Interestingly, in dogs it was discovered a mutation in the canine *DEFB103* responsible for the dominant inheritance of black coat colour, which is not signalling directly through Mc1r; this phenomenon unravelled a not previously known role of β–defensins in controlling skin pigmentation. Further studies have been conducted on human melanocytes, discovering a novel role of HBD3 as antagonist of the α-melanocyte-stimulating hormone (α-MSH, a known agonist of Mc1r, which stimulates cAMP signalling to induce eumelanin production). As HBD3 is produced by keratinocytes, it can act as paracrine factor on melanocytes modulating α-MSH effects on human pigmentation and consequently the UV response [200]. Moreover, it is known that melanocortin receptors are also involved in inflammatory and immune response modulation [201], so β-defensins could play a role also in this kind of signalling, hence suggestive of a broader range of action than expected.

2. MATERIALS AND METHODS

Buffers compositions are given in Appendix 2.

2.1 DNA samples used

2.1.1 High quality *Macaca mulatta* DNA samples

A cohort of 16 unrelated rhesus macaque DNA samples was provided from the laboratory of Professor Charles L. Bevins, University of California Davis, CA. DNA was extracted from blood and lyophilised prior to shipping. For the high-density array CGH (see section 2.4) DNA was reconstituted in water at a final concentration of 100 ng/µl. For all the PCR based assays, DNA was used at a concentration of 10 ng/µl.

2.1.2 *Macaca mulatta* DNA samples for PCR based assays

A cohort of 51 rhesus macaque DNA samples, from different incomplete trios, was provided by Dr Omer Gokcumen, Department of Biological Sciences-SUNY, Buffalo US, at concentration suitable for PCR assays.

2.2 Cell lines

6 different rhesus macaque lymphoblastoid cell lines were used for this project, established and provided by Dr Gaby Doxiadis, Biomedical Primate Research Centre Rijswijk, NL (**Table 2.1**).

Cell line name	*Macaca mulatta* subspecies	Relatedness
r00068	Indian rhesus macaque	Family member of 2BX
2BX	Indian rhesus macaque	Family member of r00068
4049	Burmese rhesus macaque	Father of 96090
96090	Burmese rhesus macaque	Offspring of 4049
2BZ	Indian rhesus macaque	Family member of 8765
8765	Indian rhesus macaque	Family member of 2BZ

Table 2.1: List of rhesus macaque lymphoblastoid cell lines used.

2.3 PCR and electrophoresis

2.3.1 Standard PCR optimisation protocols

All the PCR-based assays were optimised for the following parameters, to maximise the specificity of the reaction for the expected PCR products: annealing and extension temperature, extension time and buffer composition. For each new primer pair, a standard annealing temperature gradient between 45°C-70°C was routinely performed (six temperature points with a difference of 5 °C each) followed by a more stringent gradient (six temperature points with a difference of 1 °C each) centred on the optimal annealing temperature detected on the basis of the standard gradient. Each reaction was first performed in standard 10X Kapa Buffer A (Kapa Biosystems) with 1.5 mM magnesium chloride. In presence of non-specific products, smears or low amplification efficiency, buffers with different compositions were tried, repeating the stringent annealing temperature step: 10X low dNTPs PCR mix and 11.1 X PCR mix (see Appendix 2). In case of AT-rich products, magnesium chloride titration series were performed, on six points, with either 0.25 mM increases or 0.5 mM increases, from 1 mM. Primers were used at a final concentration of 0.5 µM. KapaTaq (Kapa Biosystems) polymerase enzyme was used at a final concentration of 0.063 U per µl of PCR reaction. 5-20 ng of genomic DNA were used for each PCR reaction. The following standard PCR amplification protocol was used for expected PCR products up to 1kb:

- 2 minutes initial denaturation at 95°C
- 35 cycles of
    - 30 seconds denaturation at 95°C
    - 30 seconds at the optimal annealing temperature for each primer pair used
    - 30 to 60 seconds of extension time, depending on the expected PCR product size at 72°C (68°C with 11.1 X PCR mix)
- 1 minute at 56°C
- 20 minutes of final extension at 72°C (68°C with 11.1 X PCR mix)

All the PCR amplifications were performed using Applied Biosystems Veriti 96 wells thermal cyclers, in a final reaction volume of 10 µl (or 25µl for long range-PCR protocols) in 0.2ml thin walls tube, strips or plates (Thermo Scientifics). Preparatory

PCR for capillary electrophoresis were performed using the same conditions herein described, between 25 and 27 cycles.

2.3.2 Long range PCR

PCR amplification of products longer than 1kb was routinely performed in 11.1X buffer, with the addition of 0.1 U *Pfu* (Thermo Scientifics CAT N° EP0501) per reaction; *Pfu* is a DNA polymerase enzyme that possesses 3' to 5' exonuclease proofreading activity, correcting nucleotide misincorporation errors. A dedicated protocol was used, in final reaction volumes of 25µl:

-initial denaturation at 94 °C for 1 minute

-16 cycles of:

- 15 seconds denaturation at 94 °C
- 10 minutes elongation at 68 °C

-12 cycles of:

- 15 seconds denaturation at 94 °C
- elongation at 68 °C with time gradient (Δ=15 seconds) starting at 10 minutes

- 10 minutes of final elongation at 72 °C

All the PCR amplifications were performed using Applied Biosystems Veriti 96 wells thermal cyclers.

2.3.3 Digital Droplet PCR

Primers specific for the *DEFB2L* gene were designed using *Macaca mulatta* reference sequence from the UCSC Genome Brower (http://genome.ucsc.edu/) January 2006 assembly (MGSC Merged 1.0/rheMac2) as described in section 2.3.4. The design of the fluorescent Taqman® probe specific for the *DEFB2L* PCR product was done using the dedicated 'hybridisation probe (internal oligo)' tool on Primer3 software (http://frodo.wi.mit.edu/).

Primer sequences and fluorescent probe sequence for the reference diploid gene *PAX9* on chromosome 7 were taken from the published data of LEE AS *et al.* 2008 [79]. Primers were manufactured by Sigma-Aldrich. Fluorescent probes were synthesised by Applied Biosystems: *DEFB2L* probe was conjugated on the 5' to 6-carboxyfluorescin (FAM; excitation length 495 nm, emission length 515 nm) fluorophore and *PAX9* probe was conjugated on the 5' to VIC (excitation length 535 nm, emission length 555 nm). Both probes were conjugated on the 3' with the quencher MGB (dihydrocyclopyrroloindole tripeptide minor groove binder); when the quencher lies in proximity to the fluorescent probe it inhibits fluorescence emission; following PCR amplification, the probe annealed to the template is degraded by the 3'-5' exonuclease activity of the DNA polymerase, releasing the fluorophore and separating it from the quencher. This results in the emission of a fluorescent signal proportional to the concentration of amplified product and detectable by a fluorescence reader. Primers and probes sequences are reported in **Table 2.2**.

Oligo name	**Oligo type**	**Sequence 5'->3'**	**Chromosomal coordinates**	**PCR product size**
DEFB2L fw	Primer	TGATATAAGGAATCCTGTTACC TGC	chr8:8072379 +8072503	125bp
DEFB2L rw	primer	ATGGCTTTTTGCAGCATTTT		
FAM-DEFB2L-MGB	probe	[FAM]GCCATATGTCATCCAGGC TT [MGB]		
PAX9 fw	Primer	GAAACACATCCGGACCTACAAG	chr7:9962030 +99620406	100bp
PAX9 rw	primer	CACGTTGTACTTGTCGCACAC		
VIC-PAX9-MGB	probe	[VIC]CATCTTCGCCTGGGAGATC [MGB]		

Table 2.2: List of primers and probes used for digital droplet PCR.

For the Digital Droplet PCR (ddPCR) each reaction was prepared in a final volume of 22µl, containing: 900nM of each primer, 250 nM of each probe, 11µl of 2X ddPCR® Supermix (Biorad) and 10ng of DNA. 20µl of the reaction mix was placed in a droplet generator cartridge (Biorad, CAT N° 186-3008), in dedicated sample wells. 70µl of droplet generation oil (Biorad, CAT N° 186-3005) was placed in the oil dedicated wells. The cartridge was then placed in a QX100 droplet generator (Biorad, CAT N° 186-3002), a device using microfluidics to combine oil with the sample reaction to generate ~200000 monodispersed, nanoliter-sized droplets for each sample, suitable for ddPCR. 40µl of the emulsion obtained per each sample was then transferred in a 96-wells plate (Thermo Scientifics), sealed with aluminium foil (Thermo Scientifics, CAT N° AB-0757) using a plate sealer device (Eppendorf, CAT N° 951023078) and amplified to end point using a standard thermal cycler, with the following protocol:

-95°C for 10 minutes, to harden the droplets

40 cycles of:

-30 seconds of denaturation at 94°C

-60 seconds of annealing/extension at 56°C

And 10 minutes at 98°C to harden the droplets.

The plate is then loaded on a QX100 droplet reader (Biorad, CAT N° 186-3003) and results were analysed using the dedicated software QuantaSoft (Biorad). Reads with more than 10000 droplets generated were considered acceptable. Each gating separating the reads in *DEFB2L*+/*PAX9*-, *DEFB2L*-/*PAX9*+, *DEFB2L*-/*PAX9*-, *DEFB2L*+/*PAX9*+ was manually checked to ensure the presence of a discrete separation between the different droplets populations. The ratio of the total number of droplets single positive for *DEFB2L* on the total number of droplets single positive for *PAX9* called the diploid copy number for each sample in analysis.

2.3.4 Primer design

All general PCR primers were designed using the *Macaca mulatta* reference sequences for the region of interest from the UCSC Genome Brower (http://genome.ucsc.edu/) January 2006 assembly (MGSC Merged 1.0/rheMac2). Primer3 software (http://frodo.wi.mit.edu/) was used to check thermodynamic properties of all the 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.

All primers used in this study were manufactured by Sigma-Aldrich. General primers were produced with a scale of 0.025 µmoles of starting material for the first base of synthesis and purified through desalting, to remove residual by-products from synthesis deprotection and cleavage. Fluorescently labelled primers, designed to have the fluorophores 6-FAM or HEX conjugated to the 5' end of the left primer, were produced with a scale of 0.005 µmoles of starting material for the first base of synthesis and purified through separation of residual by-products using reverse phase cartridges. This step guarantees an increased purity for the oligonucleotides used for capillary electrophoresis, clearing effectively unincorporated fluorophores from the synthesised oligos.

2.3.5 Standard PCR validation with gel electrophoresis

Evaluations of PCR reaction fidelity were carried out running 5-10 µl of unclean PCR products with proportional amount of loading dye containing bromophenol blue (Bioline Cat N° BIO-37045) on gel electrophoresis. Different DNA markers were used, accordingly to the PCR product size range; in particular: HyperLadder I 10000-200 bp (Bioline Cat N° BIO-33025), HyperLadder IV 1000-100 bp (Bioline Cat N° BIO-33029), HyperLadder V 500-25 bp (Bioline Cat N° BIO-33031).

Agarose gels were prepared dissolving variable concentration of agarose (SeaKem Cat N° 50005), dependent on the PCR fragment size, ranging between 0.6% and 3% (w/v) in 0.5X TBE (see appendix 2). All gels were stained adding ethidium bromide to a final concentration of 0.5 µg/ml (Sigma Cat N° E1510) to the warm molten agarose. Gels were run in 0.5X TBE buffer (saline buffer containing Tris-base 0.45M, Boric acid 0.4M, EDTA 0.1M) applying a potential difference of 70-120 V. Electronic gel pictures were taken using a Gene Flash Syngene Bioimaging device and kept for records.

2.3.6 Sequencing protocols

For sequencing reactions, a preparatory PCR was performed to enrich the product of interest, following the protocol presented in section 2.2.1, with the addition of 0.08 units of *Pfu* enzyme (Thermo Scientifics CAT N° EP0501) per each µl of PCR mix, to prevent nucleotides misincorporation that would reduce the accuracy of the sequencing process. Each PCR reaction was prepared in a final volume of 25 µl, of which 5 µl were run on gel electrophoresis (see section 2.3.5) to confirm the presence of PCR products of expected length.

When a single band was present, the remaining 20 µl were purified using two possible methods: QIAquick PCR purification kit (Qiagen CAT N° 28106), following the manufacturer's protocol, or enzymatic cleanup. For the latter, 20 µl of PCR product were cleaned up using: 4 units of Exonuclease I (*ExoI*) enzyme (Biolabs CAT N°

M0293S), 5 units of Shrimp Alkaline Phosphatase (SAP) enzyme (Usb CAT N° 70092Z) in 10X SAP buffer. Reactions were incubated at 37 °C for 2 hours. The clean-up step was necessary to remove unincorporated nucleotides and primers from the reaction.

When different PCR products were expected, all the PCR reaction was run on gel electrophoresis and the bands of interest cut from the gel, under blue light reader. The PCR products of interest were then purified using a Zymoclean Gel DNA recovery Kit (Zymo Research CAT N° D4002) following the manufacturer's protocol. In either case, PCR products yield was then estimated using NanoDrop spectrophotometer (Thermo Scientific); NanoDrop can measure the total absorbance of 1µl of DNA sample on the assumption that the maximum absorbance for DNA is at 260 nm. Alternatively, 5 µl of cleaned PCR product were run on a new agarose gel, and DNA quantification was performed comparing the intensity of the product band with a DNA ladder in the appropriate size range with known concentration for each band.

Following PCR products cleanup, the sequencing reactions were set as follows, in a total volume of 10 µl: 1 µl of Big Dye Terminator Ready Reaction mix (Applied Biosystems) containing four standard deoxynucleotides (dNTPs) and the four dideoxynucleotides (ddNTPs) each labelled with a different fluorescent dye; 1.5 µl of 5X Big Dye Terminator Buffer (50mM Tris-HCl (pH 9.0) and 2mM $MgCl_2$); 3.2 picomoles of left or right primer and 20-30 ng/kb of PCR product template. The sequencing reactions were then run using Applied Biosystems Veriti 96 wells thermal cyclers, according to the following 25 cycles protocol:

- 96 °C for 10 seconds
- 50 °C for 5 seconds
- Ramp to 60 °C over 10 seconds
- 60 °C for 4 minutes.

Sequencing reactions could be stored at this point of the protocol at 15°C or 4°C prior to processing. The sequencing reactions had to undergo a further cleaning step, to remove all the incorporated labelled nucleotides, as follows: reaction volumes were doubled with water and 2 µl of 2.2% SDS were added and carefully mixed. Reactions were then denaturated using a thermal cycler at 98 °C for 5 minutes followed by 10 minutes incubation at 25°C. The excess dye terminator was then removed by passing the reaction through a Performa DTR gel filtration column (Edge

Biosystems CAT N° 42453). Eluates were then submitted to the Protein Nucleic Acid Chemistry Laboratory of the University of Leicester (PNACL http://www.le.ac.uk/mrctox/pnacl/) and run through an Applied Biosystems 3730 sequencer. The sequencing data and the chromatograms retrieved from the PNACL sequencing service were then analysed using the software Sequence Scanner V. 1.0 (Applied Biosystems) or the software Codon Code Aligner (http://www.codoncode.com/aligner/).

2.3.7 Capillary electrophoresis

Capillary electrophoresis was carried out in this study using an ABI Genetic Analyzer 3130 XL instrumentation (Applied Biosystems, UK). 0.7 to 2 µl of fluorescently- labelled PCR products were mixed with 10µl HiDi formamide containing 1% internal size standard (GeneScan-ROX400, Applied Biosystems, UK), included for precise determination of the PCR products length. After denaturation for 3 minutes at 96°C, the products were separated on POP-7 polymer (Applied Biosystem, UK) with 30 seconds of injection time and analysed by 3130 GeneScan 3.1 software (Applied Biosystems, UK). Scanning results of fluorescent-dye-labelled PCR products (peak area and peak size) by GeneScan software (Applied Biosystems, UK) were collected and manually transferred to Excel spreadsheet files for further analysis.

2.4 High density array Comparative Genomic Hybridisation

The high-density array CGH (aCGH) technique (see paragraph 1.3.2.1) allows the diploid typing of CNVs and can help to identify the breakpoints coordinates of a copy number variable area. Hybridisation and scanning procedures were performed by CXR Biosciences Ltd. which provided the raw data for the HD array-CGH experiment.

2.4.1 Probe design

The probe groups for the regions of interest were designed using the eArray software (https://earray.chem.agilent.com/earray/; Agilent Technologies), available online. Each probe is 60 bp long and it is potentially possible to create tiling path probes spanning the region of interest. The software then filters each possible probe to guarantee similar melting temperature for better hybridisation kinetics and uniqueness in the genome, to prevent possible non-specific annealing. Each probe set was repeated 5 times, to increase the robustness of the results. The manufacturer recommended a distance between each probe of at least 150-200bp, to avoid the competition of different probes for the same hybridisation fragment. For this reason, different minimum probe spacing parameters were tried for each region in analysis, to monitor possible differences in the number of filtered probes.

2.4.2 Probe selection

The following regions of interest (**Table 2.3**) have been selected based on previous genome wide array-CGH studies [79] [81], where they were reported to be CNV.

Regions	**Chr start**	**Chr end**	**Size (bp)**	**Number of probes**	**Probes X5**
β-defensin chr8	7923933	8906098	982165	875	4375
FCGR chr1	140063021	140199300	136279	207	1035
Haptoglobin chr20	70498702	70663636	164934	284	1420
C4 chr4	31585906	31688351	102445	211	1055
UGT2B33 chr5	60459306	60851741	392435	182	910
CYP2A6	47120860	47520810	399950	249	1245
LILR KIR3DX1 chr19	60243955	60423009	179054	223	1115
SLC6A2-CES1-CES2 chr20	53849490	54227332	377842	441	2205

Table 2.3: List of genomic regions selected for the HD array CGH. All coordinates refers to January 2006 *Macaca mulatta* assembly (MGSC Merged 1.0/rheMac2).

Normalisation probes (1% of the probes, as recommended by the manufacturers), were then selected from non-CNV regions (**Table 2.4**).

Normalization probes	**Chr start**	**Chr end**	**Size (bp)**	**Number of probes**
TP53 chr5	95515208	95571818	56610	128
NFkB chr16	7.398.881	7.417.940	19059	27

Table 2.4: List of normalisation probes selected for the HD array CGH. All coordinates refers to January 2006 *Macaca mulatta* assembly (MGSC Merged 1.0/rheMac2).

Dosage positive controls were included, selecting 312 probes specific for non-CNV regions on the X chromosome (**Table 2.5**). This allows an immediate control of the signal quality, as female individuals are supposed to present a doubled signal intensity compared to males. These probes were specifically designed to span genic areas on the X chromosome not reported to be CNV. In fact, the only known CNV on rhesus macaque X chromosome is located at chrX:961142-993312; this region does not overlap segmental duplications, it is not CNV in human and contains the orthologous gene of the human *MXRA5* (matrix-remodelling associated 5, coding for an adhesion protein with leucine-rich repeats and immunoglobulin domains related to perlecan) [79]. Moreover, no probes were designed in the chromosome X pseudoautosomal regions, as they would have violated the assumption of hemizygosity.

Dosage positive controls	Chromosome start	Chromosome end	Size (bp)	Number of probes
FOXP3 chrX	47103043	47110851	7808	19
PABPC5 chrX	90190398	90193844	3446	14
ACE2 chrX	13306789	13359547	52758	88
TAZ chrX	152432485	152441707	9222	22
BRS3 chrX	134546301	134550569	4268	18
SAP chrX	122601790	122625752	23962	52
F9 chrX	137713620	137746171	32551	99

Table 2.5: List of dosage positive control probes selected for the HD array CGH. All coordinates refer to January 2006 *Macaca mulatta* assembly (MGSC Merged 1.0/rheMac2).

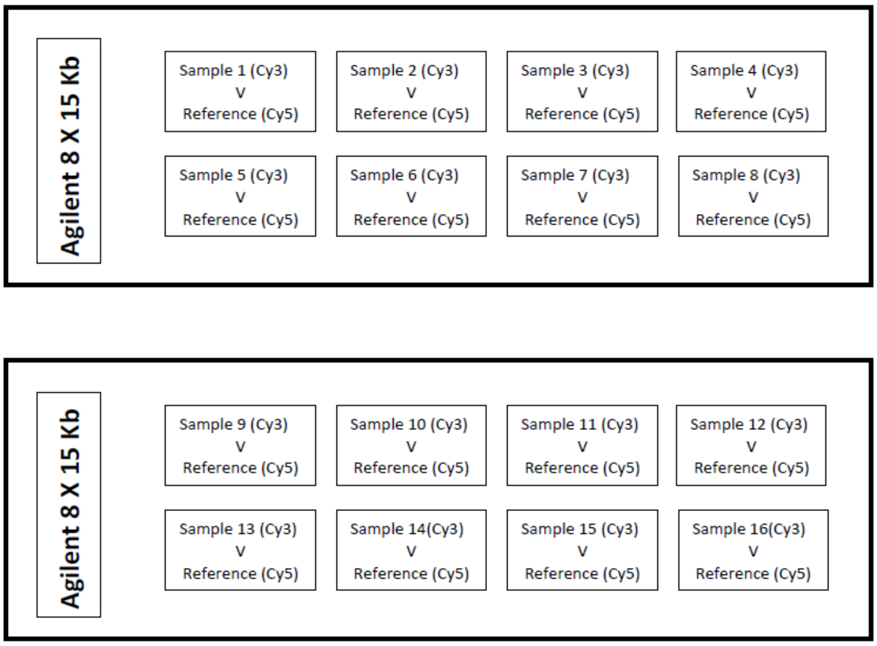
2.4.3 Experimental design

CXR Biosciences received 16 *Rhesus macaque* blood DNA samples of which one was randomly selected to represent the reference sample, obtained from 8 males and 8 females individuals. DNA samples were supplied at a concentration of 100 ng/µl for a total of 10 µg of DNA per sample, frozen on dry ice.

Samples were labelled with either cyanine 3 (Cy3) or cyanine 5 (Cy5), using the procedure outlined in the Agilent protocol ‘Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis’ (Agilent, CAT N° G4410-90010), and hybridised on Agilent Custom CGH 8x15K oligo microarray (CAT N° G4427A) in a two colour analysis (16 arrays in total). Agilent Custom CGH 8x15K oligo microarray (Agilent, CAT N° G4427A, AMADID 028407), and Agilent 8x15K gasket slides (Agilent# G2534-60014) were used for the study, sourced from Agilent Technologies Ltd. The same reference sample of rhesus macaque genomic DNA (sample N°16, Rh16) was used in each of the sixteen hybridisations.

2.4.4 Microarray Experimental procedures

The microarray analysis involved labelling DNA samples and reference samples as shown in Figure 2.1.



Agilent 8 X 15 Kb

Sample 1 (Cy3) V Reference (Cy5)

Sample 2 (Cy3) V Reference (Cy5)

Sample 3 (Cy3) V Reference (Cy5)

Sample 4 (Cy3) V Reference (Cy5)

Sample 5 (Cy3) V Reference (Cy5)

Sample 6 (Cy3) V Reference (Cy5)

Sample 7 (Cy3) V Reference (Cy5)

Sample 8 (Cy3) V Reference (Cy5)

Agilent 8 X 15 Kb

Sample 9 (Cy3) V Reference (Cy5)

Sample 10 (Cy3) V Reference (Cy5)

Sample 11 (Cy3) V Reference (Cy5)

Sample 12 (Cy3) V Reference (Cy5)

Sample 13 (Cy3) V Reference (Cy5)

Sample 14(Cy3) V Reference (Cy5)

Sample 15 (Cy3) V Reference (Cy5)

Sample 16(Cy3) V Reference (Cy5)

Figure 2.1: Two colour microarray set up.

2.4.5 Probe labelling

DNA samples (0.5µg) in 13µl of dH_20 were labelled prior to microarray hybridisation using the Agilent Genomic DNA Labelling Kit PLUS (Agilent, CAT N° 5188-3509), according to the manufacturer's protocol (Version 6.2.1, February 2010).

2.4.6 Microarray hybridisation and scanning

Agilent Custom CGH 8x15K oligo microarray (Agilent, CAT N° G4427A, AMADID 028407), were hybridised using Agilent Oligo aCGH Hybridization Kit (Agilent, CAT N° 5188-5220), according to the manufacturers protocol (Version 6.2.1, February 2010). Arrays were washed and scanned on an Agilent Microarray Scanner. Images from the scanner were processed using Agilent Feature Extraction Software v9.5.3.

2.4.7 Data normalisation

Data were produced as logarithms of the signal intensity ratio between the test and the reference. As each probe was repeated 5 times, analysis was performed on the averaged value for each data point. In order to correct for the intrinsic oscillation in the detection of the reference signal intensity, each value was corrected by subtracting the median signal intensity ratio value of the reference, for each genomic area analysed.

2.4.8 Breakpoints analysis

After normalization, each data set was analysed for breakpoints detection using BreakPtr, a computational approach for fine-mapping CNVs based on high-resolution CGH data [202]. Software available on the website: http://tiling.mbb.yale.edu/BreakPtr/.

A discrete-valued, bivariate hidden Markov model was generated applying two different training parameters. The values obtained for the normalisation probes present the standard for non-CNV calls, whereas the values for the dosage positive control probes (see paragraph 2.2.1) represent the standard for a 1-copy call for the 8 male samples and a 2-copies call for the 8 female samples. An alternative discrete-valued, bivariate hidden Markov model was generated using two training templates included with the software documentation, in order to compare the consistency of the data generated. The models were trained on human chromosome 22 high-resolution CGH data and contain known large aberrations mapping to the β-globin locus.

Upon the model generation phase, each data set from the 8 regions of interest has been scanned accordingly to the chosen parameters. For a given transition probability, the software identifies the areas with no significant deviation from zero (state 0, or diploid state), the areas showing significant signal increase (State 1, or amplification state) and the areas showing significant signal decrease (State 2, or deletion state). The authors suggest estimating the transition probability of the 3-states hidden Markov model by dividing the number of previously known/presumed aberrations (based on approximately mapped deletions/duplication) in the concatenated training set by the number of probes. As in our case an *a priori* assumption of the presumed aberrations was not applicable, a conservative transition probability of 1e^{-4} was selected.

2.4.9 Control PCR on sex attribution

Sex attribution of the rhesus macaque samples used for the array-CGH was performed accordingly to the procedure described in Wilson JF *et al.* 1998 [203]. 10ng of DNA per each sample were amplified using the following primers, at a final concentration of 0.8µM:

Rik8 5' ATTCCAGGCAGTACCAAACAG 3'

Jim9 5'CCATCAGGGCCAATAATTATTG 3'

Primers specificity has been checked with UCSC Genome Brower *In Silico* PCR tool on the January 2006 *Macaca mulatta* assembly (MGSC Merged 1.0/rheMac2). Cycling conditions were: 94°C for 2 minutes, 30 cycles of 94°C for 30s, 62°C for 30s and 68°C for 1 min 40s, and final extension at 68°C for 7 minutes.

2.5 Paralogue Ratio Test

In order to measure rhesus macaque *DEFB2L* copy number per diploid genome, the first technique optimised was a Paralogue Ratio Test [35] specific for the DEFB2L region.

2.5.1 PRT primers design

Given the complexity of the region in analysis, a bioinformatics approach was preferred to design suitable primers for PRT. For this aim, Dr Colin Veal developed a dedicated PERL script that combines in series three different bioinformatics tools. In particular, the script reads the UCSC Genome Browser *Macaca mulatta* reference sequence for the *DEFB2L* area querying Primer3 software to generate all the possible primer pairs amplifying a PCR product of suitable size (from 100bp to 400 bp) in the region of interest. Subsequently, it interrogates the reverse ePCR tool available on the NCBI website (http://www.ncbi.nlm.nih.gov/projects/e-pcr/) against all the rhesus macaque genome, to retrieve all the PCR products of similar size amplifiable with the same primer pairs identified through Primer3. The last step is a screening of all the suitable PCR products candidates against the rhesus macaque reference sequence available on the BLAST website (http://blast.ncbi.nlm.nih.gov/). This pipeline was designed to ascertain whether the PRT primer candidates matched just twice in the entire genome, on the test and on the reference locus, to guarantee the maximum specificity to the assay. This computational analysis was run using the High Performance Computing (HPC) cluster ALICE, property of the University of Leicester.

2.5.2 PRT primers optimisation

The six primer pairs with the highest score in terms of perfect matches in the area of interest and on the reference locus and with thermodynamic properties most suitable for PCR amplification, in terms of CG content, length and melting temperature, were ordered and tested according to the protocols described in section 2.3 (**Table 2.6**). A further manual control was performed, to confirm that none of the reference loci belonged to any known rhesus macaque CNV region, reported by the study of Lee *et al.* 2008 [79] and Gokcumen *et al*. (personal communication, with data published in 2011 [81]). The coordinates of all the candidate primer pairs have hence been checked against the coordinates reported for CNV regions identified by these authors.

Primer name	Sequence 5'->3'	Chromosomal coordinates	PCR product size
Alu Left	AGTGCAGTGGCACAATCTCA	chr11: 99853975-99854204 chr8: 8065922-8066139	chr 11: 229bp chr 8: 217bp
Alu Right	ACACTTTGGGAGGTTGAGGG		
PRT 8-16 Left	TGCAGGCTTAACACCACTTG	chr16: 47698513-47698893 chr8: 8075307-8075705	chr16: 380bp chr 8: 398bp
PRT 8-16 Right	TTTGGAAAATTTGCAGCCTA		
PRT 8-9 Left	ACTTGCAGGCTTAACACCAC	chr8: 8075304-8075502 chr9: 47392822-47393021	chr8: 198bp chr9 199bp
PRT 8-9 Right	GGCCTAGGAGGAAAGAATGG		
PRT 8-14 Left	GCCCTTTGAGCTGAGGCT	chr14: 70588986-70589098 chr8: 8075386-8075502	chr14: 112bp chr8: 116bp
PRT 8-14 Right	GGCCTAGGAGGAAAGAATGG		
PRT 8-3 Left	AACACCACTTGAAAGCTGCC	chr3: 21197456-21197717 chr8: 8075316-8075576	chr3: 261bp chr8: 260bp
PRT 8-3 Right	GCCAAGACAATGGGGAAA		
PRT 8-8 Left	TGTCCTTCTCACATTGCAAAA	chr8: 8074589-8074853 chr8:133814175-133814424	chr8: 264bp chr8: 249bp
PRT 8-8 Right	GACATGTGTGGGGCCTGTA		

Table 2.6: PRT primer candidates. All coordinates refer to January 2006 *Macaca mulatta* assembly (MGSC Merged 1.0/rheMac2).

After optimization, the primer pairs generating the more specific PCR products were ordered conjugated to the fluorophores 6-FAM or HEX, as described in Section 2.3.

2.5.3 PRT 8-14 conditions

All PCR reactions were performed with the following conditions: 1X 11.1X buffer, 0.5 µM of each primer, 0.63 U of KapaTaq (Kapa Biosystems) polymerase enzyme and 10 ng of genomic DNA. PCR was performed with the following cycling conditions:

- 2 minutes initial denaturation at 95°C
- 23 to 27 cycles (dependent on genomic DNA quality) of:

- 30 seconds denaturation at 95°C
- 30 seconds at 62 °C
- 30 seconds of extension time at 68°C

- 1 minute at 56°C
- 20 minutes of final extension at 68°C

Each sample was amplified in duplicate on the same PCR plate, and PCR amplification was repeated at least twice, to take into account the intra- and inter-experimental variability. PCR products were stored at 4 °C prior to processing. All the PCR products that were not run on capillary electrophoresis were maintained at 4 °C in case other duplicates from the same plate were necessary. Capillary electrophoresis was performed following the conditions described in section 2.3.7.

2.5.4 PRT 8-14 data analysis

For PRT 8-14, each peak correspondent to 112bp (chr 14) and 116bp (chr 8) was manually inspected using GeneScan software (Applied Biosystems, UK). Peak shape was an important criterion for sample inclusion in the analysis: real peaks present a sharp profile, whereas peaks with blunted top or irregular peaks indented over different base pairs are likely to be artefacts or results of insufficient PCR amplification. Saturated peaks, with fluorescence intensity over the detection scale, and peaks with heights lower than 200 in the signal intensity axis were excluded from the analysis.

Estimation of diploid copy number was calculated as the ratio between the peak area of the test peak at 116bp and the peak area of the reference peak at 112 bp, multiplied by two, under the assumption of the diploid copy number being double the copy number ratio values. Within each experiment, the average of the two duplicates was accepted if the difference between the ratios was less than 15% of their mean. The overall average between each experiment was accepted if the difference between the ratios was less than 15% of their mean.

2.6 Microsatellite assay

In order to get more information on the minimum *DEFB2L* copy number per diploid genome, a rhesus macaque specific assay was designed and optimised, to amplify a simple microsatellite repeat $(GGA)_n$ present in *DEFB2L* intron at the following coordinates: chr8:8071594-8071773.

2.6.1 Microsatellite assay conditions

Primers were manually designed using the software Primer3, as described in section 2.3.4. The primer sequences selected were the following, spanning 299bp on chromosome 8: 8071522-8071820:

Forward: 5'[6-FAM]tggcaccagagacctcaaat-3'

Reverse: 5'-cctggtttcaacctcattcttc-3'

All PCR reactions were performed in a total volume of 10 µl, with the following conditions: 1X 11.1X buffer, 0.5 µM of each primer, 0.63 U of KapaTaq (Kapa Biosystems) polymerase enzyme and 10 ng of genomic DNA. PCR was performed with the following cycling conditions:

- 2 minutes initial denaturation at 95°C
- 23 to 27 cycles (dependent on genomic DNA quality) of:
    - 30 seconds denaturation at 95°C
    - 30 seconds at 63 °C
    - 30 seconds of extension time at 68°C
- 1 minute at 56°C
- 20 minutes of final extension at 68°C

Each sample was amplified in duplicate on the same PCR plate, and PCR amplification was repeated at least twice, to take into account the intra- and inter-experimental variability. PCR products were stored at 4 °C prior to processing. All the PCR products that were not run on capillary electrophoresis were maintained at 4 °C in case other duplicates from the same plate were necessary. Capillary electrophoresis was performed following the conditions described in section 2.3.7.

2.6.2 Microsatellite assay data analysis

For the microsatellite assay, each peak in a range of 100-400bp was manually inspected using GeneScan software (Applied Biosystems, UK). Peak shape was an important criterion for sample inclusion in the analysis: real peaks present a sharp profile, whereas peaks with blunted top or irregular peaks indented over different base pairs are likely to be artefacts or results of insufficient PCR amplification Saturated peaks, and peaks with heights lower than 100 in the signal intensity axis were excluded from the analysis.

Estimation of minimum diploid copy number was calculated dividing the area of each peak by the minimum peak area. The individual ratios were then summed together to give the minimum diploid copy number estimate. Within each experiment, the average of the two duplicates was accepted if the difference between the minimum diploid copy number was less than 15% of their mean. For each sample, the overall average among different experiments was accepted if the difference between the minimum diploid copy number estimates was less than 15% of their mean.

2.7 InDel assays

The InDel Assay is designed to give preliminary indication of possible common breakpoints in the β-defensin region and the diploid copy number ratio. The region spanning the InDel is amplified with one specific primer pair. Using capillary electrophoresis (ABI 3100 Genetic Analyzer, Applied Biosystems), it is possible to discriminate the two variant alleles on size difference-base, using specific fluorescently labelled primers. In particular, the GeneScan software (Applied Biosystems) can be used to calculate the area under the curve of each peak of interest of the electropherograms resulting from the ABI run. This is a representation of the signal intensity, directly proportional to the concentration of the PCR product in the analysis. The ratio of the area under the curve for the two variants gives an indication of their copy number ratio.

InDel assays candidates for the rhesus macaque β-defensin region were designed through sequence alignment of the UCSC genome browser rhesus macaque reference Mmul_051212 against the same *Macaca mulatta* WGS reference sequence available on the BLAST database, covering 219kb on chr8 7960000-8179391. The quality check of the DNA sequences retrieved was performed using the BLAST tool 'Trace Archive'. In each case where the sequencing signal was too low or not clear the InDel was not validated. Primers tested are indicated in **Table 2.7.**

Primer name	Sequence 5'-3'	Tm (°C)	Expected size
DEFB107 L	TTCTCTCATTTTTCCTTCAA	52.6	320 bp
DEFB107 R	GAAATGTCTATTTCTCCCTTT	52	
DEFB106 L	GACAGGTAAGGAGGAGAGAGG	60	184bp
DEFB106 R	AAGGCACTGGCATTGGAAG	61	
DEFB4 (1) L	ACCATGGGACACAAGCAGAG	61.1	219bp
DEFB4 (1) R	CCCAGAGGTCGATGATATGC	60.5	
DEFB4 (2) L	GAGTGGGCCAGTGAAAGAAG	60	150bp
DEFB4 (2) R	TCATGAAATTGCCACTCACTC	58.7	
DEFB33 L	AGGCACCAAACCTCAGAACC	61.5	175bp
DEFB33 R	TGCATTCCTGATTCAAATGTTC	60	

Table 2.7: List of InDel assays primers, with expected PCR product size.

Fluorescent primers (Sigma Aldrich) were ordered for the 5 suitable InDels, with the fluorophore conjugated at the 3’ end of each forward primer. In particular; DEFB107, DEFB33 and DEFB106 were conjugated to the fluorophore FAM, whereas DEFB4 (1) and DEFB4 (2) were conjugated to the fuorophore HEX.

InDel assays were tested on *Macaca mulatta* genomic DNA. The PCR protocol to amplify the PCR products for the ABI run was set according to the manufacturers (Kapa Biosystems), with 23 cycles of extension. 1 µl of PCR product for each sample was then loaded on a 96 wells plate with 11µl of formamide + 2% GeneScan™ 400HD ROX™ Size Standard (Applied Biosystems) to proceed with capillary electrophoresis.
ABI data were then analyzed using GeneScan™ software (Applied Biosystems).

2.8 BAC library screening

Segment 1 CHORI-250 BAC library filters (http://bacpac.chori.org/rhesus250.htm) were ordered from the BACPAC Resources Center (BPRC) of the Children's Hospital Oakland Research Institute (CHORI).

2.8.1 Probe generation

In order to identify the clones containing the *DEFB2L* gene, 2 different probes specific for the *DEFB2L* region were designed (**Table 2.8**).

Primer name	**Sequence 5' to 3'**	**Chromosomal coordinates**	**PCR product size**
RhLeft1 fw	CCTCCGAGACATCACAGACA	chr8:8068951-8069850	900 bp
RhLeft1 rw	TCCTCTTATCTTGCGCCATT		
RhLeft2 fw	CAATGTCCTCACCAGCTTCC	chr8:8068812-8069761	950 bp
RhLeft2 rw	CCTGTAGCGTGGAACTCACC		
RhRight1 fw	TACCGAGCTCCTACCAGACC	chr8:8072076-8073004	929 bp
RhRight1 rw	GGAGGCGTAATTGCATAGGA		
RhRight2 fw	GGTGGAGCTGAAGTCACAAGA	chr8:8072049-8073004	956 bp
RhRight2 rw	GGAGGCGTAATTGCATAGGA		

Table 2.8: Primer pairs selected specific for the *DEFB2L* region. All coordinates refers to January 2006 *Macaca mulatta* assembly (MGSC Merged 1.0/rheMac2).

10 ng of rhesus macaque genomic DNA were used as template for each PCR reaction, to amplify the PCR products of interest. Long-range PCR protocol was used, as described in paragraph 2.1.2. Following PCR, the products were run on 0.8% agarose gel and bands were cut and purified using Zymoclean™ Gel DNA Recovery Kit (CAT N° D4002, Zymo Research). DNA concentration and purity was evaluated using a NanoDrop spectrophotometer (Thermo Scientifics).

2.8.2 Probe labelling and recovery

For each probe, 20 ng of probe DNA were radioactively labelled accordingly to the following procedure: probe DNA was denatured at 100 °C for 6 minutes in 30 µl of water and immediately cooled down on ice. DNA was then incubated overnight at room temperature with Oligonucleotide Labelling Buffer (5X OLB, see Appendix 2), 12 µg bovine serum albumin (BSA, Sigma Aldritch CAT N° A4737), 8.2 Units of DNA polymerase I Klenow Fragment (Cat N° M0210, Biolabs) and 0.02 mM of ^{32}P α-d-CTP. 70 µl of oligonucleotide stop solution (OSS, see Appendix 1) was added to stop the reaction. On ice, 10 µl of herring sperm DNA (10µg/µl stock, Sigma-Aldrich CAT N° D7290) and 30 µl of 3M NaAc (pH 5.6@25°C) were added prior to purification by ethanol precipitation. An alternative coating DNA tried to block DNA repeated elements was human placental DNA. Per each filter, 6µg of human placental DNA were diluted in 100µl of deionised water and sonicated at low intensity for 7 minutes, with 30 seconds cycles.

2.8.3 Filters hybridisation and washes

Filters were incubated in pre-hybridisation solution (6X SSC, 5X Denhardt's solution, 0.5% SDS, see Appendix2) for 2 hours at 65°C. 65 µg of herring sperm DNA was added to the ^{32}P-labelled probe and boiled for 5 minutes. 50 µl of probe solution was then added to the pre-hybridisation solution containing the filters, and incubated for 16 hours at 65°C.

Filters were then washed with four dilution of SSC starting at 2X and finishing with 0.2X containing 0.1% SDS for 15 minutes each; 2XSSC at 25°C, 2XSSC at 65°C and 0.2XSSC at 65°C, 500 ml each.

2.8.4 Positive clones identification

Positive clones were identified using two independent developing systems: filters were exposed to Fuji Super-HRE30 film at -80°C for 3 days or to phosphor imager screens (Amersham Biosciences) for 30 minutes (correspondent to 3 days of exposition with normal x-ray films) and scanned using Typhoon 9400 scanner (GE Healthcare). Positive clones were identified using a dedicated decoding grid provided by BACPAC Resources Center of the Children's Hospital Oakland Research Institute (CHORI).

2.8.5 BAC DNA extraction

Positive clones were shipped in LB agar stab culture format. The segment 1 of CHORI-250 BAC library was cloned in the low copies pTARBAC2.1 vector that confers chloramphenicol resistance (**Figure 2.2**). Colonies were picked from agar stab and grown overnight in LB containing 12.5 µg/ml of chloramphenicol at 37°C on shaker. On the following day, colonies were streaked on chloramphenicol + LB agar and incubated overnight at 37°C. Single colonies were then picked from each plate and pre-cultures were grown until OD_{600}= 5. After reaching the appropriate OD, pre-cultures were transferred into 400 ml cultures until OD_{600}= 2-3. Glycerol stocks in 20% glycerol were generated for long-term storage at -80°C.

BAC DNA was then extracted in two different ways. In the first approach anion exchange chromatography, with the gravity flow column kit NucleoBond® BAC 100 (Macherey-Nagel, CAT N° 740579) was used, aimed at the purification of large constructs (up to 300kb) without shearing. The extraction procedure was performed using the manufacturer's protocol for low-copy constructs. The second technique used, in order to increase the BAC DNA purity, was a caesium chloride gradient extraction.

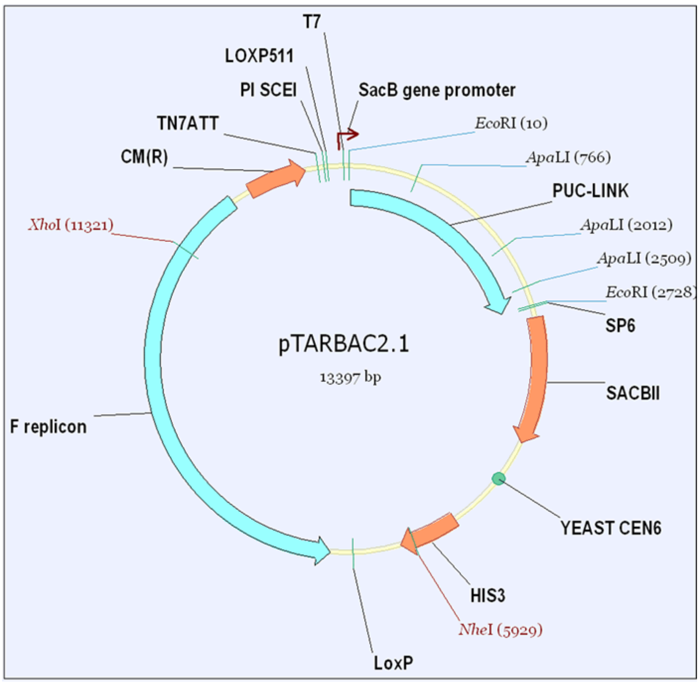


Figure 2.2: pTARBAC2.1 vector map**.**

2.8.5.1 Caesium chloride gradient BAC DNA extraction

Once reached an OD_{600}= 2-3, each 100 ml of bacterial culture were precipitated at 3220 g for 10 minutes at 4°C and pellet was resuspended in 6 ml of Solution I. After addition of 12 ml of Solution II, the solution was cleared by gentle inversion and subsequently 9 ml of ice-cold Solution III were added. Bacterial cell lysis was performed shaking the solution vigorously and, after 10 minutes of centrifugation at 3220g, it was possible to observe and remove a thick layer of cellular debris. The supernatant was placed on ice with an equal isopropanol volume for 30 minutes and then spinned at 4°C for 15 minutes. Pellet was washed in 70% ethanol, let dry to evaporate the alcohol and resuspended in 400 µl of TE buffer. Solution was brought to a 775µl volume with TE and further 775µl of 3% N-lauryl-sarkosine were added. The solution was mixed with 1.76g caesium chloride (1M in 1ml final centrifugation volume) and 60µl of 10µg/ml ethidium bromide and loaded in a ultracentrifuge sealable plastic tube. Ultra-centrifugation at 100000g was performed over-night. BAC DNA band, stained by ethidium bromide, was visible under blue light (the lower band of the gradient in case of bacterial genomic DNA contamination) and syringed out of the tube. Decontamination from ethidium bromide was performed through several washes in caesium chloride saturated isopropanol. BAC DNA was precipitated by incubation in isopropanol (0.5ml isopropanol and 0.54ml of water for each 0.4ml of DNA) at room temperature for 10 minutes and spinned at 200g for 10 minutes. Pellet was washed in 500µl of 70% ethanol, let dry and dissolved in 100µl of water.

2.8.5.2 BAC DNA quantification

BAC DNA concentration was quantified using kit Quant-iT™ PicoGreen® dsDNA reagent and kits (Invitrogen, CAT N° P7589) which is an ultra sensitive fluorescent nucleic acid stain specific for double-stranded DNA (dsDNA) in solution. This characteristic has the advantage of excluding possible co-purified RNA from the quantification. Fluorescence signals were detected using BMG Labtech Polar Optima Microplate Reader, using the following parameters: excitation filter 485p, emission filter 520p, gain 1300, 1 cycle, 10 flashes, 60 seconds of shaking time before reading.

Alternatively, NanoDrop readings (Thermo Scientifics) were performed, in order to monitor possible protein and solvents contaminations.

2.8.6 BAC sizing

For each clone, 1 µg of DNA was digested for 2 hours at 37°C with 5U of *NotI* restriction enzyme (NEB, R3189) in NEB buffer 3 (100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 1mM DTT pH 7.9@25°C) plus BSA, in a total volume of 10µl. Digested DNA was then embedded in 0.5% PFGE grade agarose blocks. BAC inserts were sized through pulsed field gel electrophoresis.

Gels for pulsed field electrophoresis were prepared using 1% PFGE grade agarose (Sigma, CAT N° A2929-25G) dissolved in 0.5% TBE (Fisher Scientific CAT N°BP13334).Gels were left equilibrate for at least 1 hour at 14°C before running in 0.5XTBE in ultrapure water and then run for 28 hours. Running conditions as follows:

5.5 V/cm, 5 seconds initial switch time, 120 second final switch time, 120 degrees switch angle, at 14 °C, on Biorad CHEF-DR®III System. Two different ladders were used in this project: λ ladder PFG marker (NEB, CAT N° N0340) and yeast chromosome PFGE marker (NEB, CAT N°N0345). Gels were then stained for 20 minutes in 0.5 μg/ml ethidium bromide solution. Electronic gel pictures were taken using a Gene Flash Syngene Bioimaging device and kept for records.

2.8.7 Validation of positive clones

For each clone, control PCRs were performed with the following primers (**Table 2.9**):

Primer name	**Sequence 5'->3'**	**Chromosomal coordinates**	**PCR product size**
SPAG11a Left	TAATGGGTTCTGCCATCCTC	chr8:8023447-8025272	1826 bp
SPAG11aRight	TTGCCAATGTTCTCAAATGC		
SPAG11b Left	AAGTGTTTGTTCCACCAGCAG	chr8:8025448-8028151	2704 bp
SPAG11b Right	TGACGGAGAATGGGTTTGAC		
DEFB103a Left	TCTGTGGCCTATTACCACGTC	chr8:8053103-8053995	893 bp
DEFB103a Right	TGGAGCAGAAGCTTACGAGAC		
DEFB103b Left	TGGCTGTGAGAGCTACAAGG	chr8:8056747-8058391	1654 bp
DEFB103b Right	CAGCCCATGTCTGATTCTTC		
DEFB103c Left	GCACTCCATCATAGGCCAAG	chr8:8058734-8059469	736 bp
DEFB103c Right	GGTGATGCCTAGTTCCCATAC		
DEFB103d Left	AGGTTTCCCAAACGTTCCTG	chr8:8059817-8060483	667 bp
DEFB103d Right	AACATTCTCGTCATGTTTCTGG		
DEFB103e Left	AAACAGGCAACTCAGCAAATC	chr8:8060621-8061699	1079 bp
DEFB103e Right	CTTCACTGGTCTGTTGCACTG		
DEFB4a Left	TTGTGTAAATAGTACCATGATTGGTG		

DEFB4a Right	GAGGATATGGACAGCCTCTACC	chr8:8068375-8070243	1869 bp
DEFB4b Left	CCTGTCTCTCTCTCCCTCTG	chr8:8070781-8071590	810 bp
DEFB4b Right	CTCCCTACTTCCCTGCCTTC		
DEFB4c Left	GAGGGAGGGACAGAGACAGG	chr8:8071764-8073577	1814 bp
DEFB4c Right	GGCTCTGTCATCACATCTGG		
DEFB4d Left	TCAGTTGCACTCACCTACCTTG	chr8:8073771-8074143	373 bp
DEFB4d Right	TGTGTTGTGTGAGCTTTGACC		

Table 2.9: List of primers used for BAC clones validation and mapping. All coordinates refer to January 2006 *Macaca mulatta* assembly (MGSC Merged 1.0/rheMac2).

2.8.7.1 BAC end sequencing

BAC end sequencing was performed using 1 µg of BAC DNA as template for each 30µl reaction. T7 promoter and SP6 promoter primers were used. Sequencing reactions were performed with the following modifications:

-1.2 µl of Big Dye Terminator Ready Reaction mix (Applied Biosystems);

-6 µl of 5X Big Dye Terminator Buffer (50mM Tris-HCl (pH 9.0@25°C) and 2mM $MgCl_2$)

-4mM extra $MgCl_2$ (4.8µl)

-20 picomoles of primer. The sequencing reactions were then run using Applied Biosystems Veriti 96 wells thermal cyclers, using the following protocol:

2 minutes denaturation at 96°C, followed by 30 cycles of:

10 seconds denaturation at 96°C

5 seconds of annealing at 50°C

4 minutes of extension at 60°C.

2.8.7.2 Whole BAC sequencing

Whole BAC sequencing on selected clones has been performed by GATC Biotech (Konstanz, Switzeland) using a Pacific Bioscience PacBio platform, based on single-molecule real-time sequencing. This technology, used for *de novo* genome sequencing, has the advantage of generating long sequence reads >3kb and up to

10kb. This allows to reduce the number of contigs generated and hence errors in the assembly phase, especially for complex genomic area that cannot be accurately resolved with sequencing technologies based on smaller reads. A dedicated *de novo* assembly algorithm, with quality values optimised for BAC assemblies using ultra long reads has been used in the assembly phase, called Allora (A Long Read Assembler). The output file will represent the final linear contig for the BAC clone in analysis.

2.8.8 *DEFB2L* sequencing

To proceed with *DEFB2L*sequencing from each BAC clone, a first long range PCR was performed (protocol described in section 2.1.2) using the following primers (**Table 2.10**):

Primer name	Sequence 5'->3'	Chromosomal coordinates	PCR product size
1 Left	GAACGTATCAATCAGGAGTGATG	chr8:8068005-8072685	4681bp
17 Right	ATCTCTAACATCCTTGGAACACC		

Table 2.10: List of primers used for long range PCR on the region spanning *DEFB2L* gene. All coordinates refer to January 2006 *Macaca mulatta* assembly (MGSC Merged 1.0/rheMac2).

Long range PCR products were purified as described in section 2.3.2 and used as templates for each single sequencing reaction, using the following primers (**Table 2.11**):

Primer	Sequence	Chromosomal coordinates	PCR product size
Exon1 Left	ATAAGATGGACGGCTTGGTG	chr8:8070054-8070243	190bp
Exon1 Right	GAGGATATGGACAGCCTCTACC		
Exon2 Left	CACGCTGTTTGCTCTTTGTG	chr8:8072332-8072610	279bp
Exon2Right	TCCGTAAATCTGAACACCACAG		
DEFB4b Left	CCTGTCTCTCTCTCTCCCTCTG	chr8:8070781-8071590	810 bp
DEFB4b Right	CTCCCTACTTCCCTGCCTTC		
DEFB4c Left	GAGGGAGGGACAGAGACAGG	From chr8:8071764	

Table 2.11: List of primers used for *DEFB2L* exon 1 and exon 2 sequencing.

Sequencing reactions were performed as described in section 2.3.6.

2.8.9 *DEFB2L* BAC variants analysis

Exon sequences retrieved from each clone were aligned with every *DEFB2L* variant already annotated on UCSC genome browser and BLAST website (http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9544) and further analysis were performed with the software CodonCode Aligner (available on: http://www.codoncode.com/aligner/). All sequences used, in FASTA format, are reported in APPENDIX 3. Alignments were performed using slow pairwise alignment option, IUB DNA weight matrix, gap open: 10, gap extension: 0.1.

Protein sequences were inferred using the Translate tool from ExPASy Bioinformatics Resource Portal (http://web.expasy.org/translate/) , multiple alignment were performed using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) and graphic representations of the levels of conservation of the different protein sequences were created using EsPript 2.2 tool available online (http://espript.ibcp.fr/ESPript/ESPript/).

Possible 3D representations for each DEFB2L protein variant were modelled upon the structure deposited for human hBD2 mature peptide (http://www.rcsb.org/pdb/explore/explore.do?pdbId=1FD3) using PyMOL software (http://pymol.org/).

2.8.10 Ka/Ks ratio calculation

The estimator used to evaluate the selective force acting on the *DEFB2L* gene was the Ka/Ks ratio. The KaKs_calculator [204] was used for this purpose. The software calculates the Ka/Ks ratio in three steps: on two aligned sequences, it is first necessary to count the numbers of synonymous (S) and nonsynonymous (N) sites (S + N = n) and the numbers of synonymous (S_d) and nonsynonymous (N_d) substitutions ($S_d + N_d = m$). A correction for multiple substitutions is then performed, so that (N_d/N) and (S_d/S) could represent Ka and Ks, respectively. This correction is necessary, given that the observed number of substitutions underestimates the real number of substitutions as sequences diverge over time. The probability of the substitution to occur depends on the substitution model applied; for this project, a Model Averaging (MA) maximum likelihood method implemented by the creators of KaKs_calculator was applied. MA model is based on the average of the parameters applied by different maximum likelihood candidate models. Ka/Ks ratio gives information about the type of selection acting on the sequences in analysis: Ka/Ks=1 indicates an equal proportion of non-synonymous and synonymous substitution, suggesting neutral evolution; Ka/Ks<1 indicates an excess of synonymous substitutions over non-synonymous, indicative of purifying selection; Ka/Ks>1 indicates a higher proportion of non-synonymous substitutions over synonymous, with the presence of more amino acidic changes than expected by chance, suggestive of diversifying selection.

2.9 MEGA analysis

Phylogenetic trees were reconstructed from both cDNA sequences and protein sequences from every *DEFB2L* orthologous sequences deposited online. Lower threshold for sequence similarity with *Macaca mulatta DEFB2L* was set at $5e^{-04}$.
Phylogeny reconstruction was performed using the software MEGA6 (Molecular Evolutionary Genetics Analysis http://www.megasoftware.net/).

For phylogenetic trees based on coding sequences, the following parameters were selected (**Table 2.12**):

Analysis	
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	Bootstrap method
N° of Bootstrap Replications	500
Substitution Model	
Substitutions Type	Nucleotide
Model/Method	General Time Reversible model (GTR)
Rates and Patterns	
Rates among Sites	Uniform rates
Data Subset to Use	
Gaps/Missing Data Treatment	Use all sites
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically
Codons Included	1st

Table 2.12: List of parameters used in MEGA5 software to generate phylogenetic trees based on coding sequences for all DEFB2L orthologous sequences deposited online

For phylogenetic trees based on protein sequences, the following parameters were selected (**Table 2.13**):

Analysis	
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	Bootstrap method
N° of Bootstrap Replications	500
Substitution Model	
Substitutions Type	Amino acid
Model/Method	Dayhoff model
Rates and Patterns	
Rates among Sites	Uniform rates
Data Subset to Use	
Gaps/Missing Data Treatment	Partial deletion
Site Coverage Cutoff (%)	95
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically

Table 2.13: List of parameters used in MEGA5 software to generate phylogenetic trees based on protein sequences for all hBD2 orthologous sequences deposited online.

2.10 FLUORESCENCE *IN SITU* HYBRIDISATION

2.10.1 BAC DNA Probe generation for FISH

Prior to probe generation, 50 ng of BAC DNA from each clone was amplified through Whole Genome Amplification protocol, using the GenomePlex® Single Cell Whole Genome Amplification (WGA) Kit (CAT N° WGA4-10RXN, Sigma), to increase the yield of each BAC extraction. Amplified products were then purified using Amicon® Ultra-0.5 ml 30K membrane centrifugal filter devices (CAT N° UFC503096, Millipore).

Each clone was labelled in 2 colours, biotin and digoxygenin, using two different kits: for biotin, 500ng of WGA DNA was labelled with BioPrime® Purification

Module (CAT N°18095-011, Invitrogen) using biotinilated dUTPs; for digoxygenin, 1 µg of WGA DNA was labelled with BioPrime® Array-CGH Genomic Labeling System (CAT N°18095-011, Invitrogen) using digoxygenated-dUTPs. Both labelled products were purified using PureLink™ purification columns (CAT N°18095-011, Invitrogen) following the manufacturer's protocols. For the chromosome 8 staining, a commercial whole chromosome 8 painting was used (Cytotest, Cat N° CT-WPP08-20).

2.10.2 PCR products Probe generation for FISH

Labelled PCR products amplified using BAC DNA as templates were also used as probes for fluorescence *in situ* hybridisation experiments. Each region of interest was separated into smaller PCR products that did not contained repeated elements, accordingly to the *Macaca mulatta* reference sequences from the UCSC Genome Brower (http://genome.ucsc.edu/) January 2006 assembly (MGSC Merged 1.0/rheMac2). Primers used for PCR amplification are listed in Table 2.14. Following PCR amplification, PCR products were run on agarose gel and specific bands were cut and purified with Zymoclean Gel DNA recovery Kit (Zymo Research CAT N° D4002) following manufacturer protocol. Each purified PCR product was quantified with NanoDrop Spectrophotometer (Thermo Scientific).

Prior to labelling, probes spanning the same region of interest were pooled together in the same tube, with the same molecular ratio, calculated from DNA concentration and PCR product length. This calculation has been necessary to ensure that smaller labelled products would not be over represented in the hybridisation reaction and to get homogeneous signal from the entire area of interest. Each probe mixture was labelled in 2 colours, using two different kits: for biotin, 500ng of WGA DNA were labelled with BioPrime® Purification Module (CAT N°18095-011, Invitrogen); for digoxygenin, 1 µg of WGA DNA was labelled with BioPrime® Array-CGH Genomic Labeling System (CAT N°18095-011, Invitrogen). Both labelled products were purified using PureLink™ purification columns (CAT N°18095-011, Invitrogen) following the manufacturer's protocols.

2.10.3 Testing for incorporation of labels

The efficiency of labels incorporation was tested using the following protocol: a piece of Hybond™ Nucleic acids blotting membrane (CAT N° 95038-336, GE

Healthcare), of 3cm x 3 cm, was placed in Buffer 1 (See APPENDIX 1 for buffers composition) for 5 minutes and blotted dry with filter paper. 1µl of each probe was allowed to dry on the membrane that was then placed in Buffer 1 for 1 minute, followed by 30 minutes of incubation in Buffer 2, with gentle shaking. antibody-AP mixture was prepared with Anti-Digoxygenin-AP FAB fragment (CAT N° 11093274910, Roche Diagnostics) and Anti-Biotin-AP FAB fragment(Roche) both 1:500 in Buffer 1. Membrane was then incubated in the dark for 30 minutes at 37°C in cling film with 0.5 ml of antibody-AP mixture. Membrane was washed in Buffer 1 for 15 minutes and then transferred to Buffer 3 for 2 minutes. Detection reagents INT/BCIPC (INT:(2-[4-iodophenyl]-3-[4-nitrophenyl]-5-phenyltetrazolium chloride) BCIP: (5-bromo-4-chloro-3-indolyl phosphate, toluidine salt in DMSO)) (CAT N°11681460001, Roche) was prepared in Buffer 3 (17.6 µl INT/BCIP stock solution in 5 ml Buffer 3) and used for membrane incubation for 10 minutes in the dark at room temperature. The incorporation efficiency is directly visible as dots on the membrane.

2.10.4 FISH on metaphase spreads

2.10.4.1 Cytogenetic harvest and generation of metaphase spreads from lymphoblastoid cells in culture

Rapidly dividing lymphoblastoid cell lines were incubated for 1 hour in 0.05 µg/ml colcemid (CAT N° 15210-040Gibco®) which depolymerises microtubules arresting the cells in metaphase. Cell clumps were gently broken with a Pasteur pipette before centrifuging at 100 rcf for 10 minutes. After medium aspiration, cell pellets were gently broken flicking the side of the tubes and 2 ml of hypotonic KCl solution at 37°C (0.075 M KCl) were slowly added in agitation. Further 8 ml of KCl hypotonic solution were added, and cell solutions were mixed by inverting the tubes. Cells were incubated at 37°C for 12 minutes and then centrifuged at 100xg for 8 minutes. Hypotonic solution was carefully removed and cell pellets resuspended by gently flicking. A few drops of fixative (methanol: glacial acetic acid = 3:1) at 4 °C were added to help resuspension and to avoid the presence of cell clumps. 2 ml of fixative were slowly added with agitation and other 8 ml more rapidly. Cell solutions were mixed by flicking the tubes. Cells were fixed for 20 minutes at 4°C. Cells were centrifuged at

100xg for 8 minutes and fixation procedure was repeated twice. After the last centrifugation, cells were resuspended in 200 µl of fresh fixative.

To generate metaphase spreads, one drop of fixed cells per slide was let fall from 15 cm height on a polylysine-coated slide (Cosmos Biomedical), held at with a 45° angle. Slides were then quickly fixed with 100% methanol in a Coplin jar and air dried. For long term storage slides were kept at -20°C with silica gel.

2.10.4.2 Post-fixation of air dried slides

Slides were fixed in 3 parts of ethanol and 1 part glacial acetic acid for 30 minutes, dehydrated twice for 5 minutes with 100% ethanol and let air-dry. Slides were then incubated for 1 hour at 37 °C in humid chamber with 100µg/ml RNase solution (CAT N° EN0531, Fermentas) in 2X SSC, under a plastic coverslip.

2.10.4.3 Paraformaldehyde fixation

Slides were fixed with 4% paraformaldehyde (pH 7@25°C) for 10 minutes at room temperature, in chemical hood and then washed twice in 2X SSC for 2 minutes and 10 minutes respectively.

2.10.4.3 Dehydration

Slides dehydration step was conducted with washes in increasing ethanol concentrations, respectively 70%, 85% and 100%, for 2 minutes each. Slides were then air-dried.

2.10.4.4 Hybridisation

Hybridisation probe mixes were prepared with the following components (**Table 2.14**):

Component	**Final concentration in hybridisation mix**
100% formamide	40%
20X SSC	1X
50X Dextran sulphate	10%
Salmon sperm DNA	0.025µg
100 mM EDTA	1.25 mM
10% SDS	0.125 %

Table 2.14: Components for the hybridisation mix.

40-100 ng of probe were added to the probe mix and, when needed, 1 µg of Cot-1 DNA (CAT N° 15279-011, Invitrogen™). Probe mixes were denatured at 75°C for 7 minutes, and immediately dispensed on the slides and covered with a small plastic coverslip. Metaphase spread slides were then denaturated at 72°C for 6°C, then cooled to 37°C and left hybridising in an *in situ* thermal cycler system (Thermo Scientific) for 16-20 hours.

2.10.4.5 Post-hybridisation washes

Slides were transferred in 2X SSC at 35-40°C in a Coplin jar on shaker to float off coverslips. For the following steps, temperature was accurately recorded, being critical for the quality of the *in situ* experiments. A first wash was done at 42 °C in 2X SSC for 2 minutes, followed by two washes of 5 minutes each in high stringency solution (20% formamide and 0.1X SSC) at 40-45°C. Slides were then washed twice at 42°C on shaker in low stringency solution (0.1X SSC) for 2 minutes and 10 minutes respectively. Slides were then washed in 2X SSC on shaker at 42°C for 5 minutes and let cool down at room temperature.

2.10.4.6 Detection

Slides were transferred in detection buffer (4X SSC and 0.2% Tween) for 5 minutes. 200 µl of blocking solution (5% BSA in detection buffer) were added per each slide, covered with a plastic coverslip and incubated at 37°C for 30 minutes. After draining, 50µl of antibody solution was applied on each slide; in this study the following antibodies were used: FITC antidigoxygenin (D8029-50FUS Biological) 1:100 and Alexa Fluor® 546 streptavidin 1:200 (CAT N° S-11225, Invitrogen) in blocking buffer. Slides were incubated with a plastic coverslip at 37°C in humid chamber for 1 hour and then washed in detection buffer for 2 minutes at 40 °C on shaker, followed by other two washes in detection buffer for 8 minutes each at 40 °C on shaker.

2.10.4.7 DAPI staining and mounting

Slides were then incubated with 100 µl of DAPI solution (4µg/ml) under a plastic coverslip for 30 minutes in the dark and then rinsed in detection buffer. Two drops of anti-fade solution (R1320, Agar Scientific) were added to each of the slides, covered with a large glass coverslip (Glass N° 0 24 mm X 40 mm) and stored at +4°C at least overnight prior to analysis. Slides were analysed using a Nikon fluorescence microscope, and multichannel pictures were acquired at 100X.

3. Analysis of rhesus macaque β-defensin region using high-density array CGH

The first approach utilised to assess the presence, extent and distribution of copy number variation in the rhesus macaque β-defensin region and in other disease-associated genomic areas of interest has been a high-density comparative genomic hybridisation array (Section 1.3.2.1). The rationale of this experiment was based on previous findings [79] [81].

In Lee *et al* 2008, the authors performed a genome wide array-CGH (with an average probe spacing of 6.5 kb) on 10 unrelated rhesus macaques, that identified a copy number variable area spanning the β-defensin locus (chr8: 8043509-8707577, 664 kb, covered by 33 probes). In particular, 4 macaques out of 10 showed variable signals of gain/loss without a consistent breakpoint pattern among individuals, as presented in **Figure 3.1**.

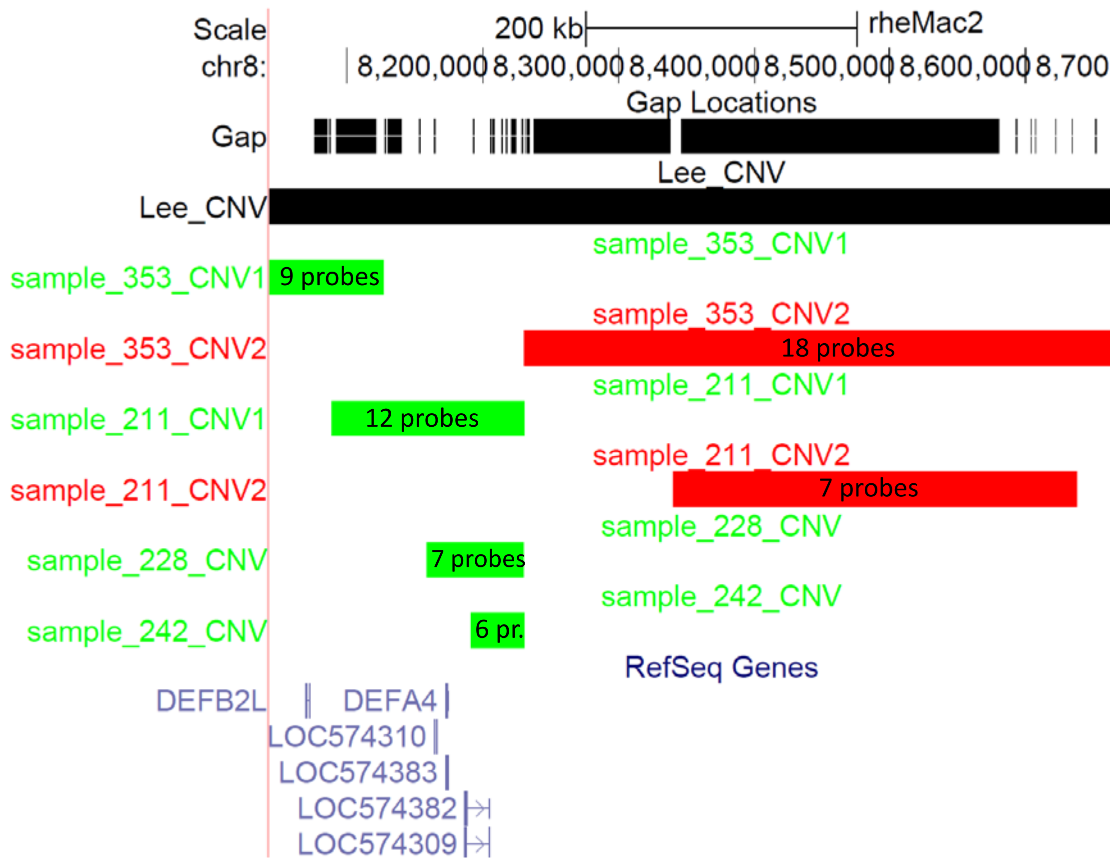


Figure 3.1: Boundaries of the CNV region spanning rhesus macaque *DEFB2L* and the α-defensin cluster as identified by Lee *et al* 2008 [79] with a dedicated genome-wide array CGH platform. From the top, the black track represents the CNV extent (chr8:8043509-8707577, 644 kb). A total of 4 samples (353, 211, 228 and 242) showed signs of gains (green track) or losses (red track) compared with the reference. The number of probes calling gains or losses is indicated inside each track, for a total of 33 probes.

In a following paper of Gokcumen *et al* 2011, the authors performed a second-generation genome-wide aCGH on 17 non-related rhesus macaque samples, with an average probe spacing of 3kb and an effective resolution for CNV identification of 15kb. This approach highlighted 1 Mb as copy number variable, with more than 200 probes and an average spacing of 5kb, spanning the β-defensin area with the following breakpoints: chromosome 8: 7923933-8906098. A comparison between the CNV size identified by Lee *et al* and the one reported by Gokcumen *et al* is shown in **Figure 3.2**.

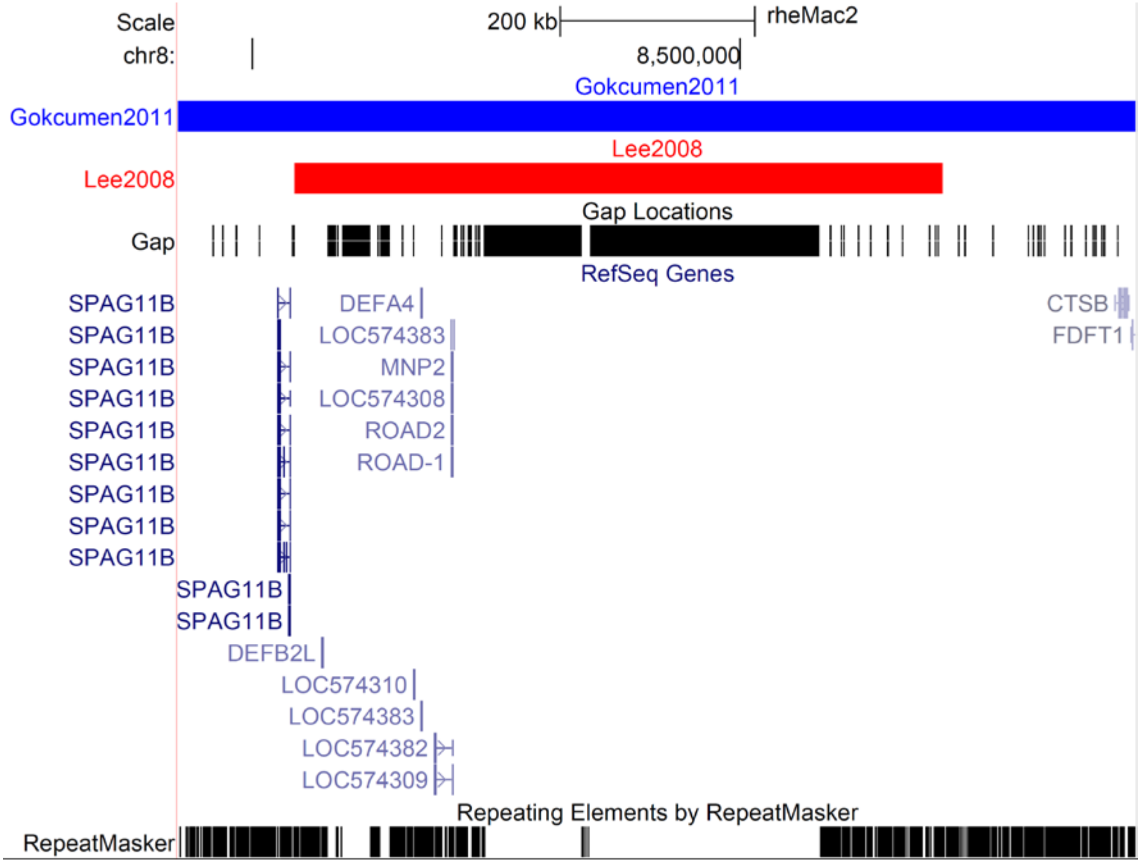


Figure 3.2: Comparison of the CNV boundaries identified by Lee *et al* [79] (blue track: 644kb called by 33 probes) and Gokcumen *et al* (red track: 857 kb called by235 probes).

From Figure 3.2 it is possible to observe important discrepancies in the identification of the CNV boundaries: the data of Lee *et al* suggest that the distal CNV breakpoint includes just the *DEFB2L* gene and not the rest of the β-defensin cluster; on the contrary, Gokcumen data indicate a CNV area spanning the entire β-defensin region. For these reasons, the creation of a new customised array-CGH platform targeting just this complex area seemed necessary to resolve its CNV boundaries.

Given the higher probe coverage obtained by Gokcumen *et al*, the log intensity values for this area of interest (provided by Omer Gokcumen (Department of Biological Sciences-SUNY, Buffalo, US) in a personal communication) have been used as a starting point for this project.

These data were smoothed over a 10-probe sliding window and the resulting averages were plotted (**Figure 3.3**), using as comparison a window with the same chromosomal coordinates and annotated genes from UCSC genome browser (January 2006 *Macaca mulatta* assembly (MGSC Merged 1.0/rheMac2)). From the plot it is possible to observe how the β-defensin locus is within a variable area, right at the edge of a region with no probe coverage, because of multiple gaps in the rhesus macaque genomic assembly and the presence of segmental duplications that impeded the probe design. Despite the lack of information immediately downstream of *DEFB2L*, these promising results constituted the basis for further analysis.

150 kb from the annotated *DEFB2L* gene, there is a noticeable variation in signal intensity in a region containing the α- defensin cluster and, 500kb proximal, it is observable another variable area corresponding to zinc finger family genes (*ZFN*) and *FAM90A* gene family. This latter family, in particular, has been previously annotated and characterised as copy number variable in human, and is a primate-specific gene family on chr8p23.1, with unknown function [205]. These data constitute a further confirmation of the high polymorphism rate of this genomic area, likely to be a CNV hotspot.

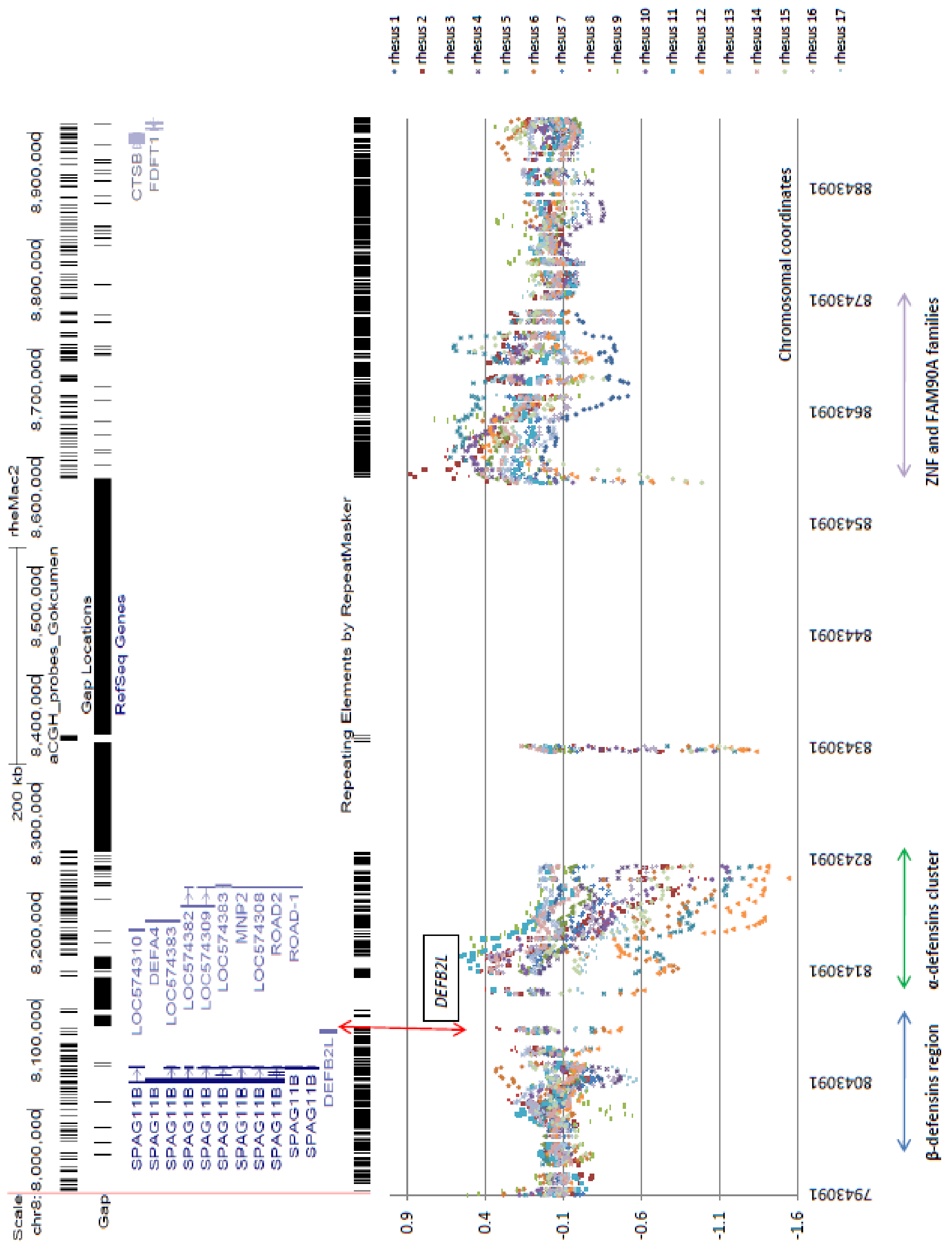


Figure 3.3: Distribution of log intensity ratio spanning the β-defensins locus from the array-CGH data of Gokcumen *et al*. 2011 [81].

The X axis indicates the chromosomal coordinates; Y axis indicates the log intensity ratios averaged 10-probe sliding window. The red arrow shows the position of the *DEFB2L* gene, within the β-defensins cluster indicated by the blue arrow.

3.1 Genomic regions of interest and probes selection

The choice of a 15K Agilent customised high-density CGH array was made to allow the simultaneous CNV detection of different genes of interest, beside the β-defensins cluster, that constituted the main target of this project. The selection of other plausible CNV candidate areas to analyse was done on the basis of the CNV regions identified by Lee AS *et al.* [79] and Gokcumen O *et al.* [81]. Among the regions reported to be CNV in both rhesus macaque and human, priority was given to disease-associated genomic areas reported from literature.

All probes designed have been tested to avoid bias due to the competition for the same hybridisation fragment. For all the regions in analysis, a minimum space of 200bp has been kept between adjacent probes, to remove the risk of cross-hybridisation derived from probes overlapping.

3.2 Data normalisation and filtering

All the data output are normalised against the reference sample (sample Rh16), and expressed as logarithm of the ratio of the processed signal intensity detected in the green channel (Cy3: test sample) on the signal intensity detected on the red channel (Cy5: reference sample).

3.2.1 Quantification of self-self hybridisation

The first step, in order to assess the hybridisation efficiency, has been the analysis of the output data generated from the self-self hybridisation of the reference individual Rh16. The log intensity ratios resulting from the hybridisation of the same DNA sample labelled with two different fluorophores was expected to cluster around 0, reflective of no bias in the fluorophores efficiency and high signal detection accuracy. All the values deviating from 0 should be considered as systematic measurement error; this background noise should be subtracted from all the test samples reads, to normalise the dataset (**Figure 3.4**). For all the areas considered as dosage positive controls, mapping to the X chromosome (312 probes) and diploid controls (155 probes), the values clustered around 0 as expected. As quality control, the percentage of probes giving a signal between -0.1 and 0.1 has been indicated for each gene of interest, in **Table 3.1**, together with average values and standard deviations. These results confirmed high hybridisation efficiency for the reference sample; the averaged fluorescence background noise of the reference individual has been calculated for each diploid control and dosage positive control area, and each resulting value has been subtracted to the log intensity ratios of each correspondent region for all the test individuals in analysis.

Dosage positive control on chrX	**Average log ratios**	**Standard Deviation**	**Percentage of probes with -0.1<log ratio<0.1**
SAP	0.002	0.1	79%
ACE	0.004	0.12	82%
BRS3	0.0001	0.09	89%
F9	0.0003	0.1	87%
TAZ	-0.008	0.17	50%
FOXP3	-0.11	0.25	42%
PABPC5	0.04	0.23	36%
	Global average dosage positive controls	**Global standard deviation dosage positive controls**	
Dosage positive controls	-0.009	0.14	78%
Diploid control	**Average**	**Standard Deviation**	
NFκB chr16	0.02	0.23	48%
TP53 chr5	0.009	0.12	80%
	Global average diploid controls	**Global standard deviation diploid controls**	
	0.012	0.14	75%

Table 3.1: Average log intensity ratios of the reference sample Rh16 for all the dosage positive control and diploid control probes included in the array-CGH. For each gene considered, the average log intensity ratio, the standard deviation and the percentage of probes clustering between -0.1 and 0.1 are indicated. Global averages and standard deviations are indicated for the dosage positive controls and the diploid controls.

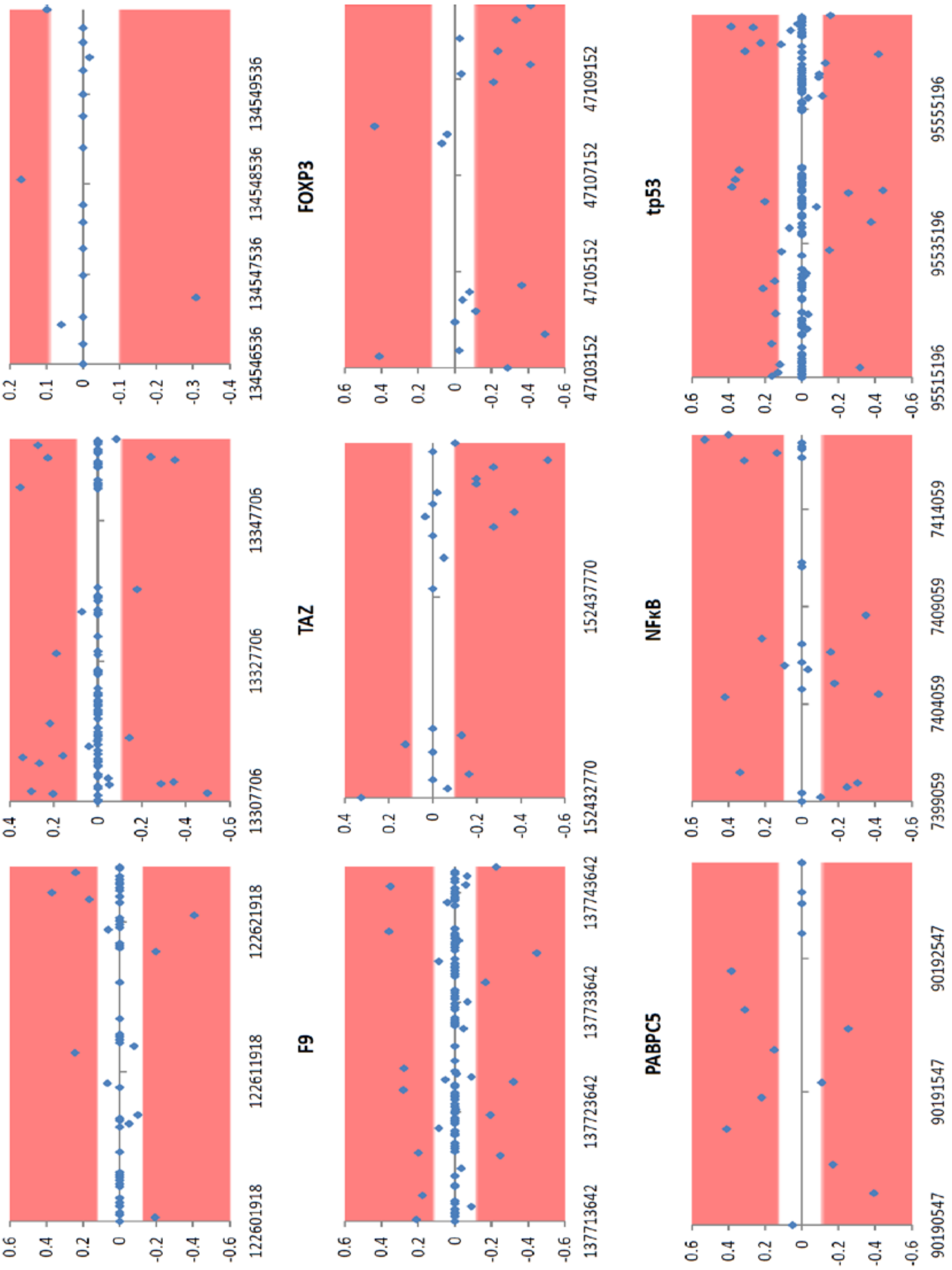


Figure 3.4: Distribution of the dosage positive controls and the diploid controls (*NFkB* and *TP53*) for the male reference individual Rh16. The X axis shows the chromosomal coordinates for each gene considered, the Y axis indicates the log intensity ratio between two detections of the reference sample Rh16.

3.2.2. Diploid control normalisation

The average Rh16 background noise values reported for *NFkB* and *TP53* were subtracted from the correspondent log intensity ratio for all the samples in analysis. The results are plotted in **Figure 3.5** and **Figure 3.6** and show consistent clustering around 0, indicative of equal copy number between the test and the reference sample. Single coordinate probes deviating from 0 in all the samples, in absence of flanking probes variation, are likely to reflect probe non-uniqueness and not copy number variation.

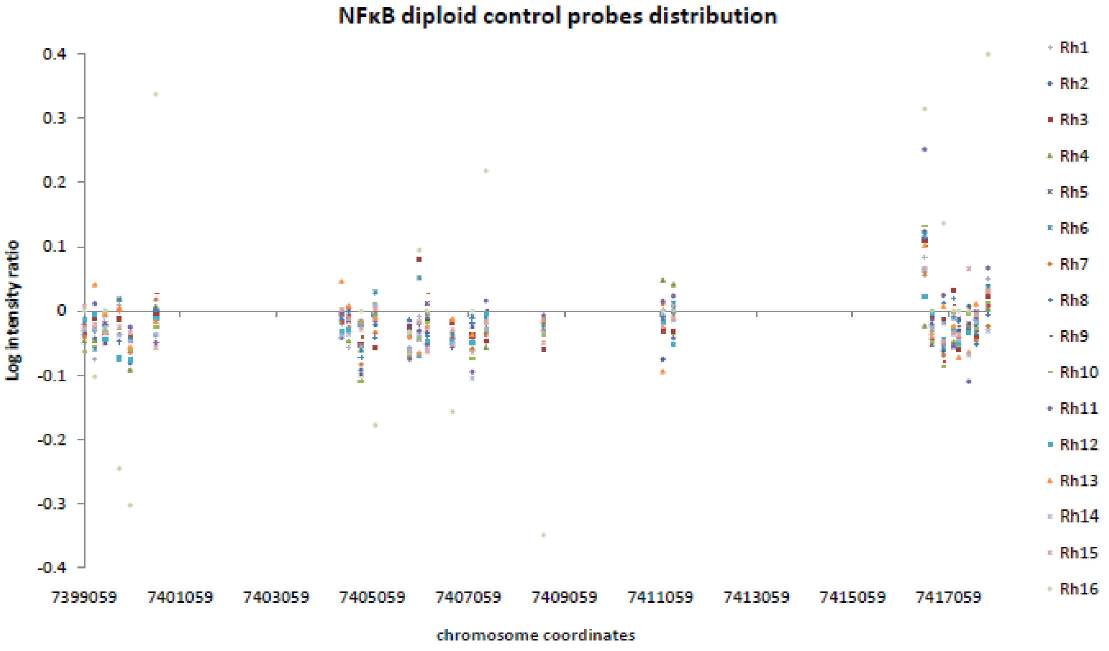


Figure 3.5: *NFκB* diploid control probe distribution for rhesus macaque samples 1 to 16.

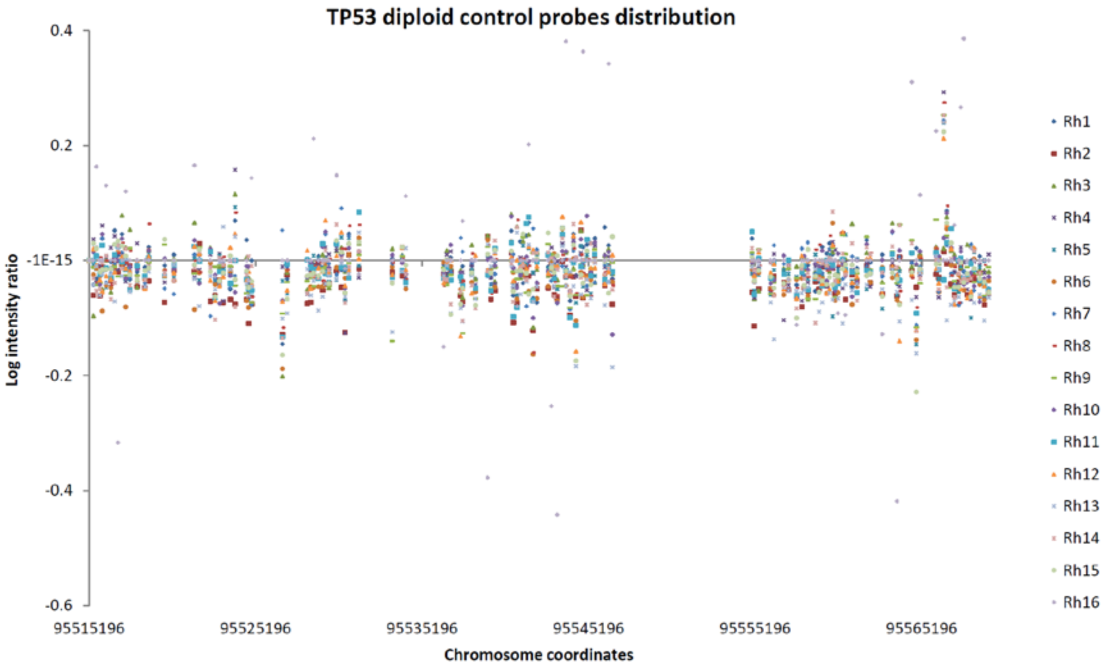


Figure 3.6: *TP53* diploid control probe distribution for rhesus macaque samples 1 to 16.

3.2.3 PCR validation of rhesus macaque samples sex

A control PCR was performed on all the samples in the analysis, aimed at discriminating males from females for a correct attribution of the dosage positive control probes listed in **Table 3.1**. A standard procedure to amplify an intron of the X-linked zinc finger protein gene (*ZFX/ZFY*) was selected, based on a publication of Wilson JF *et al.* 1998 [203]. The rationale behind this technique is based on the fact that *ZFX* and *ZFY* are two homologous genes mapping to the non-pseudoautosomal region of the sex chromosomes, with *ZFX* presenting a 422bp *Alu* insertion into its last intron, dating before the old world monkey-new world monkey split [206]. Primers amplifying this intron (Rik8 and Jim9, see paragraph 2.4.9) generate a 1151bp fragment for *ZFX* and a shorter 729bp fragment on *ZFY*, distinguishable on gel electrophoresis. PCR amplification for the rhesus macaque samples Rh1 to Rh16 plus one male and one female human positive control has been performed following the protocol described in 2.4.9. Results of the PCR amplification are shown in **Figure 3.7**.

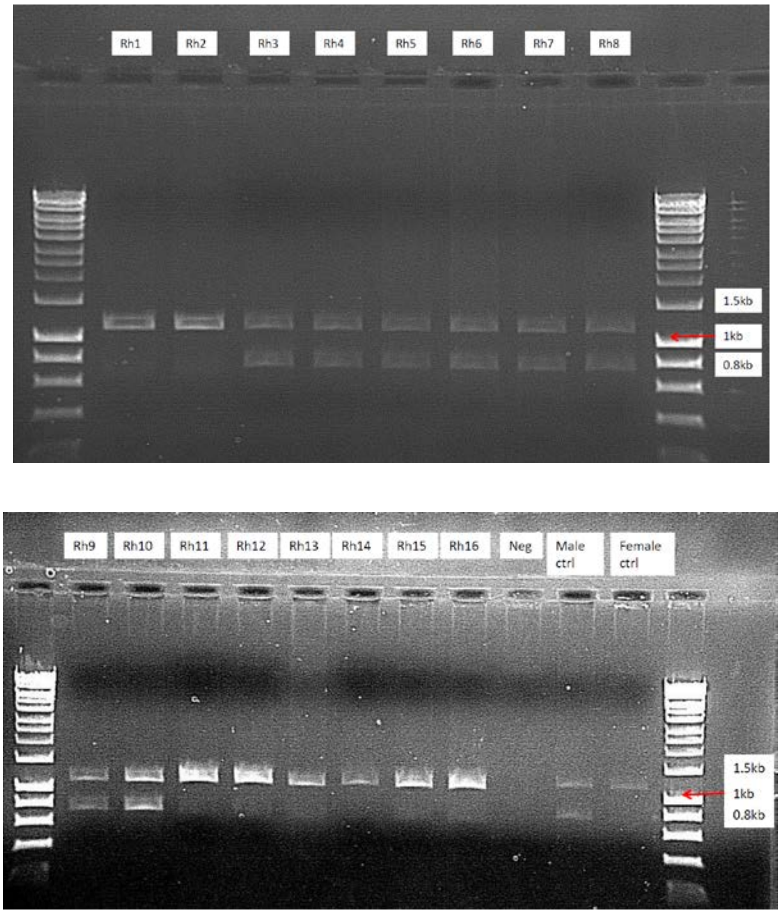


Figure 3.7: 0.8% agarose gel showing PCR amplified ZFX/ZFY fragments. Samples analysed are in order: Rh1 to Rh16, negative control, male human control, female human control. Males present one ZFX band of 1151bp and one ZFY band of 729bp; females present one 1151bp band from the two X chromosome copies. 441ng of Hyperladder I are used for band sizing.

The results of the control PCR allowed the separation of the macaque cohort used for the array CGH in a male and a female population (**Table 3.2**), to be used as internal calibrator of the experiment accuracy.

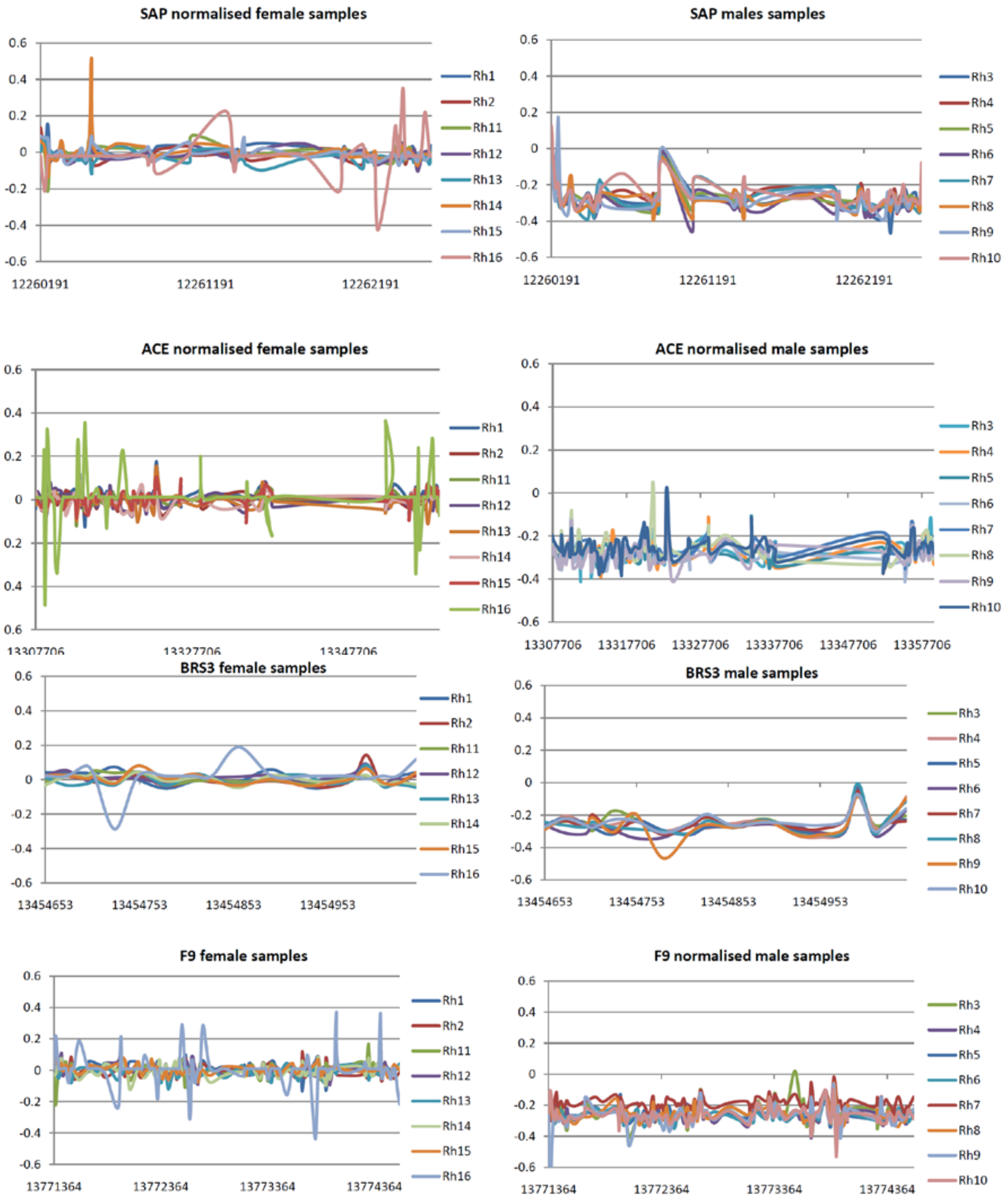
Sample name	Sex
RH1	Female
RH2	Female
RH3	Male
RH4	Male
RH5	Male
RH6	Male
RH7	Male
RH8	Male
RH9	Male
RH10	Male
RH11	Female
RH12	Female
RH13	Female
RH14	Female
RH15	Female
RH16	Female

T**able 3.2:** Sex attribution of the rhesus macaque samples used for the array-CGH experiment, based on the results of the ZFX/ZFY control PCR.

3.2.3 Dosage positive controls normalisation

The average Rh16 background noise values reported for each dosage positive control gene were subtracted from the correspondent log intensity ratio for all the samples in the analysis. The resulting values were then used as hybridisation quality control for each sample in the analysis: as the reference sample was a female, the logarithm of female/female ratios (1) was expected to be 0, whereas the logarithm of male/female ratios (0.5) was expected to cluster around -0.3.

Normalised log intensity ratio values were analysed separately for each dosage positive control area, in order to monitor possible unreported variations that may bias the dataset, and values for each individual were plotted on two separate graphs, accordingly to the sex of the samples (**Figure 3.8**). For each gene considered, two different patterns were expected: probes from female samples clustering around 0 and probes from male samples clustering around -0.3.


SAP normalised female samples
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
12260191
12261191
12262191
Rh1
Rh2
Rh11
Rh12
Rh13
Rh14
Rh15
Rh16
SAP males samples
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
12260191
12261191
12262191
Rh3
Rh4
Rh5
Rh6
Rh7
Rh8
Rh9
Rh10
ACE normalised female samples
0.6
0.4
0.2
0
0.2
0.4
0.6
13307706
13327706
13347706
Rh1
Rh2
Rh11
Rh12
Rh13
Rh14
Rh15
Rh16
ACE normalised male samples
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
13307706
13317706
13327706
13337706
13347706
13357706
Rh3
Rh4
Rh5
Rh6
Rh7
Rh8
Rh9
Rh10
BRS3 female samples
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
13454653
13454753
13454853
13454953
Rh1
Rh2
Rh11
Rh12
Rh13
Rh14
Rh15
Rh16
BRS3 male samples
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
13454653
13454753
13454853
13454953
Rh3
Rh4
Rh5
Rh6
Rh7
Rh8
Rh9
Rh10
F9 female samples
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
13771364
13772364
13773364
13774364
Rh1
Rh2
Rh11
Rh12
Rh13
Rh14
Rh15
Rh16
F9 normalised male samples
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
13771364
13772364
13773364
13774364
Rh3
Rh4
Rh5
Rh6
Rh7
Rh8
Rh9
Rh10

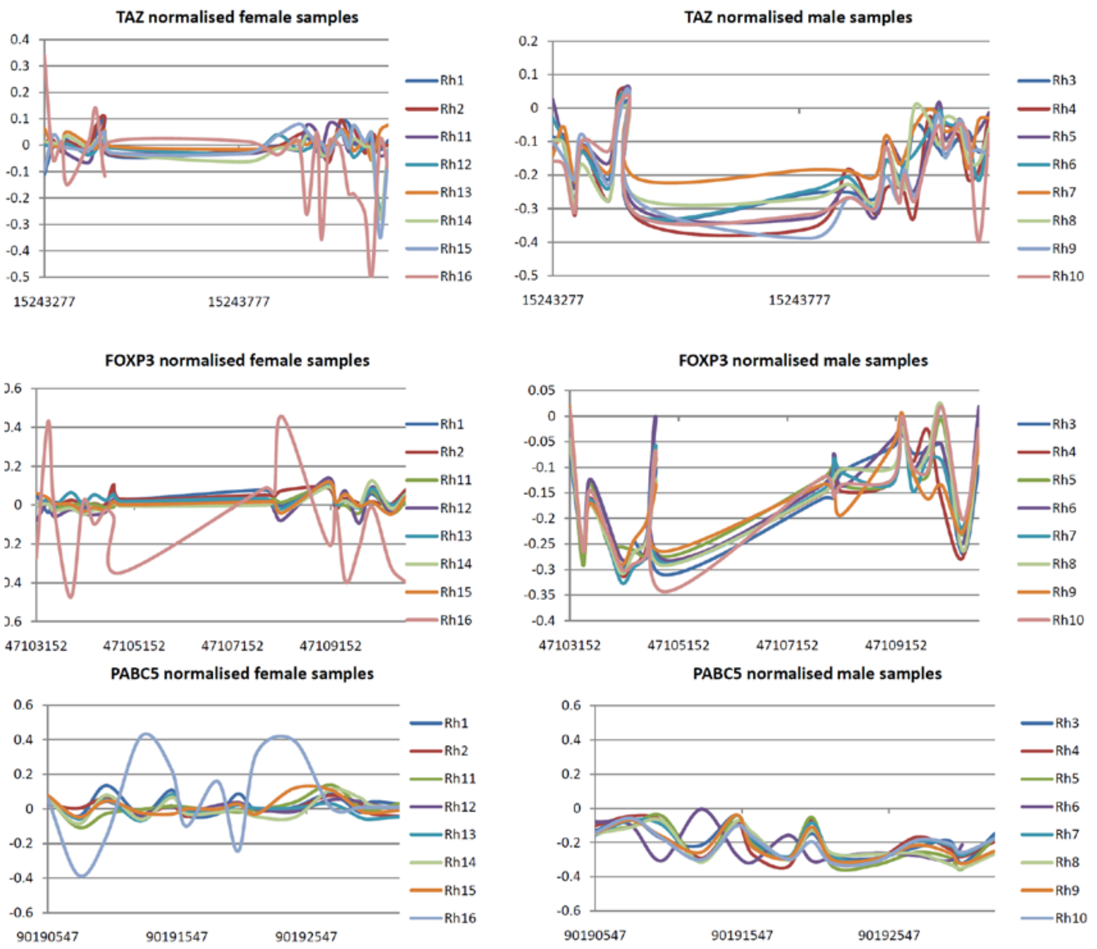


Figure 3.8: Dosage positive controls probe distribution for females (left column) and males (right column). The X axis shows the chromosomal coordinates for each gene considered, the Y axis indicates the log intensity ratio between the test sample in analysis and the reference sample Rh16 (female).

A log intensity distribution around -0.25 was consistent with a male/female ratio. As shown in **Figure 3.8**, it was possible to observe for each area considered a clear difference in the distribution pattern between males and females, with the former clustering around -0.25 and the latter around 0. Increases in probe coverage, as for SAP, ACE and F9, corresponded to more homogeneous probe clusterings, whereas small probe numbers were associated with increased signal fluctuation, despite being still significant (**Table 3.2**), as shown by TAZ, BRS3, FOXP3 and PACB5. Hence for further experiments it would be advisable to include controls from fewer areas, with higher probe coverage, instead of analysing more genes with fewer probes.

The box plot presented in **Figure 3.9** was indicative of a good separation of the samples in two classes, with respect to their sex.

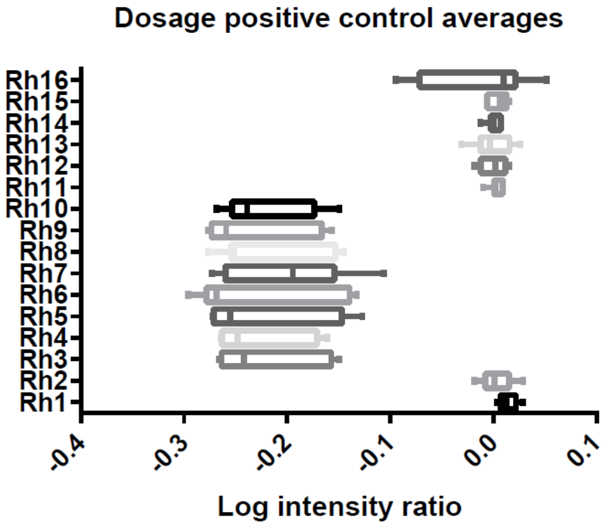


Figure 3.9: Box plot presenting the dispersion of the average ratio intensity values for all the dosage positive control areas in analysis.

In order to confirm the accuracy of the dosage positive controls data, a two-way ANOVA statistics has been performed per each area in analysis, separating the samples in two groups "females" and "males", accordingly to the results of the control PCR (**Figure 3.7**). A multiple comparisons correction performed with Sidak's multiple comparison test has been applied, to compare each probe mean for a given coordinate between the two groups (**Table 3.3**).

Dosage positive control	Row factor F	Column factor F	Row factor x column factor	P value	Total probes	Significant probes after Sidak's correction
SAP	F (51, 357) = 12.08	F (1, 7) = 1367	F (51, 357) = 7.928	P < 0.0001	51	93%
ACE	F (87, 609) = 6.097	F (1, 7) = 3638	F (87, 609) = 2.986	P < 0.0001	87	100%
BRS3	F (17, 119) = 18.12	F (1, 7) = 4030	F (17, 119) = 6.968	P < 0.0001	17	100%
F9	F (98, 686) = 4.396	F (1, 7) = 441.4	F (98, 686) = 3.744	P < 0.0001	98	99%
TAZ	F (21, 147) = 13.41	F (1, 7) = 211.0	F (21, 147) = 6.482	P < 0.0001	22	55%
FOXP3	F (18, 126) = 9.950	F (1, 7) = 95.37	F (18, 126) = 7.423	P < 0.0001	18	73%
PABC5	F (13, 91) = 8.059	F (1, 7) = 1036	F (13, 91) = 9.265	P < 0.0001	14	93%

Table 3.3: Summarising table for each two-way ANOVA test performed on the dosage positive control areas in analysis.

These results show strong concordance for all the probes in the analysis, indicative of reliable hybridisation efficiency.

3.3 β-defensin analysis

After signal normalisation and quality control checking of dosage positive controls and diploid controls, it was possible to proceed with the analysis of the β-defensin area, given the confirmation of good DNA quality and good hybridisation efficiency.

3.3.1 Data normalisation and filtering

The first phase of the analysis focussed on ascertaining the hybridisation efficiency of the reference individual Rh16, for the 875 probes spanning the β-defensin area. Because of the low magnitude of log intensity values deriving from the self comparison of Rh16, standard deviations among replicates were higher than expected (0.15). Hence an extra filtering step was performed, excluding all the log intensity ratio values >0.1 and <-0.1. 28% of the probes were lost, but, as each position was analysed with 5 independent measurements, it was possible to maintain information for the 99% of the chromosomal coordinates considered, even filtering out some of the quintuplicates (**Figure 3.10**). After filtering, the standard deviation dropped to 0.025 and the average background noise was used as a normalisation value for all the other individuals in the analysis.

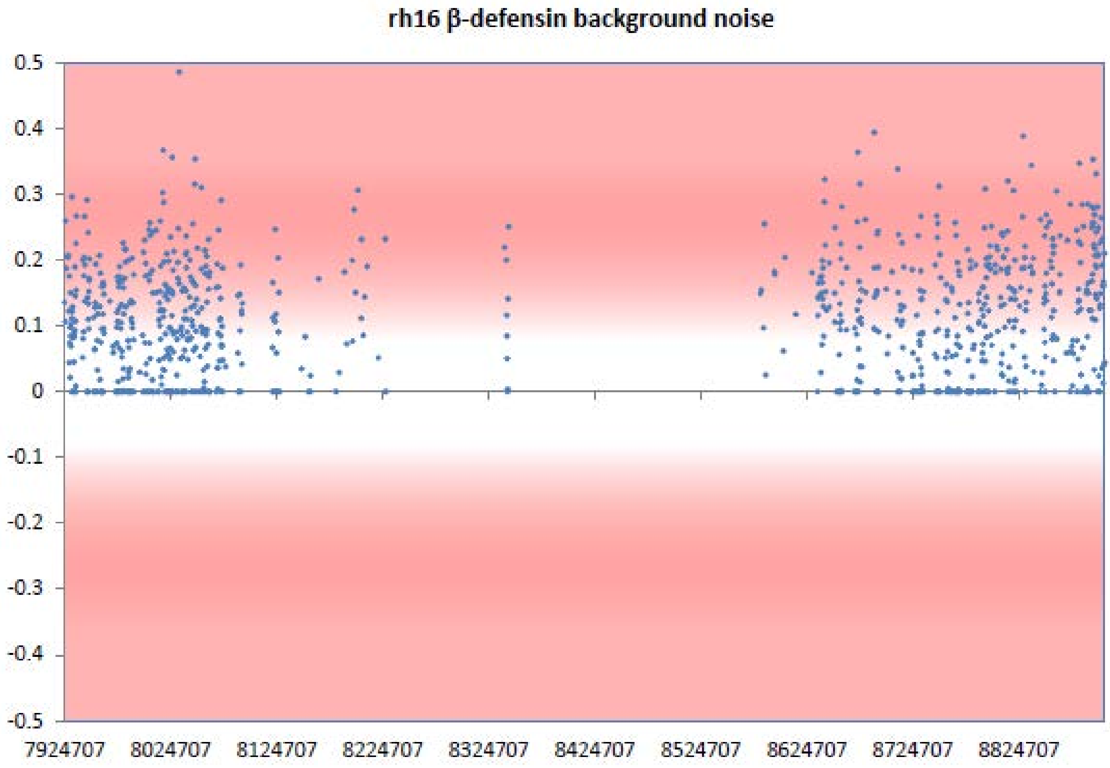


Figure 3.10: Log intensity ratio across the β-defensin region for the individual Rh16, reflecting the background noise level of the array-CGH experiment. Probes with log intensity ratio >0.1 and <-0.1.were excluded from the analysis. The red banding indicates the exclusion threshold. The X axis shows the chromosomal coordinates spanning the β-defensins area, the Y axis indicates the log intensity ratio generated through comparison of the reference sample Rh16 with itself.

After normalisation of all the test individuals, it was possible to proceed with the data filtering for Rh1 to Rh15. As each probe was repeated 5 times, to increase the measurement accuracy, for each chromosomal coordinate the average log intensity value was considered, if the standard deviation among each 5 replicates was <0.05. This extra filtering was performed to guarantee the concordance among replicates; **Table 3.4** summarises the effects of the probe filtering on the dataset.

Sample	St dev before filtering	St dev after filtering	Percentage of excluded probes	Number of excluded probes
Rh1	0.021	0.020	2%	17
Rh2	0.020	0.019	1%	9
Rh3	0.024	0.021	6.4%	56
Rh4	0.023	0.021	3.4%	30
Rh5	0.020	0.019	2.1%	18
Rh6	0.018	0.018	1.1%	10
Rh7	0.019	0.019	0.8%	7
Rh8	0.019	0.019	2.6%	23
Rh9	0.025	0.022	5.4%	47
Rh10	0.018	0.017	1.4%	12
Rh11	0.022	0.020	3%	26
Rh12	0.025	0.021	7%	61
Rh13	0.021	0.019	3.8%	33
Rh14	0.022	0.020	4.1%	36
Rh15	0.021	0.019	3.2%	28

Table 3.4: Effects of probe filtering on the probe distribution of samples Rh1-15. With a standard deviation cut off of 0.05 for replicates validation, between 0.8% and 7% of the chromosomal coordinates were excluded from the analysis.

These results demonstrate an overall strong concordance among replicates, with low average standard deviation before and after filtering and minimal chromosomal coordinates loss. **Figure 3.11** is the merged probe distribution for all 15 rhesus macaque samples in analysis.

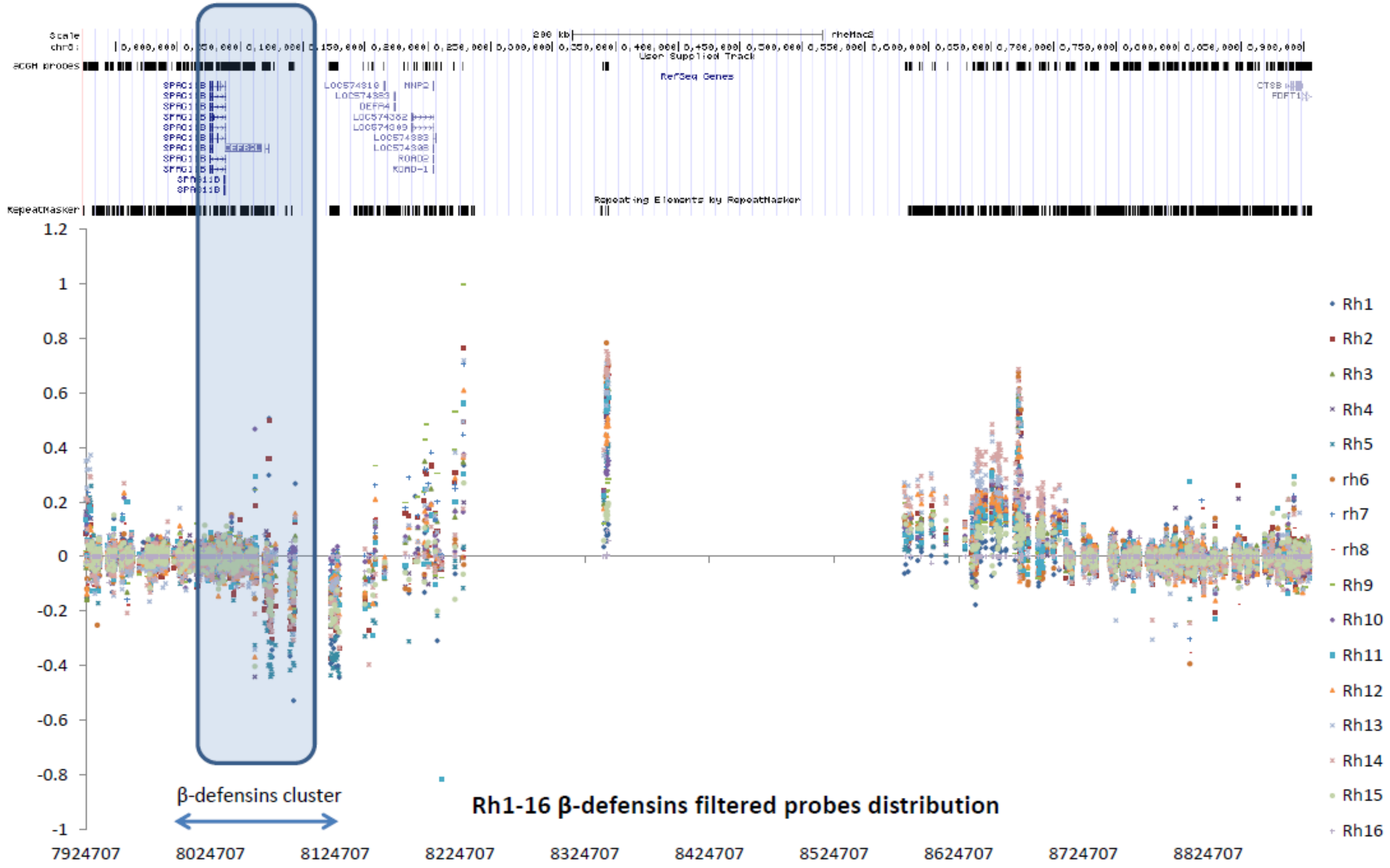


Figure 3.11: Probe distribution around the β-defensin area for all the rhesus macaque samples included in the array-CGH study. The upper panel shows the UCSC Genome Brower coordinates for chr8:7924707-8906039 with the annotated genes for the rhesus macaque genomic assembly of Jan2006. The blue rectangle indicates the location of the β-defensin cluster, highlighting the variation among the 16 samples**.**

3.3.2 Breakpoints analysis

Following data normalisation and filtering, the resulting data set was used with the aim of identifying possible breakpoints on chr8p23.1, to narrow down the copy number variation boundaries. The software BreakPtr was used to perform this analysis (See section 2.4.8). The first phase focussed on the generation of an appropriate hidden Markov model (HMM) to be used for the identification of unknown CNVs. To train the HMM, two different approaches were tried: the first used a large set of high-resolution array-CGH data generated on human chromosome 22 by Korbel *et al.* [202], with 186177 probes spanning large aberrations with known breakpoints; the second made use of an "artificial CNV area" specifically constructed for this analysis from the control probe results of my array CGH experiment. Upon generation of the hidden Markov model, for a given transition probability the software identifies the areas with no significant deviation from zero (state 0, or diploid state), the areas showing significant signal increase (State 1, or amplification state) and the areas showing significant signal decrease (State 2, or deletion state). For the β-defensin area under analysis the situation was more complex, as the presence of multiallelic copy number states were expected; hence in this case the 'diploid state' would correspond to a real diploid state for non-CNV areas and to a CNV region with the same copy number between test and reference samples. Analogously, 'deletion states' were expected to correspond to copy number decreases compared to the reference and 'amplification states' to copy number increases.

3.3.2.1 Bivariate hidden Markov model on validated large aberrations

For this approach, a positive control training file with 186177 probes was used, with 4 known breakpoints, representing a CNV area on human chromosome 22; the negative control was constituted by 186175 probes specific for diploid areas of the same chromosome. Based on these files training, 4 different hidden Markov matrices were built, applying different transition probabilities (**Table 3.5**). The matrix structure is presented in **Table 3.6**.

Transition probability deletion → deletion	Transition probability normal state → deletion	Transition probability amplification→deletion
Transition probability deletion →normal state	Transition probability normal state → normal state	Transition probability amplification→ normal state
Transition probability deletion → amplification	Transition probability normal state → amplification	Transition probability amplification → amplification

Table 3.5: Transition probability matrix structure.

	Deletion	**Normal state**	**Duplication**
First state probability	0.334	0.333	0.333
Transition matrix for transition probability of 0.1	0.8	0.1	0.1
	0.1	0.8	0.1
	0.1	0.1	0.8
Transition matrix for transition probability of 0.01	0.98	0.01	0.01
	0.01	0.98	0.01
	0.01	0.01	0.98
Transition matrix for transition probability of 0.001	0.998	0.001	0.001
	0.001	0.998	0.001
	0.001	0.001	0.998
Transition matrix for transition probability of 0.0001	0.9998	0.0001	0.0001
	0.0001	0.9998	0.0001
	0.0001	0.0001	0.9998
Transition matrix for transition probability of 0.00001	0.99998	0.00001	0.00001
	0.00001	0.99998	0.00001
	0.00001	0.00001	0.99998
State 0 Diploid state	Mean vector	0.00158	
	Covariance matrix	0.02745	
State 1 Amplification state	Mean vector	0.1757	
	Covariance matrix	0.0482	
State 2 Deletion state	Mean vector	-0.1824	
	Covariance matrix	0.0496	

Table3.6: Transition probability matrixes used to generate 4 different bivariate hidden Markov models, with, in order, transition probabilities of: 10^{-1}, 10^{-2}, 10^{-3} and 10^{-4}. Mean vector and covariance matrix values are constant because they are based on the same input datasets.

All these models have been run on the filtered β-defensins datasets generated for the 16 rhesus macaque samples used for the array-CGH, using the Finder-core tool of BreakPtr software. The breakpoints identified do not vary when the transition probabilities are changed, from the more relaxed filtering of 0.1 to the stricter one of 10^{-4}, suggesting the breakpoint calls are robust. The breakpoints are presented in **Table 3.7**. **Figure 3.12** shows the breakpoints identified for each individual plotted on the correspondent genomic area spanning the β-defensin area (chr8:7924707-8906098), interrogated with the array-CGH. From **Figure 3.12** it is also possible to notice the Repeat Masker track, showing the presence of repeated elements, and a Gap track, indicating the presence of extensive gaps in the rhesus macaque genomic assembly for the area in analysis.

Sample	Chromosome	Start	End	State	Probes number	Mean log intensity ratio
Rh1	chr8	7924707	8071452	0	345	0.002
	chr8	8071942	8156567	2	39	-0.252
	chr8	8156729	8906039	0	470	0.005
Rh2	chr8	7924707	8071452	0	347	-0.015
	chr8	8071942	8156567	2	42	-0.212
	chr8	8156729	8674538	1	101	0.227
	chr8	8674746	8906039	0	376	-0.005
Rh3	chr8	7924707	8210058	0	376	-0.015
	chr8	8220547	8674538	1	80	0.215
	chr8	8674746	8906039	0	363	0.008
Rh4	chr8	7924707	8227117	0	395	-0.009
	chr8	8227621	8674538	1	80	0.208
	chr8	8674746	8906039	0	370	-0.000
Rh5	chr8	7924707	8069678	0	339	-0.005
	chr8	8069931	8183690	2	51	-0.315
	chr8	8190871	8227621	0	14	-0.017
	chr8	8339894	8675141	1	85	0.170
	chr8	8675296	8906039	0	367	0.003
Rh6	chr8	7924707	8227621	0	407	-0.022
	chr8	8339894	8674538	1	81	0.220
	chr8	8674746	8906039	0	377	-0.007
Rh7	chr8	7924707	8156567	0	389	-0.004
	chr8	8156729	8343622	1	29	0.265
	chr8	8581056	8906039	0	450	0.020
Rh8	chr8	7924707	8227621	0	401	-0.015
	chr8	8339894	8674380	1	79	0.228
	chr8	8674746	8906039	0	372	0.004
Rh9	chr8	7924707	8148362	0	373	-0.013
	chr8	8156729	8604233	1	33	0.245
	chr8	8614432	8906039	0	421	0.016
Rh10	chr8	7924707	8220547	0	405	0.003
	chr8	8220702	8709919	1	116	0.187
	chr8	8710744	8906039	0	341	-0.010
Rh11	chr8	7924707	8206378	0	392	-0.012
	chr8	8210058	8674538	1	86	0.204
	chr8	8674746	8906039	0	370	-0.003
Rh12	chr8	7924707	8206378	0	378	-0.015
	chr8	8220702	8674746	1	81	0.227
	chr8	8674970	8906039	0	355	0.001
Rh13	chr8	7924707	7928392	1	6	0.328
	chr8	7928877	8220547	0	384	-0.028
	chr8	8220702	8674970	1	86	0.316
	chr8	8675141	8906039	0	365	-0.005
Rh14	chr8	7924707	8069931	0	332	-0.008
	chr8	8070117	8151591	2	43	-0.174
	chr8	8154787	8220702	0	18	-0.008
	chr8	8227117	8709919	1	112	0.320
	chr8	8710503	8906039	0	334	-0.016
Rh15	chr8	7924707	8906039	0	847	0.000

Table 3.7: List of breakpoint coordinates in the β-defensin area calculated on the log intensity ratios for the 15 rhesus macaque samples screened with the array-CGH. State 0 indicates a diploid state, or the same copy number compared to the reference; state 1 indicates an amplification state, or copy number increase compared to the reference Rh16; state 2 indicates a deletion state, or copy number decrease compared to the reference Rh16; the sixth column indicates the number of probes falling in the state described; the last column indicates the average log intensity ratio for each calculated state.

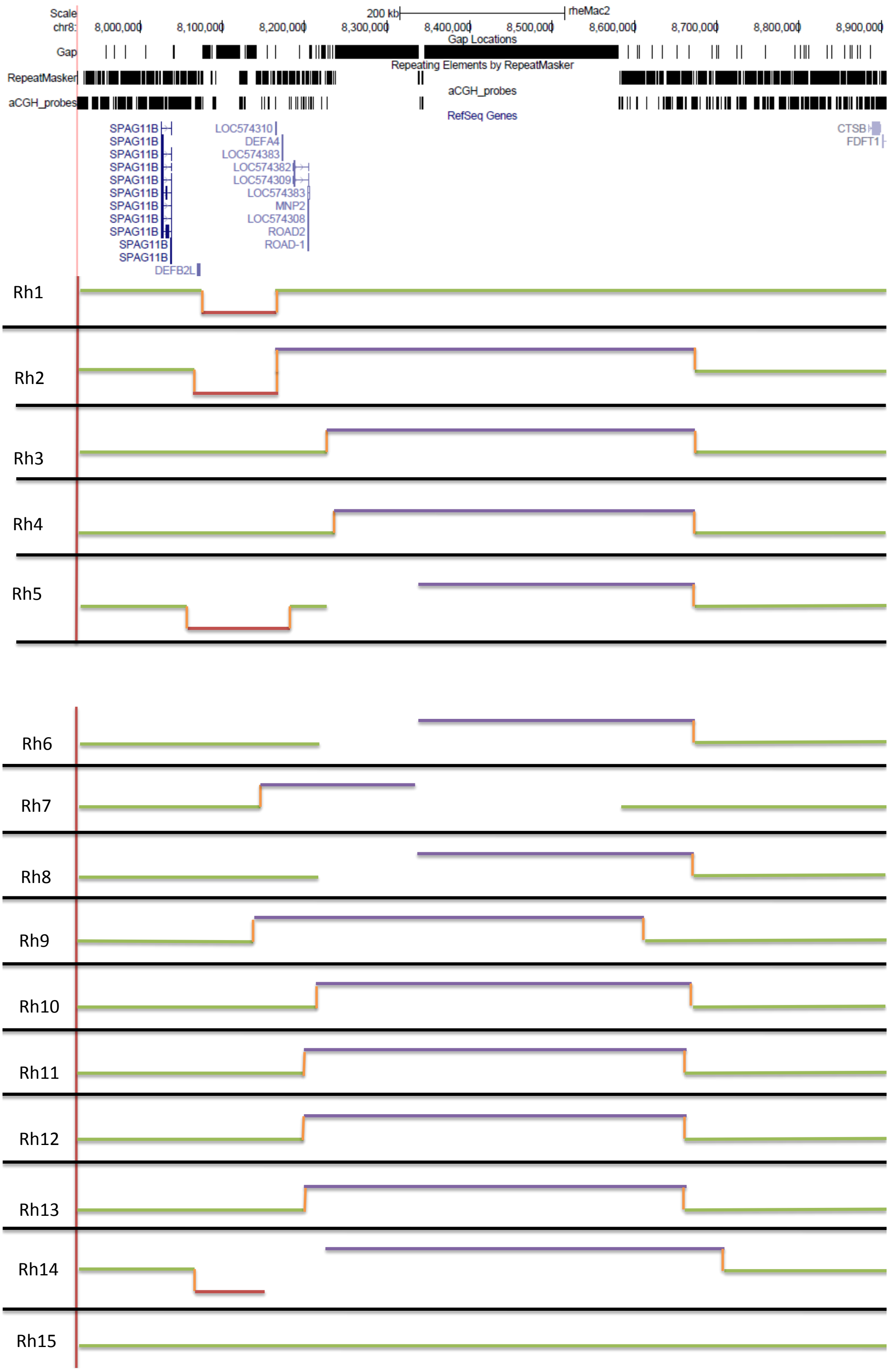


Figure 3.12: Breakpoint coordinates identified with the software BreakPtr, for the 15 samples in analysis. The green track indicates same CN compared with the reference; red indicates CN decrease and blue marks CN increase compared with the reference Rh16.

With this approach it was possible to identify a CNV boundary located exactly at the start of the *DEFB2L* gene for Rh1, Rh2, Rh5 and Rh14. The CNV identified has an approximate size of 81-84kb for the four samples (chr8: 8070117-8151591). This CNV size is extremely inaccurate, as the presence of gaps in the genomic assembly is likely to have impeded an accurate estimation of the true size of this genomic area. Nevertheless, as is clear from **Figure 3.13**, this genomic area is a low probe coverage region, with just 3 probes clusters in 84kb, for a total of 29 to 51 probes, due to the presence of gaps in the assembly and to repeated elements that impeded a tiling-path probe design.

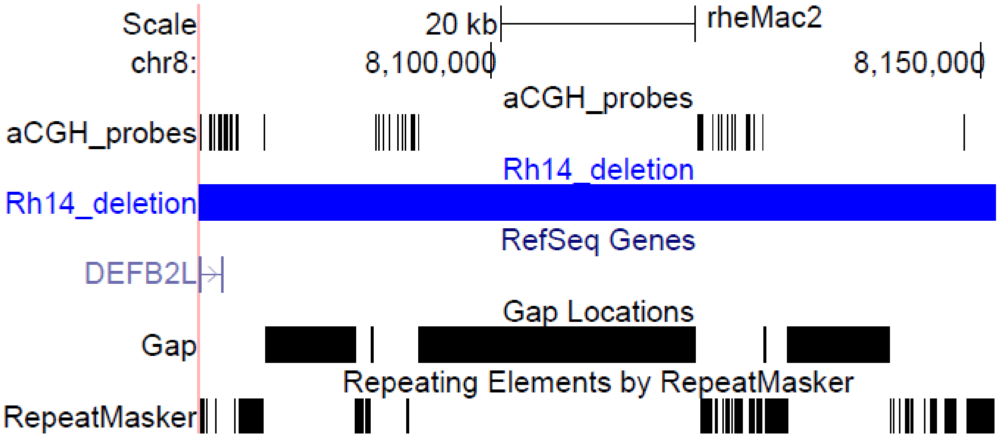


Figure 3.13: probe coverage for the CNV area identified through BreakPtr, for the Rh14 sample. *DEFB2L* gene is located at the distal edge of the variable area. The Repeat Masker track shows a high concentration of repeated elements and the Gap track covers 48.5kb out of 81.5kb (60% of the total area).

The telomeric area to the deletion highlighted, presented on the contrary a high probe coverage, as presented in **Figure 3.13**. It is hence expectable that this breakpoint, around chr8: 8070117, has been accurately identified, as confirmed by a consistent coordinate call for Rh1, Rh2, Rh5 and Rh14, likely to reflect a real transition between diploid and deletion state compared to the reference. Accordingly to these data, Rh1, Rh2, Rh5 and Rh14 may present a lower copy number compared to the reference; Rh15 and the reference Rh16 are expected to have the same copy number for all the area considered, whereas the other individuals may present the same copy number of the reference Rh16 for the *DEFB2L* gene but vary at the centromeric end of the region considered; in particular, Rh2, Rh7 and Rh9 show a copy number increase in correspondence with the α-defensin cluster or just distal to it (around chr8:8156729). Rh5, Rh6, Rh8, Rh10, Rh11, Rh12, Rh13 and Rh14 on the contrary, show a copy number increase in the area centromeric to the α-defensin cluster, without encompassing it. This last result may reflect a bias in the probe coverage, as this variable signal has been generated by just 10 probes clustered in 4kb, between two sequence gaps of 28.4kb and 10.4kb, which presented a huge inter-sample variation. This signal may be due to probe non-specificity, with likely cross-hybridisation to repeated elements not annotated in the rhesus macaque genome. If the effects of this dramatic signal intensity would have directly interfered with the detection of real copy number signals with lower intensities around the area, the breakpoint analysis of the probes intensity signals distal to the sequence gap would have generated a different breakpoints pattern. This hypothesis has been tested on all the samples but did not result in any variation in the breakpoints distribution.

Intriguingly, the consistency of the amplification coordinates for samples Rh2, Rh7 and Rh9, for the genomic area spanning the α-defensins cluster, would suggest the presence of a real amplification, at least until the sequence gap present at chr8:8237937-8337936.

From these data it could be possible to hypothesise the presence of two distinct CNVs: one encompassing the *DEFB2L* gene, in agreement with the distal breakpoint identified by Lee *et al* 2008, and a centromeric CNV unit in proximity of the α-defensin cluster, as indicated in **Figure 3.14**. For both CNV units, the characterisation of precise breakpoints was hampered by the presence of large gaps in the rhesus macaque sequence assembly with no probe coverage. Moreover, it should be noticed that the array CGH data generated in this study did not possess enough resolution power to rule out the hypothesis of the presence of more complex rearrangements in the area in analysis

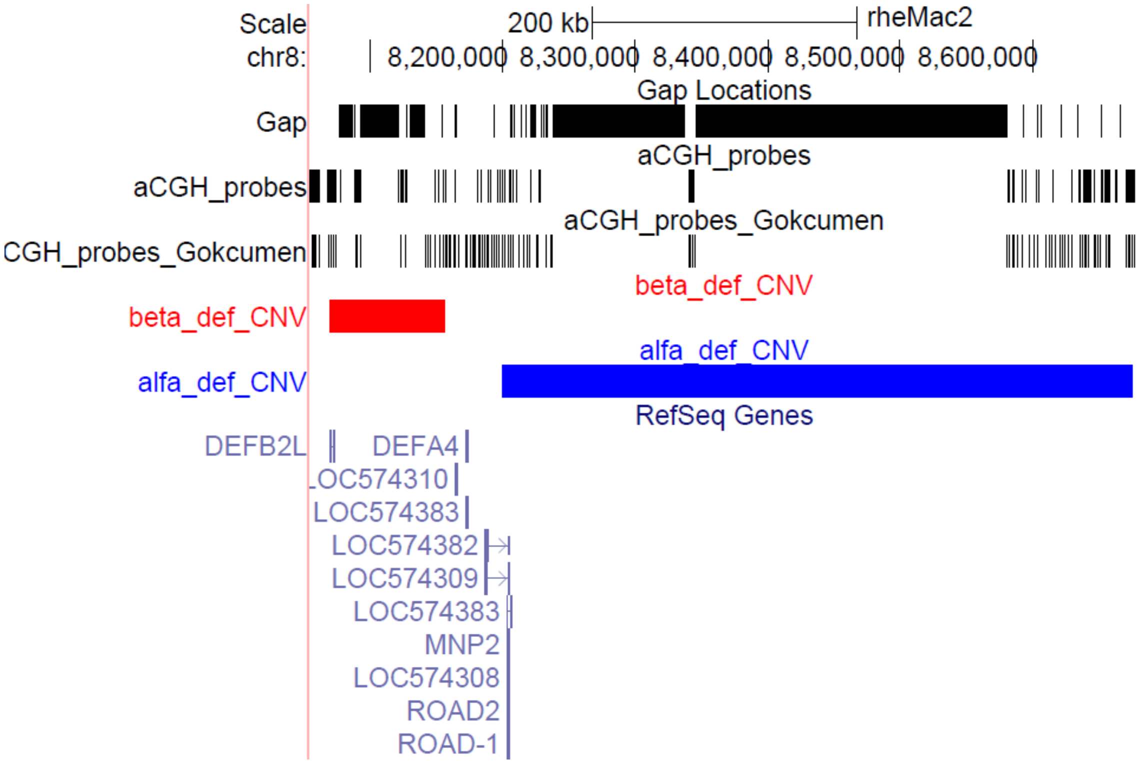


Figure 3.14: Hypothesised location of two distinct CNV units: one encompassing the *DEFB2L* gene (red track) and one in proximity of the α-defensin cluster (blue track). From the top, it is also indicated: the location of gaps in the rhesus macaque assembly; the position of the array CGH probes used in this study; the position of the array CGH probes used in the study of Gokcumen *et al* 2011 [81].

Nevertheless, these results have been generated using a model trained on extremely large human datasets, covering the entire chromosome 22, with large aberration spanning several megabases, whereas the dataset generated with the high-density array-CGH covers a dramatically smaller area. This may have affected the detection of small CNVs, or the call of amplifications and deletions in case of little copy number differences between the test and the reference. For these reasons, an alternative approach has been designed and tested, as described in the next chapters.

3.3.2.2 Bivariate hidden Markov model on rhesus macaque validated controls

Ideally, a more unbiased way to generate breakpoint information would have come from the use of training data set from the same array-CGH used, generated on rhesus macaque genomic areas for which previously known different size copy number and large diploid control regions were available. For this study, this information was not available, and an artificial alternative was developed. For the generation of a bivariate hidden Markov models through BreakPtr, three input files were needed for the training phase:

- a diploid region
- a region with known aberrations
- a file with known breakpoint coordinates and the indication of amplification or deletion state referred to the known aberrations file.

With this aim, three artificial input files were generated, combining the diploid controls and the dosage positive controls data from all the 16 rhesus macaque samples. The diploid region was generated merging all the data coming from *p53* and *NFκB* diploid controls for the 16 samples analysed, plus *SAP*, *ACE*, *BRS3* and *F9* dosage positive controls data from the females samples, for a total of 4393 control probes; progressive chromosomal coordinates from one BreakPtr template were attributed to each probe, covering over 35 Mb. The aberration file was created combining dosage positive controls data from the male samples, flanked by diploid control probes to mimic non-CNV areas. Amplification states were emulated by inverting the sign of the dosage positive control probes from the male samples, to simulate 3 copies values clustering around +0.25 (**Figure 3.15**). Progressive chromosomal coordinates taken from one BreakPtr template were applied to the probe series, to span 480kb, for a total of 3916 probes, with 4 declared deletion states and 4 amplification states, that were then used as breakpoint coordinates for the third file needed for the software training.

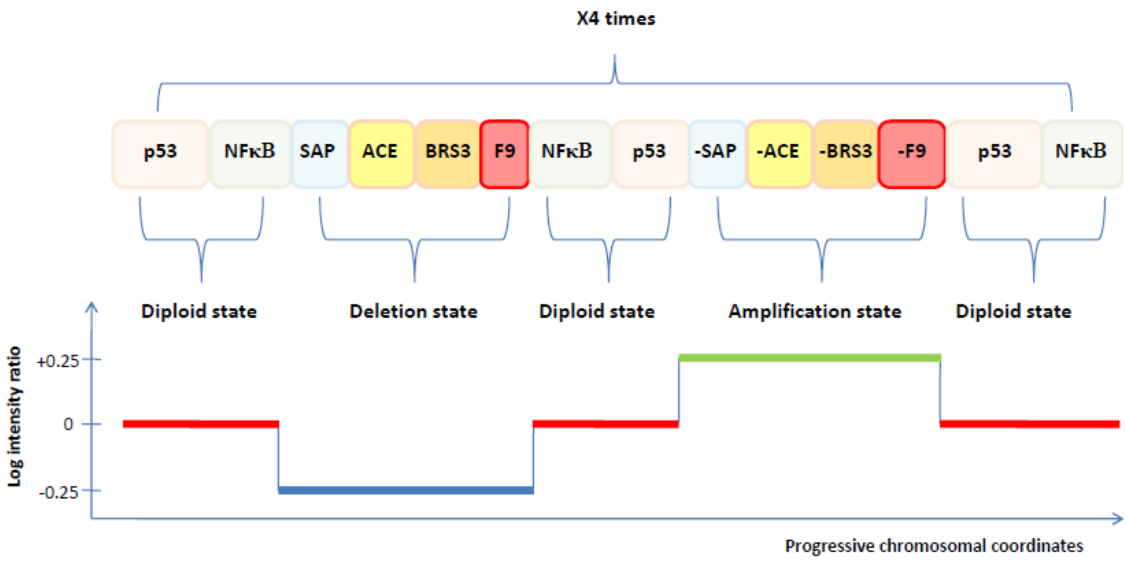


Figure 3.15: Scheme of the artificial genomic area of 480kb generated from the male samples Rh3, Rh4, Rh5 and Rh6, using a total of 3916 probes. Log intensity ratio values from *p53* and *NFκB* are used as 0, to identify diploid states (red line); log intensity ratios from dosage positive controls genes cluster at -0.25, to identify deletion states; the same log intensity ratios from dosage positive controls, with inverted sign, have been used to call amplification states, at +0.25.

This approach did not generate a suitable bivariate hidden Markov model for breakpoints identification, as the BreakPtr Trainer Tool was not able to compute any mean vector and covariance matrix to build the model, based on the dataset provided. This impeded the transition state calculations, and hence this approach was discarded. There are two concurrent explanation for the failure of this approach: the first regards the size of the CNV dataset used for the training; Korbel JO *et al.* utilised more than 180000 probes to generate a suitable hidden Markov model, used to identify 232 putative CNVs, highlighting 464 breakpoints flanking 121 duplications and 111 deletions; their median CNV size was 15kb and they reported a mean CNV size of 85kb. Conversely, the array CGH performed for this project did not have any *a priori* CNV information generated with similar high-density probe coverage. For the same reason, data information coming from lower-density arrays (as performed in Gokcumen O *et al.*) have not been used as training reference, as the differences in probe distribution may have generated an excessive bias in the breakpoint calculation algorithm. Secondly, it is possible that small variations in copy number, as 2 to 1 deletions or 2 to 3 duplications, were too subtle to be detectable with the BreakPtr software and should have not been used as template for the model training. Moreover, differently from the large aberration dataset provided with the software, the customised aberration file here used as alternative training approach was based on smaller CNV areas, as the final aim was the detection of small sized CNVs that could have been excluded with any larger scale training. In particular, the main goal would have been a more precise localisation of the centromeric breakpoint for the deletion encompassing the *DEFB2L* gene, to evaluate the presence of finer state transitions in the area distal to the gaps in the sequence assembly. Moreover, this approach is useful to estimate dosage losses or gains, indicative of CNV presence in the area in analysis, but it is not suitable for accurate copy number calls. In order to fulfil this aim, different approaches have been subsequently tried, as it will be presented in the next sections.

3.4 Chapter summary

This chapter describes the analysis performed on high-density array CGH data obtained for 15 non-related rhesus macaque DNA samples, against a rhesus macaque reference (Rh16), for a total of 8 males and 8 females. 155 diploid controls and 312 dosage positive controls designed on the X chromosome have been included in the array design per each individual, to be used as quality control of the hybridisation efficiency.

The genomic area most intensively interrogated with this customised array has been the β-defensin cluster, with ±200kb of flanking regions (chr8:7924707-8906098), with the aim of obtaining information on the presence and extension of potential multiallelic CNVs falling in this region. A breakpoint analysis has been performed using the software BreakPtr, to indentify the CNV boundaries, based on the generation of a hidden Markov model trained on a high-density array-CGH human dataset. Our results confirm the presence of a breakpoint in correspondence with the *DEFB2L* gene, for 4 individuals out of 15 (Rh1, Rh2, Rh5, Rh14). The position of the proximal breakpoint for this putative CNV remains unclear, for the lack of sequence information in the *Macaca mulatta* assembly in this complex region. Nevertheless, 3 individuals out of 15 (Rh2, Rh7, Rh9) also showed signs of copy number gain in correspondence with the α-defensin cluster, located centromeric to the β-defensin locus.

An alternative approach has been tried, with the aim of identifying CNVs of smaller size and detecting finer variations in dosage gains or losses, based on the creation of an 'artificial genomic region' built with diploid and dosage positive controls from our rhesus macaque HD array-CGH. This approach demonstrated an insufficient resolution power and was abandoned, in favour of the previous HMM. These data highlighted the presence of a multiallelic CNV and constituted the basis for the development of dedicated assays for a more accurate CNV typing.

DEFB2L copy number typing with PCR-based assays

Following the results of the breakpoint analysis on the β-defensins cluster, highlighting a signal of copy number variation in at least 4 individuals out of 16 (Rh1, Rh2, Rh5 and Rh14), the following step was the development of suitable PCR based methods to achieve an accurate *DEFB2L* copy number typing. With this aim, four different approaches were used: InDel assays, Paralogue Ratio Test (PRT), a microsatellite assay and digital droplet PCR.

4.1 InDel assays

These sets of assays are based on the identification of short insertion/deletion variants among different copies in the CNV area in analysis (**Figure 4.1**), performed through sequence alignment of the UCSC genome browser rhesus macaque reference Mmul_051212 against the WGS sequences generated from the same *Macaca mulatta* individual, available on the BLAST database. With this approach, short InDels identified through alignment of the same individual could represent copy specific variants, helpful for copy number attribution. Assuming *DEFB2L* gene to be part of the CNV area, a total of 100kb upstream and 100kb downstream this gene were aligned and manually screened, covering 219kb on chr8:7960000-8179391. In order to separate sequencing artefacts from real variant, each position correspondent to small insertion or deletion of 2-3bp was visualised with the BLAST tool 'Trace Archive', to check the electropherogram quality of the shotgun sequence for each putative InDel. In each case of too low or not clear sequencing signal, the InDel was not validated. After screening, 5 possible InDel candidates were selected (**Figure 4.2**). Each InDel assay was named after the nearer orthologous β-defensin gene manually mapped in the area and, because of the high degree of sequence homology for β-defensin orthologues, likely to be present in the rhesus macaque genome, despite the lack of gene annotations for all genes but *DEFB2L* (**Figure 4.3**).

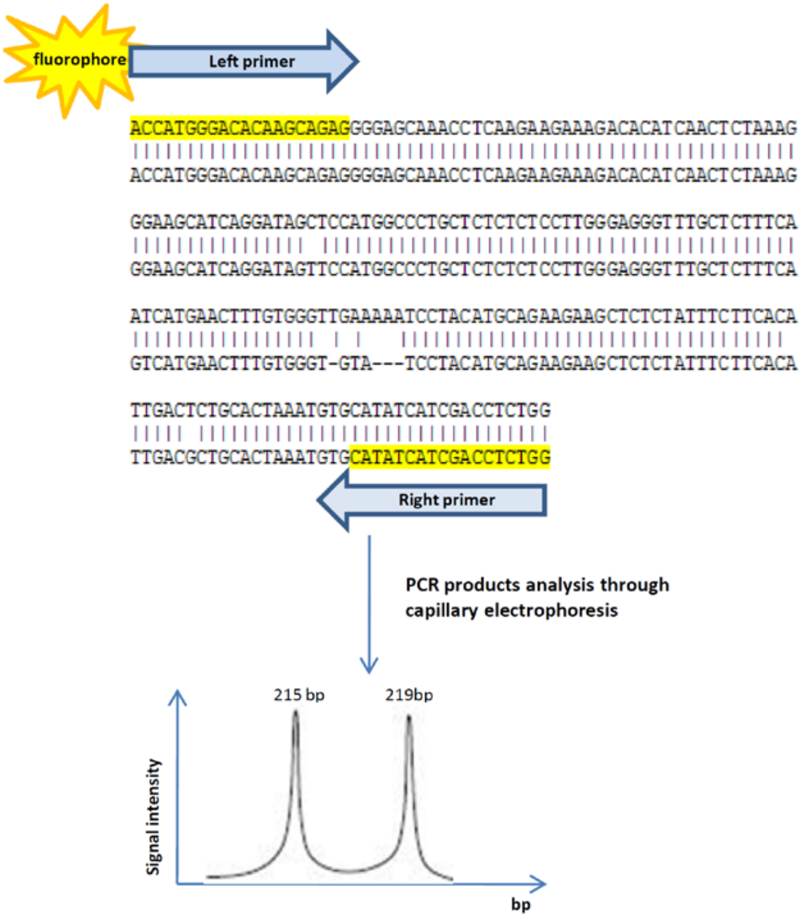
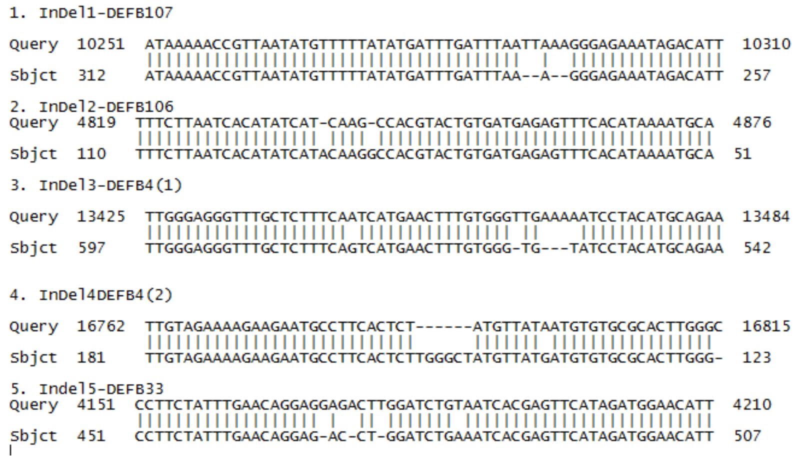

fluorophore
Left primer
ACCATGGGACACAAGCAGAGGGGAGCAAACCTCAAGAAGAAAGACACATCAACTCTAAAG
ACCATGGGACACAAGCAGAGGGGAGCAAACCTCAAGAAGAAAGACACATCAACTCTAAAG
GGAAGCATCAGGATAGCTCCATGGCCCTGCTCTCTCTCCTTGGGAGGGTTTGCTCTTTCA
GGAAGCATCAGGATAGTTCCATGGCCCTGCTCTCTCTCCTTGGGAGGGTTTGCTCTTTCA
ATCATGAACTTTGTGGGTTGAAAAATCCTACATGCAGAAGAAGCTCTCTATTTCTTCACA
GTCATGAACTTTGTGGGT-GTA---TCCTACATGCAGAAGAAGCTCTCTATTTCTTCACA
TTGACTCTGCACTAAATGTGCATATCATCGACCTCTGG
TTGACGCTGCACTAAATGTGCATATCATCGACCTCTGG
Right primer
PCR products analysis through capillary electrophoresis
215 bp
219bp
Signal intensity
bp

Figure 4.1: Scheme presenting the rationale of InDel assays: after the identification of suitable indel positions, through DNA sequence alignment of the UCSC genome browser rhesus macaque reference Mmul_051212 against the WGS sequences generated from the same *Macaca mulatta* individual, available on the BLAST database. Primers specific for the amplification of the area immediately spanning the indel are designed, with the left primer conjugated to a fluorophore. Following PCR amplification, PCR products are size-based separated through capillary electrophoresis.



```
1. InDel1-DEFB107

Query  10251  ATAAAAACCGTTAATATGTTTTTATATGATTTGATTTAATTAAAGGGAGAAATAGACATT  10310
              |||||||||||||||||||||||||||||||||||||||  |  ||||||||||||||||
Sbjct  312    ATAAAAACCGTTAATATGTTTTTATATGATTTGATTTAA--A--GGGAGAAATAGACATT  257

2. InDel2-DEFB106
Query  4819  TTTCTTAATCACATATCAT-CAAG-CCACGTACTGTGATGAGAGTTTCACATAAAATGCA  4876
             ||||||||||||||||||| |||| |||||||||||||||||||||||||||||||||||
Sbjct  110   TTTCTTAATCACATATCATACAAGGCCACGTACTGTGATGAGAGTTTCACATAAAATGCA  51

3. InDel3-DEFB4(1)

Query  13425  TTGGGAGGGTTTGCTCTTTCAATCATGAACTTTGTGGGTTGAAAAATCCTACATGCAGAA  13484
              ||||||||||||||||||||| |||||||||||||||| ||    |||||||||||||||
Sbjct  597    TTGGGAGGGTTTGCTCTTTCAGTCATGAACTTTGTGGG-TG---TATCCTACATGCAGAA  542


4. InDel4DEFB4(2)

Query  16762  TTGTAGAAAAGAAGAATGCCTTCACTCT------ATGTTATAATGTGTGCGCACTTGGGC  16815
              ||||||||||||||||||||||||||||      ||||||| ||||||||||||||||| 
Sbjct  181    TTGTAGAAAAGAAGAATGCCTTCACTCTTGGGCTATGTTATGATGTGTGCGCACTTGGG-  123

5. Indel5-DEFB33
Query  4151  CCTTCTATTTGAACAGGAGGAGACTTGGATCTGTAATCACGAGTTCATAGATGGAACATT  4210
             ||||||||||||||||||| |  || ||||||| ||||||||||||||||||||||||||
Sbjct  451   CCTTCTATTTGAACAGGAG-AC-CT-GGATCTGAAATCACGAGTTCATAGATGGAACATT  507
|
```

Figure 4.2: InDel position for each of the InDel candidates selected.

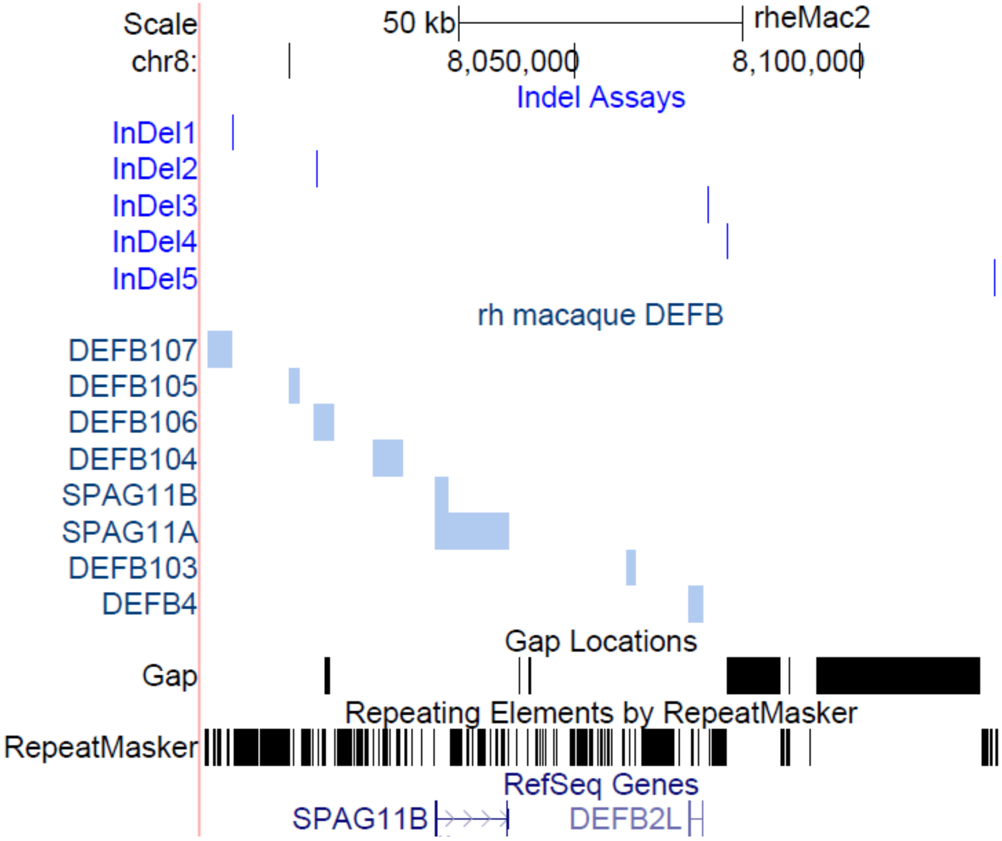


Figure 4.3: Localisation of the InDel assays on UCSC Genome Browser. From the top to the bottom, the tracks included describe: in dark blue the localisation of all the InDel assays, named from 1 to 5; in light blue are shown manually annotated rhesus macaque defensins genes.

InDel Assay	**Genomic coordinates**	**PCR product size (bp)**	**Fluorophore used**
DEFB107	Chr8:7989992-799031	320	6-FAM
DEFB106	Chr8:8004717-8004900	184	6-FAM
DEFB4(1)	Chr8:8073228-8073446	219	HEX
DEFB4(2)	Chr8:8076629-8076778	150	HEX
DEFB33	Chr8:8123473-8123647	175	6-FAM

Table 4.1: Genomic coordinates for the InDel assays, expected PCR product sizes and fluorophores used. As reference, *DEFB2L* spans chr8:8070118-8072503

Each assay has been optimised with a PCR gradient using non-fluorescent primers, as described in Section 2.3.1, on two control rhesus macaque samples. When specific amplification was achieved, it was possible to proceed with the setting of preparatory PCR reactions for capillary electrophoresis, using dedicated fluorescent primers.

Each assay was performed in duplicate on all the 16 rhesus macaque samples used for the array-CGH, with at least three independent PCR amplifications on separate plates. In concordance with the array results (Chapter 3), DEFB107-InDel1 and DEFB106-InDel2 always presented one single peak, consistent with their position outside the putative CNV area. The indels observed at these position highlighted through sequence alignment may hence be ascribed to sequence polymorphisms between the two copies of chromosome 8 for the reference rhesus macaque individual Mmul_051212, not present in the cohort of samples analysed.

The analysis of the capillary electrophoresis profiles for DEFB4(1)-InDel3, DEFB4(2)-InDel4 and DEFB33-InDel5 revealed for some of the samples in analysis the presence of two peaks, with the same electrophoretic size across the samples. For each sample, the ratio between the areas under the curve for the two peaks was calculated. The average of each duplicate pair was used as normalised value for each independent run, on which the global average ratio, with the respective standard deviation and coefficient of variation, was then derived. **Figure 4.4** presents examples of the peaks distribution observed. **Table 4.2**, **4.3** and **4.4** present the data summary for each of the three assays; each table indicates: the average ratio between the areas under the curve for the two peaks; the standard deviation among the replicates; the variation coefficient, calculated as ratio of the standard deviation on the average; the presumed copy number ratio, manually calculated on the base of the average ratio, the derived copy number estimate and the number of replicates. **Table 4.5** shows the copy number comparison among the three assays. It should be noticed that, as these results have been calculated as ratios between two signal intensities, it is not possible to discriminate among multiples of the copy numbers inferred; for example, 1:1 ratios could correspond to copy numbers of 2, 4, 6 and so forth. For this reason, just the copy number derived from the lowest ratio is indicated.

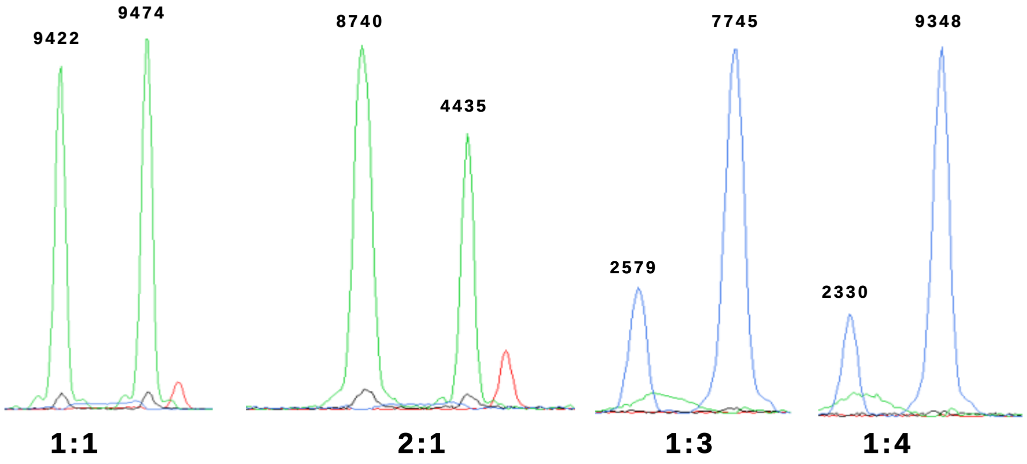


Figure 4.4: examples of peak heights and areas (indicated per each peak) observed through capillary electrophoresis. The ratio between the two peak areas is an indicator of diploid copy number.

Sample name	Average ratio	Standard deviation	Coefficient of variation	Presumed CN ratio	CN	Replicates number
rh1	uninformative	uninformative	uninformative	uninformative	-	6
rh2	2.5	0.15	0.06	5:2	7	5
rh3	2.5	0.20	0.08	5:2	7	7
rh4	uninformative	uninformative	uninformative	uninformative	-	5
rh5	uninformative	uninformative	uninformative	uninformative	-	5
rh6	uninformative	uninformative	uninformative	uninformative	-	5
rh7	2.2	0.23	0.10	2:1 or 5:2 or 7:3	3, 7 or 10	6
rh8	uninformative	uninformative	uninformative	uninformative	-	5
rh9	0.4	0.03	0.08	2:5	7	7
rh10	0.5	0.01	0.02	1:2	3	6
rh11	0.3	0.02	0.07	1:3	4	6
rh12	0.4	0.03	0.08	2:5	7	6
rh13	uninformative	uninformative	uninformative	uninformative	-	4
rh14	2.0	0.09	0.04	2:1	3	6
rh15	2.5	0.17	0.07	5:2	7	6
rh16	2.2	0.02	0.009	2:1 or 5:2 or 7:3	3, 7 or 10	5

Table 4.2: DEFB4(1)-InDel3 results for rhesus macaque samples 1 to 16.

Sample name	Average ratio	Standard deviation	Coefficient of variation	Presumed copy number ratio	CN	Replicates number
rh1	0.85	0.04	0.04	1:1, 4:5, 5:6, 6:7 or 7:8	2,9,11, 13,15	8
rh2	uninformative	uninformative	uninformative	uninformative	-	6
rh3	uninformative	uninformative	uninformative	uninformative	-	6
rh4	1.7	0.07	0.04	7:4 or 5:3	11 or 8	6
rh5	uninformative	uninformative	uninformative	uninformative	-	3
rh6	uninformative	uninformative	uninformative	uninformative	-	6
rh7	uninformative	uninformative	uninformative	uninformative	-	7
rh8	uninformative	uninformative	uninformative	uninformative	-	6
rh9	uninformative	uninformative	uninformative	uninformative	-	7
rh10	uninformative	uninformative	uninformative	uninformative	-	4
rh11	uninformative	uninformative	uninformative	uninformative	-	5
rh12	1.45	0.19	0.13	3:2	5	7
rh13	1.56	0.05	0.03	3:2	5	9
rh14	uninformative	uninformative	uninformative	uninformative	-	8
rh15	1.01	0.007	0.007	1:1	2	8
rh16	1.9	0.05	0.03	2:1	3	8

Table 4.3: DEFB4(2)-InDel4 results for rhesus macaque samples 1 to 16.

Sample name	**Average ratio**	**Standard deviation**	**Coefficient of variation**	**Presumed copy number ratio**	**CN**	**Replicates number**
rh1	uninformative	uninformative	uninformative	uninformative	-	7
rh2	uninformative	uninformative	uninformative	uninformative	-	4
rh3	3.25	0.24	0.07	2:7 or 3:10	9 or 13	3
rh4	3.3	0.18	0.05	2:7	9	5
rh5	0.96	0.02	0.02	1:1	2	3
rh6	3.1	0.19	0.06	1:3	4	6
rh7	4	0.62	0.15	1:4	5	4
rh8	3.1	0.18	0.06	1:3	4	6
rh9	3.16	0.19	0.06	1:3	4	4
rh10	1.5	0.06	0.04	2:3	5	5
rh11	2	0.09	0.04	1:2	3	6
rh12	3.17	0.07	0.02	1:3	4	6
rh13	uninformative	uninformative	uninformative	uninformative	-	6
rh14	1.5	0.07	0.04	2:3	5	6
rh15	2	0.13	0.06	1:2	3	5
rh16	uninformative	uninformative	uninformative	uninformative	-	5

Table 4.4: DEFB33-InDel5 results for rhesus macaque samples 1 to 16.

Sample name	**DEFB4(1)-InDel3 CN**	**DEFB4(2)-InDel4 CN**	**DEFB33-InDel5 CN**
rh1	uninformative	2,9,11,13,15	uninformative
rh2	7	uninformative	uninformative
rh3	7	uninformative	9 or 13
rh4	uninformative	11 or 8	9
rh5	uninformative	uninformative	2
rh6	uninformative	uninformative	4
rh7	3, 7 or 10	uninformative	5
rh8	uninformative	uninformative	4
rh9	7	uninformative	4
rh10	3	uninformative	5
rh11	4	uninformative	3
rh12	7	5	4
rh13	uninformative	5	uninformative
rh14	3	uninformative	5
rh15	7	2	3
rh16	3, 7 or 10	3	uninformative

Table 4.5: Comparison among the copy number estimates obtained with DEFB4(1)-Indel3, DEFB4(2)-InDel4 and DEFB33-Indel5 assays. Uninformative data were generated each time just one peak was observed.

The dataset generated showed discordant results between assays, with a high rate of non-informative samples. DEFB4(1)-Indel3 was uninformative for 6 individuals out of 16, and DEFB4(2)-InDel4 was not informative for 10 samples out of 16. As information on the frequencies of the indels in analysis in the rhesus macaque population is not available, one possible explanation could be that these indels are actually present with low frequency. The only other sequence information available for this area came later in the project from the whole sequencing of two BAC clones spanning the β-defensin area, presented in chapter 5.7. The alignment of the assembled contigs with the expected PCR products confirmed the presence of the shorter DEFB4(1)-Indel3 variant, the longer DEFB4(2)-InDel4 variant and the longer DEFB33-Indel5 variant. Moreover, the presence of two peaks in some of the genomic samples analysed for this project, for the three assays, suggest that the InDels identified were not sequencing artefacts.

Another consideration regards the possibility that DEFB33-InDel5 would fall on another copy number cluster, in proximity of the α-defensin area. This would be consistent with the breakpoint analysis presented in section 3.3.2 and could partially explain the discrepancies in the data observed. Nevertheless, the InDel assays presented provided a further indication of copy number variation in correspondence with *DEFB2L* and may be used in conjunction with more accurate CNV typing assays.

4.2 Paralogue Ratio Test (PRT)

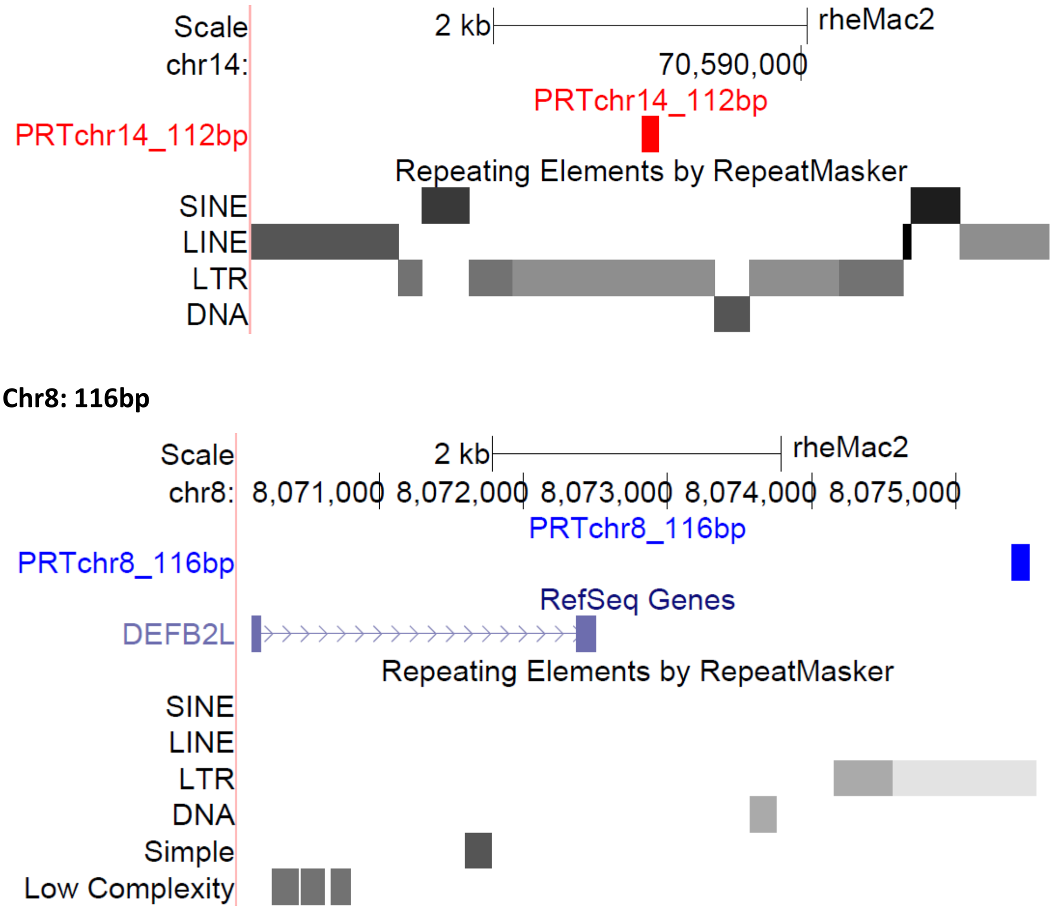
The next assays designed with the aim of performing an accurate copy number typing of the β-defensin locus were Paralogue Ratio Tests. This method is based on the simultaneous PCR amplification of one element within the variable repeat unit and one other unlinked reference locus of similar size, using the same primer pair.

Through capillary electrophoresis it is then possible to perform size-based discrimination of the two PCR products to calculate the ratio of the area under the peaks between the test and the reference. This would allow a precise detection of the diploid copy number of the region of interest for each sample analysed. Nevertheless, in order to obtain primer pairs specific for just two genomic regions of similar but not identical size, primers need to be designed within repeated DNA elements, ideally divergent enough to prevent unspecific amplification across the genome. This task was extremely challenging, also given the presence of gaps in the rhesus macaque assembly and possibly non annotated repeats, which tend to be compressed in genomic assemblies generated through shotgun sequencing. All attempts of manual primers design failed, with all the possible primer candidates generating unspecific products non suitable for PRT. For this reason, a bioinformatics approach was preferred (refer to Section 2.5.1), allowing the design of 6 PRT candidates (**Table 4.6**).

PRT assay	**Chromosomal coordinates**	**PCR product size**
PRT-Alu	chr11: 99853975-99854204 chr8: 8065922-8066139	chr 11: 229bp chr 8: 217bp
PRT 8-16	chr16: 47698513-47698893 chr8: 8075307-8075705	chr16: 380bp chr 8: 398bp
PRT 8-9	chr8: 8075304-8075502 chr9: 47392822-47393021	chr8: 198bp chr9 199bp
PRT 8-14	chr14: 70588986-70589098 chr8: 8075386-8075502	chr14: 112bp chr8: 116bp
PRT 8-3	chr3: 21197456-21197717 chr8: 8075316-8075576	chr3: 261bp chr8: 260bp
PRT 8-8	chr8: 8074589-8074853 chr8:133814175-133814424	chr8: 264bp chr8: 249bp

Table 4.6: candidate assays for PRT. All coordinates refers to January 2006 *Macaca mulatta* assembly (MGSC Merged 1.0/rheMac2).

Given the small size difference (1 bp) between the two expected PCR products, PRT8-9 and PRT8-3 were excluded. PRT-ALU showed a strong bias in the amplification of the PCR product on chr8 (271bp) when run through capillary electrophoresis, consistently producing a second peak of expected size (219bp) with weak intensity not suitable for ratio calculations; it is possible that this difference in intensity was produced by a mismatch in one of the priming sites on chr11, reducing the PCR reaction efficiency. PRT8-8, PRT 8-16 and PRT8-6 produced non suitable electrophoretic profiles, with different peaks produced instead of one expected of size similar to the reference locus, suggestive of non-specific amplification from other repeated elements. PRT 8-14 gave two consistent peaks of expected size (112 and 116bp), and was hence used as copy number typing assay for the β-defensin region. **Figure 4.5** shows the chromosomal coordinates of the two PCR products amplified with PRT8-14 assay.



Chr14: 112bp

Chr8: 116bp

Figure 4.5: chromosomal coordinates for the two PCR products co-amplified with PRT 8-14. Both products are amplified from LTR repeats. In the lower panel the blue Genome Browser track evidences the location of *DEFB2L* gene (chr8:8070118-8072503).

A first set of PRT8-14 was performed on the same rhesus macaque samples used for the array CGH (Rh1-Rh16). Resulting ratios are averaged from six independent PCR reactions and correspondent electrophoretic runs. Copy number is calculated multiplying by two the copy number ratio, under the assumption of copy number double of the mean ratio. **Table 4.7** shows the averaged results obtained for the samples in the analysis. **Figure 4.6** presents the copy number distribution obtained from the six replicates.

sample	average ratio	standard deviation	variation coefficient	Copy number
rh1	3.69	0.26	0.07	7.4
rh2	2.85	0.62	0.22	5.7
rh3	3.95	0.32	0.08	7.9
rh4	4.29	0.21	0.05	8.6
rh5	2.54	0.43	0.17	5.0
rh6	4.01	0.32	0.08	8.0
rh7	4.91	0.18	0.04	9.8
rh8	4.16	0.14	0.03	8.3
rh9	4.47	0.23	0.05	9
rh10	5.15	0.22	0.04	10.3
rh11	4.55	0.27	0.06	9.1
rh12	4.64	0.25	0.05	9.3
rh13	4.77	0.29	0.06	9.6
rh14	5.00	0.10	0.02	10
rh15	4.13	0.24	0.06	8.2
rh16	4.28	0.33	0.08	8.6

Table 4.7: PRT 8-14 results on rhesus macaque samples Rh1-Rh16.

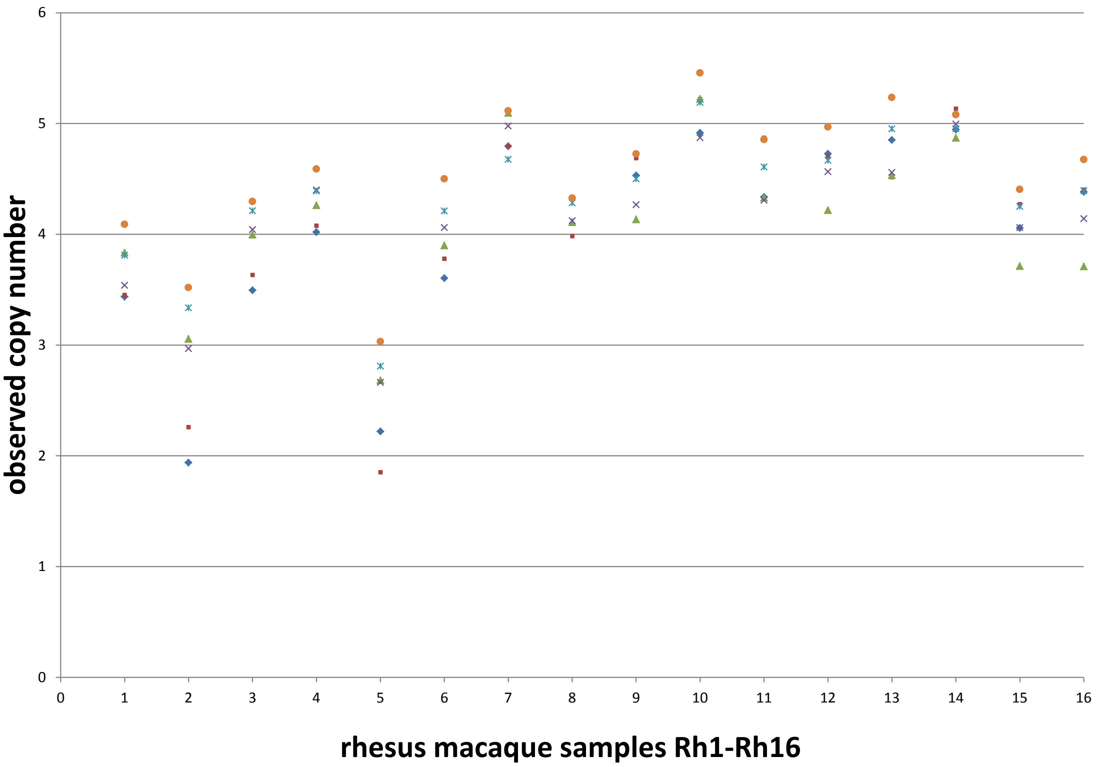


Figure 4.6: Observed copy number distribution for 5 independent PRT 8-14 experiments on 16 rhesus macaque samples in analysis. Numbers on X axis correspond to Rh1-Rh16. Y axis indicates the copy number values obtained for each replicate.

The PRT 8-14 assay successfully called copy number in all the samples in analysis, with overall concordance among replicates. Copy number values ranged between 5 and 10. Nevertheless, several samples did not cluster around an integer copy number value, as it would have been biologically expected.

The assay was also unsuccessfully tested on *Macaca sylvanus* DNA (4 samples, (EURPRIMAT Consortium) with the amplification of non specific bands; it is likely that, despite the closeness between the two species, secondary annealing sites for the PRT8-14 primers could be present in *Macaca sylvanus*.

A second series of PRT 8-14 assays were performed on a cohort of 51 rhesus macaque DNA samples provided by Dr Omer Gokcumen (Department of Biological Sciences-SUNY, Buffalo, US); these DNA samples were sent in concentration suitable for PCR amplification, despite poor DNA quality for part of the cohort, which in some cases gave insufficient amplification. A further 15 rhesus macaque DNA samples (11 samples property of Dr Edward Hollox, University of Leicester, and 4 samples extracted from rhesus macaque lymphoblastoid cell lines, provided by Dr Gaby Doxiadis, Biomedical Primate Research Centre Rijswijk, NL) were added to the cohort, to increase the sample size to a total of 66 samples.

PRT 8-14 assay on all the Gokcumen's cohort samples was repeated four times, despite for 30 samples it was not possible to obtain four replicates, due to insufficient amplification. Just one sample failed in all the runs. The rest of the cohort was repeated three times, with concordant results among replicates. Copy number distribution for this cohort is shown in **Figure 4.7**.

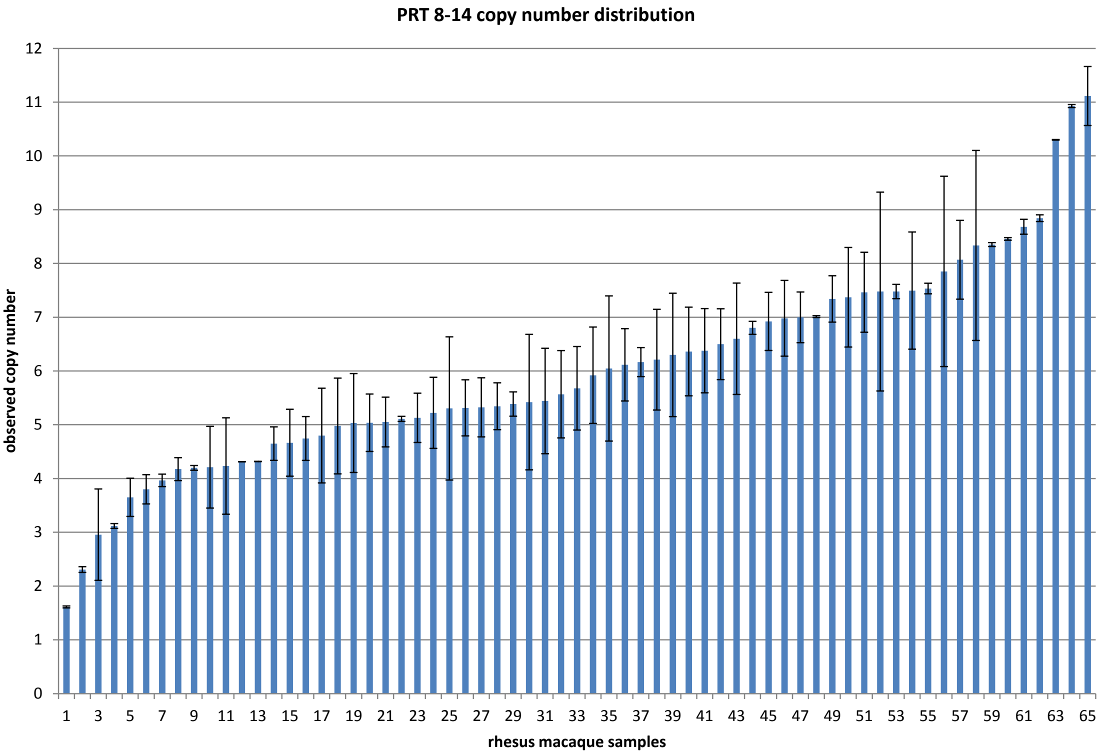


Figure 4.7: Copy number distribution for a cohort of 65 rhesus macaque samples called with PRT 8-14. Numbers on the X axis correspond to the cohort's samples, ordered for increasing copy number. Y axis indicates the observed copy number. Error bars show the standard deviation among replicates. Standard deviations are missing for the samples that failed 3 runs out of 4.

PRT 8-14 results from a larger cohort of rhesus macaque samples showed a copy number distribution ranging between 2 and 11. In two of the runs, four samples in duplicate from high quality DNA extracted from rhesus macaque lymphoblastoid cell lines was used as positive controls, with high concordance between intra- and inter-experiment duplicates. Interestingly, a similar copy number distribution has been reported for the human β-defensin cluster [29], despite the larger size of the human repeat unit of approximately 240kb, compared with 8.5kb in rhesus macaque (as will be presented in Chapter 5). These findings may be consistent with a deleterious effect of higher BD2 copy numbers in both lineages, possibly as consequence of excessive immune system activation. Moreover, considering that the arise of CNV in rhesus macaque and human β-defensin region are likely to have happened as consequence of independent events of duplication and divergence, these data may be read under the perspective of a functionally convergent copy number distribution in the two species.

4.3 Microsatellite Assay

As described in the previous section, despite PRT8-14 was demonstrated to call consistently copy number, one main limitation of this assay was the lack of copy number clustering around integer values. In order to circumvent this problem, a new assay was developed that could constitute a valid support for the interpretation of copy number calls in between two round values.

The 'Microsatellite Assay' was based on the presence of a $(GGGA)_n$ simple repeat microsatellite in the intron of *DEFB2L* gene, with an annotated size of 180bp (**Figure 4.8**). Under the assumption that different copies of *DEFB2L* may present different microsatellite repeat lengths, a primer pair flanking the repeat was designed, with the left primer conjugated to a fluorophore. After PCR amplification, microsatellite repeats of different size could be separated through capillary electrophoresis, allowing at the same time a relative quantification of the resulting peaks (**Figure 4.9**). Minimum copy numbers are estimated summing all the peak areas and dividing the result by the peak with the smallest area. It should be noticed that this technique is not able to discriminate among multiples of the minimum copy number predicted. Nevertheless, it could be suitable to discriminate between even and odd copy number, in case of decimal copy number calls.

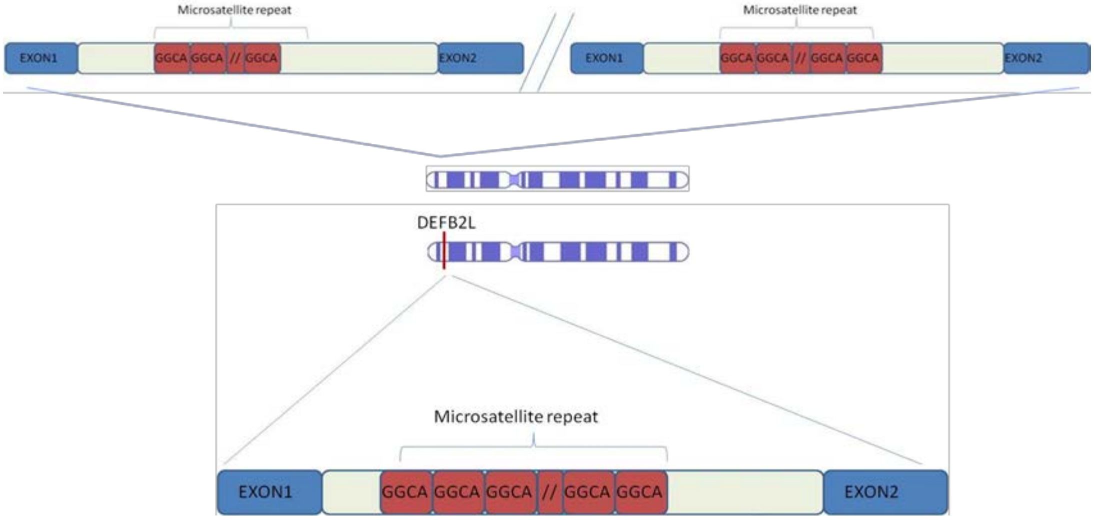


Figure 4.8: Cartoon of the (GGGA)$_n$ microsatellite repeat present in the *DEFB2L* intron. In the case shown, one chromosome 8 presents two *DEFB2L* copies, that will generate two fragments of different length, whereas the other chromosome 8 presents one copy of the gene, that will generate a third fragment. Hence, through capillary electrophoresis, three peaks would be expected.

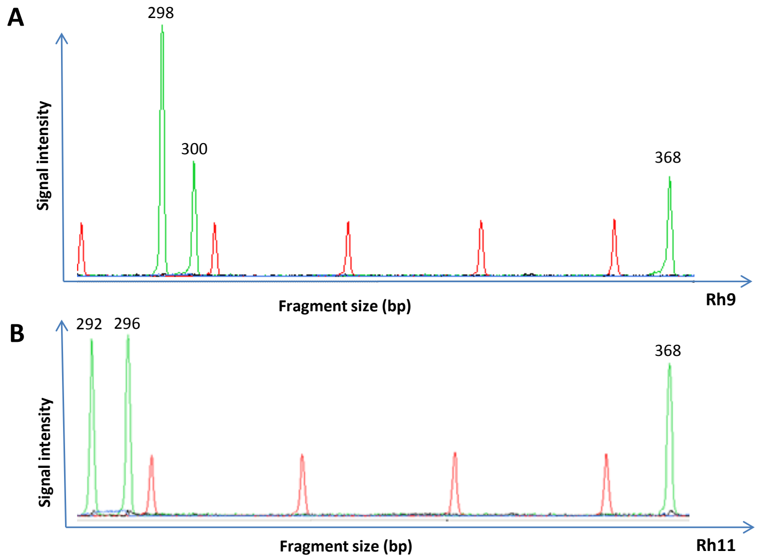


Figure 4.9: Examples of electrophoretic profiles obtained for rhesus macaque samples Rh9 and Rh11. X axis shows the fragment size; Y axis shows the signal intensity. Panel **A** presents a case of two copies of *DEFB2L* bearing two microsatellite repeats of the same length (298 bp) that generate a peak of double intensity and further two gene copies with microsatellites of different size (300bp and 368bp). Panel **B** shows three gene copies with microsatellites of different sizes (292bp, 296 bp and 368 bp).

The Microsatellite Assay was first used on the cohort of high quality DNA samples used for the array CGH (Rh1-Rh16), with three independent replicates. Results are shown in **Table 4.8**.

sample	**Estimated CN**	**standard deviation**	**variation coefficient**	**PRT8-14 Copy number**	**Concordance**
rh1	1	0.00	0.00	7.4	+ (any)
rh2	3	0.18	0.06	5.7	+ (6)
rh3	4.2	0.31	0.07	7.9	+ (8)
rh4	2	0.13	0.06	8.6	+ (8)
rh5	4.8	0.00	0.00	5.0	+ (5)
rh6	4.1	0.15	0.04	8.0	+ (8)
rh7	4.9	0.34	0.07	9.8	+ (10)
rh8	4	0.29	0.07	8.3	+ (8)
rh9	4.2	0.10	0.03	9	- (8?)
rh10	3.8	1.34	0.55	10.3	- (8 or 12?)
rh11	3.1	0.22	0.07	9.1	+ (9)
rh12	2.1	0.07	0.03	9.3	- (8 or 10?)
rh13	2	0.03	0.02	9.6	+ (10)
rh14	3.9	0.26	0.07	10	-(8 or 12)
rh15	2.8	0.32	0.12	8.2	- (9?)
rh16	2.1	0.11	0.05	8.6	+ (8)

Table 4.8: Microsatellite Assay results of rhesus macaque samples Rh1-Rh16. Columns present: the estimated minimum copy number from the assay; standard deviation among the replicates and the correspondent variation coefficient; the copy number calls derived from PRT8-14 assay; in the last column, + indicates concordance between PRT and Microsatellite Assay copy number calls, - indicates non-concordance.

These results show concordance in 11 samples out of 16 and are hence indicative of the technique potential to help discriminate between decimal copy number calls. Nevertheless, the inaccuracy of the Microsatellite Assays showed the tendency to increase in case of copy numbers of 8-9-10. This aspect could be explained by the fact that, in parallel with the copy number increase, there will be statistically an increased number of microsatellites of the same size; for high copy numbers, the resolution power of capillary electrophoresis to detect a one-copy difference decreases, posing problems in calculating the relative ratio.

4.4 digital droplet PCR (ddPCR)

An alternative method that just recently became available for copy number detection is digital droplet PCR (ddPCR, Section 1.3.3.6), based on the generation of discrete droplet partitions containing randomly distributed DNA template, to the point where some contain no template and other contain one or more copies of nucleic acid. In order to confirm the copy number calls typed with PRT 8-14, a new ddPCR-based assay was developed and tested, on the same cohorts of rhesus macaque samples previously described. A primer pair amplifying 125bp of the second *DEFB2L* exon was designed, together with a Taqman® probe specific for the amplified sequence; a previously described assay specific for *PAX9* was selected as reference assay [79]. **Table 4.9** presents the averaged copy numbers obtained with four independent runs; the software Quatasoft™ directly calls copy number, as the ratio between single positive droplets for the test on single positive droplets for the reference. **Figure 4.10** presents the copy number distribution obtained from the four replicates. Replicates were included in the analysis if the total number of accepted droplets was >10000.

sample	Copy number	standard deviation	variation coefficient
rh1	4	0.68	0.17
rh2	3.9	0.33	0.09
rh3	4.9	0.69	0.14
rh4	5.5	0.36	0.07
rh5	2.9	0.92	0.32
rh6	2.9	0.90	0.31
rh7	5.1	0.45	0.09
rh8	4.9	0.59	0.12
rh9	5.4	0.50	0.09
rh10	5.5	0.74	0.13
rh11	4.6	0.39	0.09
rh12	4.9	0.30	0.06
rh13	4.7	0.48	0.10
rh14	5.7	0.64	0.11
rh15	5.0	0.09	0.02
rh16	4.3	0.71	0.16

Table 4.9: ddPCR results on rhesus macaque samples Rh1-Rh16.

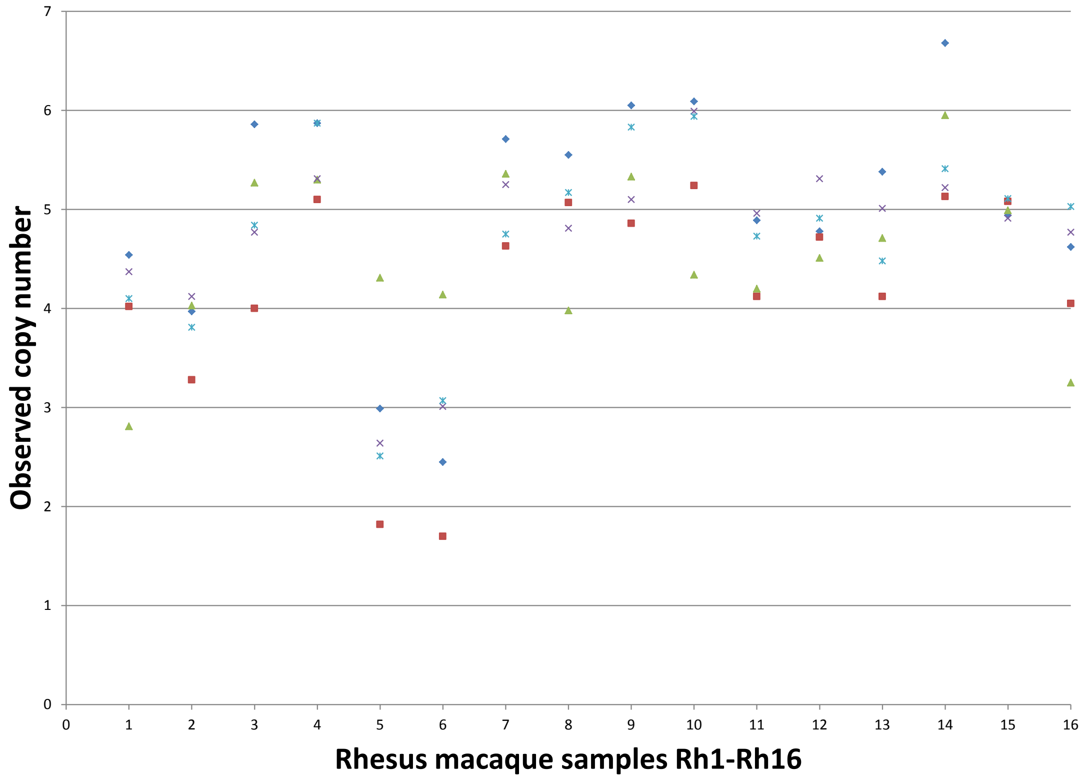
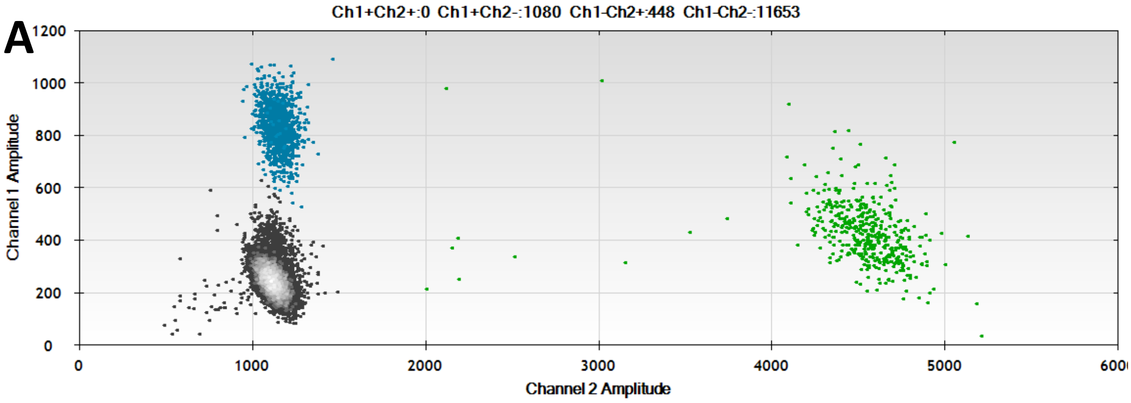


Figure 4.10: Observed copy number distributions for 4 independent ddPCR runs on the high-quality cohort of 16 rhesus macaque samples. Numbers on X axis correspond to Rh1-Rh16. Y axis indicates the copy number values obtained for each replicate.

These results evidenced a higher dispersion of the data among replicates compared with PRT8-14, with excessive standard deviation values. Copy number values called with ddPCR ranged between 4 and 5.7. Moreover, despite the principle of ddPCR was based on the readings of four different populations of droplets, according to the possible binary states of presence/absence of test/template, the double positive population was never observed for the ddPCR-*DEFB2L* assay designed. All the reads were hence manually screened, to ensure a correct gating among droplet populations, but no separations of the +/+ population was visible (**Figure 4.11**).



B

test +/ reference -	**test +/ reference +**
test -/ reference -	**test -/ reference +**

Figure 4.11: Droplets population clusters observed for the ddPCR *DEFB2L* assay. Panel **A** shows the clustering pattern obtained for all the samples: in blue test +/reference -, in black test- / reference -, in green test - / reference +. Panel **B** indicates the expected droplet population clusters: the double-positive population test+ / reference +is missing in panel **A.**

A second set of ddPCR-*DEFB2L* assays was made on 66 samples from the lower-quality DNA cohort, in order to evaluate the copy number distribution of *DEFB2L* and compare it with the results obtained through PRT 8-14. The assay was run just once, with all the 66 samples in one plate, plus 14 *Macaca sylvanus* DNA samples (provided by Dr Patricia Balaresque, CNRS Toulose) and one *Macaca fuscata* DNA sample. The copy number distribution for the *Macaca mulatta* samples is shown in **Figure 4.12**. All the *Macaca sylvanus* samples did not show signs of amplification; nevertheless, the only *Macaca fuscata* DNA sample available presented an amplification comparable with all the *Macaca mulatta* samples, and showed a copy number of 5. Also for this dataset, in all the samples the population of double positive droplets +/+ was not detected.

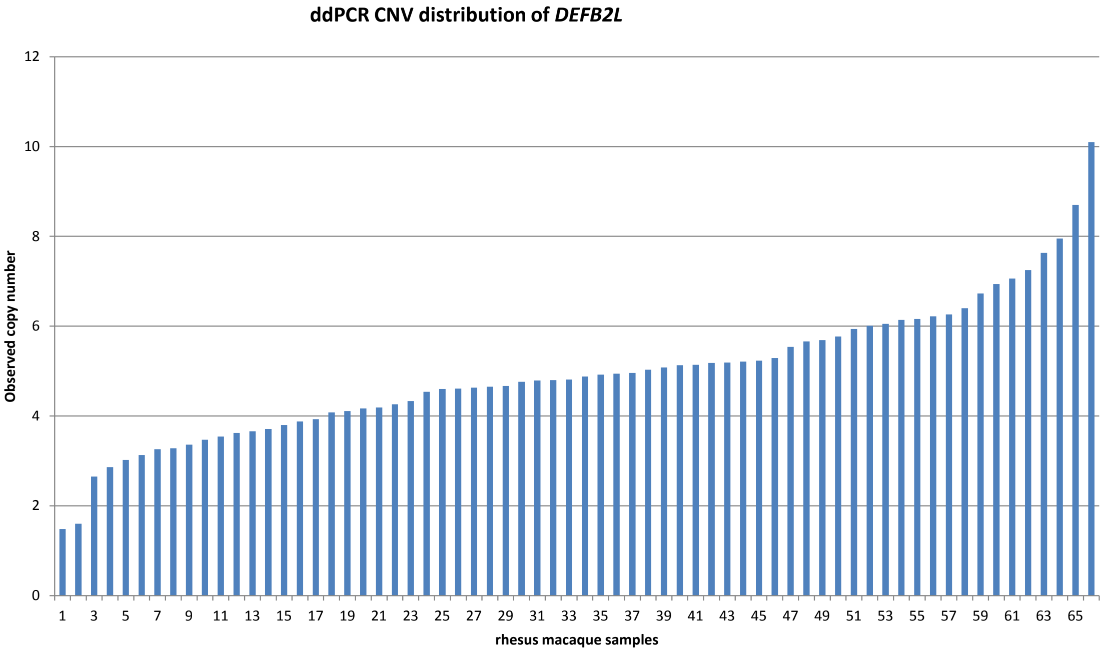


Figure 4.12: *DEFB2L* copy number distribution for a cohort of 66 rhesus macaque samples called with ddPCR. Numbers on the X axis correspond to the cohort's samples, ordered for increasing copy number.

The copy number distribution of the cohort of 66 low-quality DNA samples ranged between 2 and 10. Copy numbers were positively called in all the samples, despite it was not possible to observe a clustering around integer values.

4.5 Comparison of *DEFB2L* copy number distribution called with PRT8-14 and ddPCR

Given the results obtained with PRT 8-14 and digital droplet PCR, a comparative analysis between the two datasets was necessary to establish the reliability of the two techniques. The comparison of the copy number calls for Rh1-Rh16 samples are shown in **Figure 4.13** whereas **Figure 4.14** presents the compared *DEFB2L* copy number distribution for the cohort of 66 low-quality rhesus macaque DNA samples.

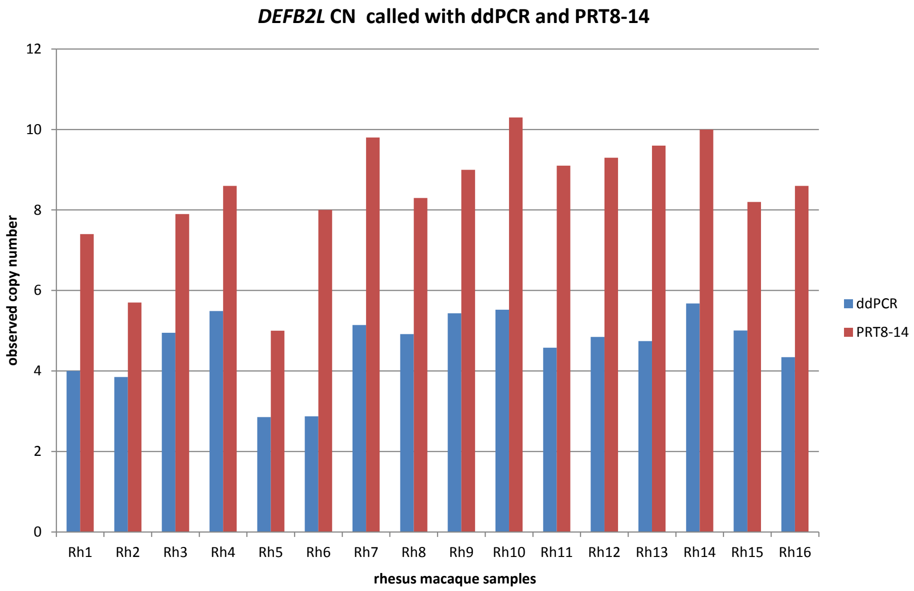


Figure 4.13: Comparison of *DEFB2L* copy number called with PRT8-14 (red) and digital droplet PCR (blue) for the cohort of high-quality DNA samples used for the array CGH. On the X axis all the 16 samples are indicated.

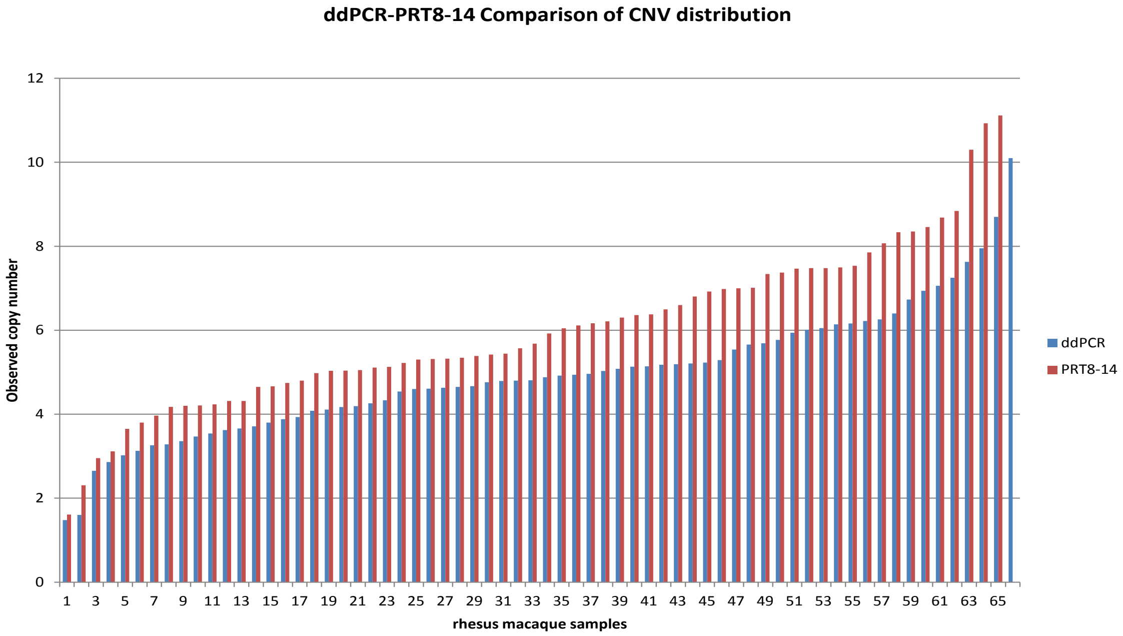


Figure 4.14: Comparison of *DEFB2L* copy number distribution called with PRT8-14 (red) and digital droplet PCR (blue) for the same cohort of 66 rhesus macaque samples. Numbers on the X axis correspond to the cohort's samples, ordered for increasing copy number.

From this comparative analysis it was possible to observe a systematic difference between the two techniques, with PRT8-14 calling 1-2 copies more than ddPCR. This difference showed a tendency to accentuate for higher copy number calls, especially visible from Figure 4.14. Nevertheless, the comparison of the two datasets obtained for Rh1-Rh16 using the Pearson's correlation coefficient gave a value of 0.75, hence suggesting the presence of a positive correlation between the two techniques.

One aspect that should be taken into account regarding the accuracy of ddPCR, is the lack of +/+ droplets population in all the samples analysed. This population was statistically expected to be present but for unknown reasons it did not cluster separately from the others. Hence, in case of misassignment, the +/+ population may have been included in the test-/reference+ subgroup, determining a systematic decrease in the copy number call, calculated as single positive for the test/single positive for the reference.

Moreover, the replicate distribution shows a consistently higher standard deviation compared with PRT, suggesting that different factors could impair the efficiency of ddPCR, despite the great number of intra-sample replicates generated through emulsion PCR.

4.5 Concordance of PRT 8-14 and digital droplet PCR with array CGH data for the *DEFB2L* region

An alternative way to compare the efficiency of the two endpoint PCR assays used for copy number calling of the rhesus macaque β-defensin region came from the results obtained from the same area analysed with high density array CGH data.

If PRT8-14 and ddPCR were calling real copy number values, then for both datasets the logarithm of the copy number ratio between each sample in the analysis and the copy number value of Rh16, used as reference sample for the array CGH, should give comparable values. These transformed values were compared with the averaged aCGH intensity value for the putative *DEFB2L* copy number unit (chr8:8068749-8076651) for each individual, using the Pearson's correlation coefficient. Correlations of PRT8-14 dataset with the array CGH dataset gave a correlation coefficient of 0.73, whereas correlation of ddPCR dataset with the array CGH dataset gave a lower correlation of 0.61.

Another estimate of correlation came from a Principal Component Analysis (PCA) performed on the array-CGH data spanning the putative *DEFB2L* copy number variable area (20 probes, chr8:8068749-8076651). This mathematical approach was chosen in order to transform the series of log intensity ratios into linearly uncorrelated variables, called principal components. In particular, the first orthogonal component obtained (principal component 1) is calculated in order to account for the largest possible variance. The principal component 1 can then be used as non correlated variable to compare the array-CGH data with data generated with different techniques that assume different variance.

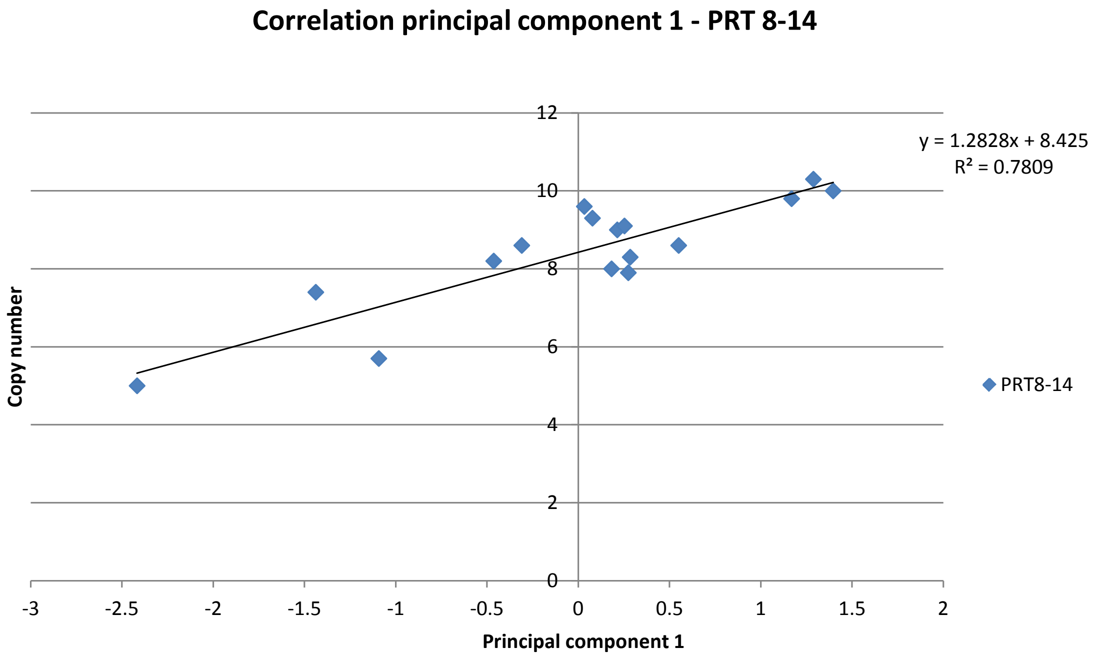


Figure 4.15: Graph showing the Principal component 1 from the array CGH principal component analysis performed on the β-defensin region, and the copy number calls for the correspondent rhesus macaque samples detected through PRT 8-14. The trend line presents an r^2 -squared value=0.78.

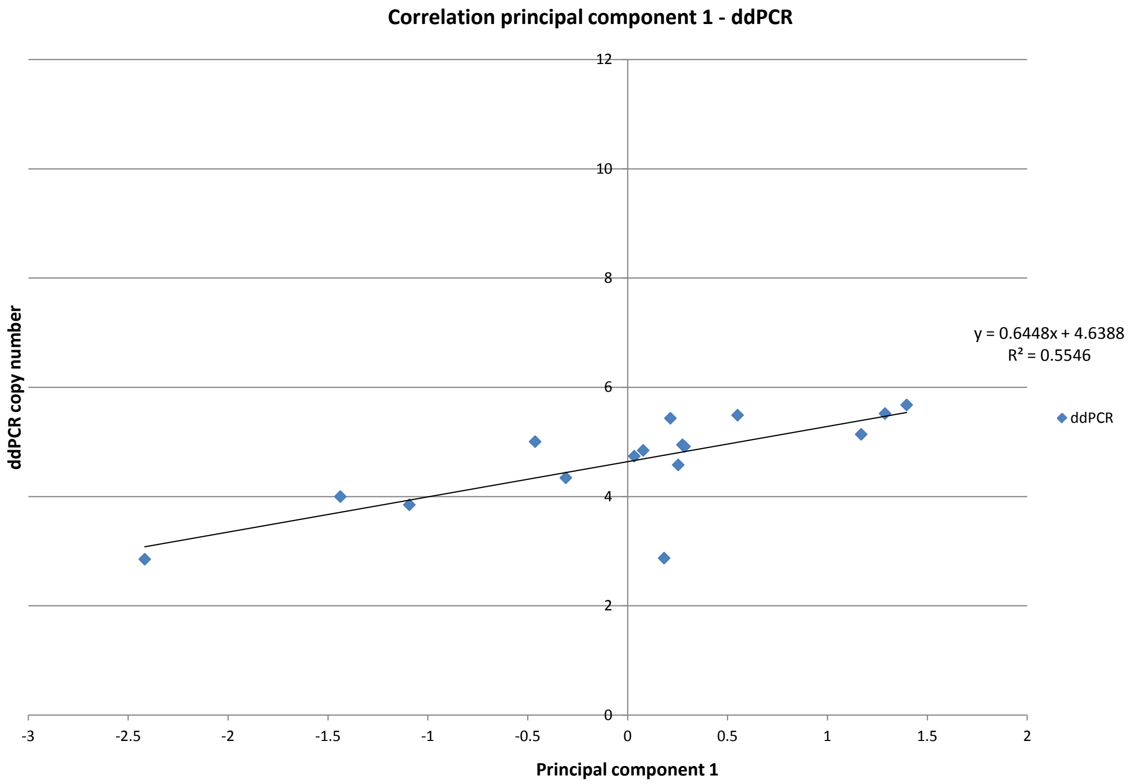


Figure 4.16: Graph showing the Principal component 1 from the array CGH principal component analysis performed on the β-defensin region, and the copy number calls for the correspondent rhesus macaque samples detected through digital droplet PCR. The trend line presents r^2-squared value=0.55.

These results evidenced a higher correlation of PRT8-14 with the array CGH dataset, compared with digital droplet PCR, suggestive of higher precision. Nevertheless, from **Figure 4.16** it was possible to observe the presence of just an outlier deviating from the values distribution, correspondent to Rh6. If this sample was taken out from both datasets, the resulting r values became perfectly comparable between PRT8-14 (r^2=0.79) and ddPCR (r^2=0.84) (**Figure 4.17**).

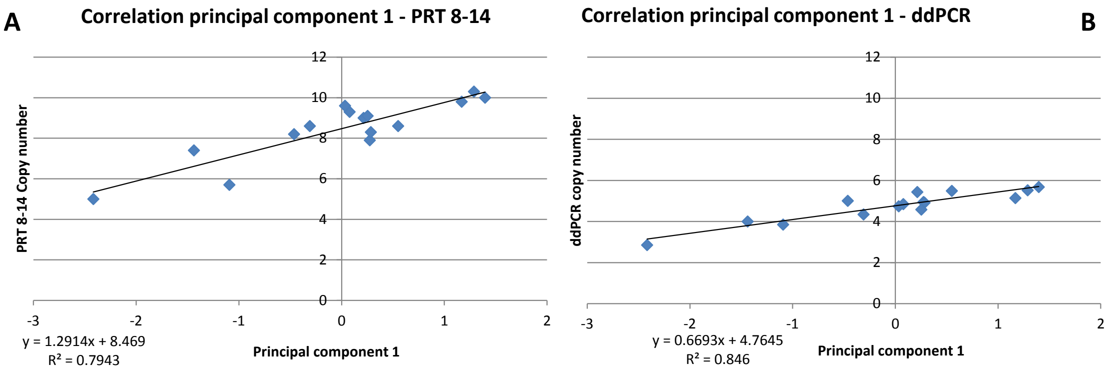


Figure 4.17: Graph showing the new Principal component 1 from the array CGH principal component analysis performed on the β-defensin region, and the copy number calls for the correspondent rhesus macaque samples detected through PRT8-14 (Panel **A**) and digital droplet PCR (Panel **B**). The sample Rh6 has been excluded from both datasets. The trendline presents r^2-squared value = 0.79 for PRT8-14 and r^2-squared value=0.84 for ddPCR.

These results demonstrated comparable performances for both PRT8-14 and digital droplet PCR that allowed to obtain more accurate information on the copy number distribution of *DEFB2L* in rhesus macaque.

4.6 Summary

- This chapter presents four PCR based methods that have been used for copy number typing of the rhesus macaque β-defensin region, illustrating limitations and advantages of each.

The InDel assays, based on the presence of short indels in paralogous copies, did not present a good level of accuracy, possibly for the low frequency distribution of the indels selected for the assay design.

Paralogue Ratio Test (PRT) presented a good level of concordance among replicates for copy number calling of a cohort of 66 low-quality rhesus macaque DNA samples and the same16 high-quality DNA samples used for the array-CGH.

The Microsatellite Assay, based on the presence of a microsatellite with variable length of $(AGGG)_n$ repeats in the intron of *DEFB2L*, could be a useful test to discriminate between odd and even copy numbers, despite a loss of accuracy in case of higher copy number call.

Digital droplet PCR is a method with ease of use comparable with PRT, but it presented a higher variability among replicates. Nevertheless, it is possible that higher levels of accuracy could be reached optimising or changing the ddPCR *DEFB2L* assay, as the absence of the population of double positive droplets in each read may have biased the copy number calls.

Notwithstanding, on a total of 82 rhesus macaque samples, it was possible to call copy numbers ranging between 2 and 11, demonstrating that the area encompassing *DEFB2L* is copy number variable in rhesus macaque monkey.

Analysis of the β-defensin region using large insert clones

This section presents the results obtained through the sequence analysis of bacterial artificial chromosome (BAC) clones carrying the *DEFB2L* gene, after screening of the *Macaca mulatta* BAC library CHORI-250 (BACPAC Resources Center). BAC constructs have the advantage of containing stable inserts of large size (150-350kb) and hence constituted an ideal approach to obtain more accurate sequence information on the β-defensin area. Moreover, BAC constructs provide haploid information, allowing retrieval of separate information from the two chromosomal copies of the area of interest.

5.1 Probe selection and library screening

For the BAC library screening, primers for four different probe candidates were designed, with the aim of amplifying PCR products not containing repeated elements (Section 2.8.1). This aspect was crucial, as the procedure of radioactive labelling needed to be performed with the DNA polymerase I Klenow fragment that generates a pool of short fragments instead of full length products. The presence of repeated elements in the probe sequence would have resulted in high non-specific hybridisation. As shown in **Figure 5.1**, all the PCR amplifications of the candidate probes gave specific bands of expected size, as predicted with the *in silico* PCR tool of UCSC Genome Browser.

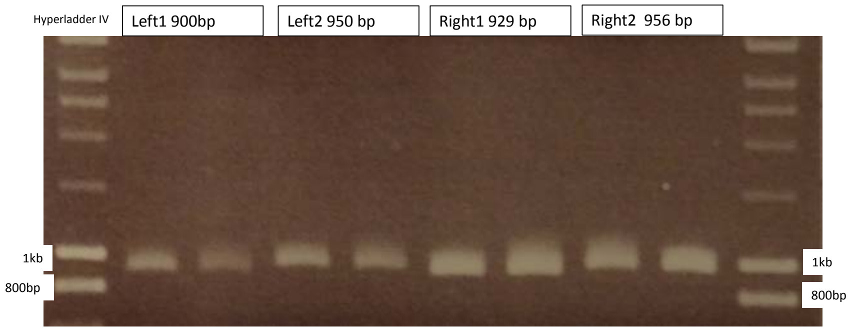


Figure 5.1: 0.8% agarose gel electrophoresis showing a single band of expected size per each candidate probe for the BAC library screening. Each primer pair has been tested on two different rhesus macaque samples. 5 µl of PCR product was run on each lane. 366ng of HyperladderIV were used as DNA marker.

All the four assays performed similarly and hence one distal and one proximal probe to *DEFB2L* were selected. The PCR products Left1 and Right1 were cleaned with Quiaquick DNA purification kit, as described in Section 2.8.1, quantified by absorbance at 260 nm and pulled together to be labelled in equal concentration. **Figure 5.2** shows the position of the two probes on chromosome 8, flanking *DEFB2L.*

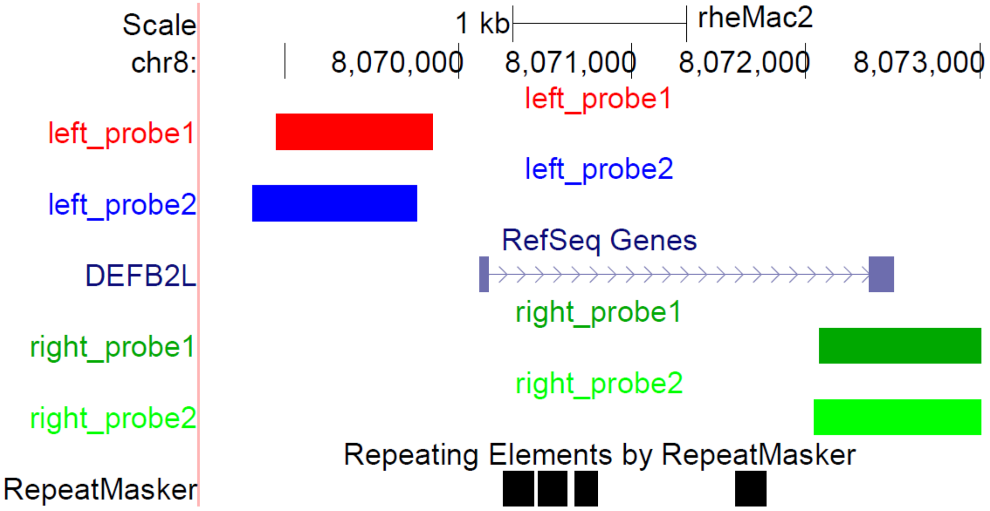


Figure 5.2: position of the probes used to screen clones positive for the *DEFB2L* gene: Left1 (red), Left2 (blue), Right1 (dark green) and Right2 (light green). *DEFB2L* position is indicated in blue. None of the probes contains repeated elements (as shown by the repeat masker track, in black).

Probe labelling and filters hybridisation was performed as described in Sections 2.8. **Table 5.1** lists the positive clones identified.

Filter number	Clone name
1	47B11
	47D7
	47F7
	45E17
2	65I2
3	104C3
	135L4
	121P8
4	148I5
	151M14
5	212G13
	198C4
	201P10
	213L6
	217D13
6	262P20
	246K23
	275L13
	251J8
	243E20
	279A15

Table 5.1: List of clones positive for *DEFB2L.*

Before selecting the clones to be ordered, the High Throughput Genomic Sequences (HTGS) database of BLAST was interrogated, using Left1 and Right1 sequences as queries. Two hits were produced, from two BAC clones whose sequence has been deposited online: 65I2 (accession number: AC193549.4, complete sequence, 99% identity with Left1 and 100% identity with Right1) and 243E20 (accession number: AC191454.4, working draft sequence, 98% identity with Left1 and 100% identity with Right1). Both of these clones contain the *DEFB2L* sequence, and they have been positively identified with the procedure hereby described (on filter 2 and filter 6 respectively). This constituted a good positive control for the experiment accuracy. One clone per filter has been selected for further analysis (**Table 5.2**) with the exception of Filter 2, as its only positive clone was already described. Two clones were selected from Filter 6 instead.

Filter number	Clone name
1	47B11
3	135L4
4	148I5
5	217D13
6	201P10
	246K23

Table 5.2: list of clones ordered for further analysis.

5.2 BAC DNA extraction

Upon arrival, clones were grown as described in Section 2.8.5. The main goal was the characterisation of the β-defensin region to retrieve separate information from the two chromosomal copies of the area of interest. With this aim, different approaches were used: PFGE, BAC-end sequencing, whole BAC sequencing for selected clones and metaphase spreads fluorescence *in situ* hybridisation, as it will be presented in this chapter. All these techniques required BAC DNA of high purity and with high yield to succeed. With this aim, four different methods for BAC DNA extraction have been used: column-based extraction kit, alkaline lysis-based maxiprep, column-based on gravity flow anion exchange chromatography and alkaline lysis-based extraction with separation on caesium chloride ultracentrifugation gradient. This section presents an overview on the performances and limitations of these approaches.

5.2.1 E.Z.N.A.® Endo-Free Plasmid Maxi Kit

The first approach tried made use of the E.Z.N.A.® Endo-Free Plasmid Maxi Kit, a column-based method designed for the isolation of large-scale constructs. With this approach, the BAC DNA yield was apparently high, with spectrophotometric quantifications at 260nm giving DNA concentrations in the order of 0.5µg/µl. Repeated failure of down-stream BAC-end sequencing using this DNA suggests carry-over of bacterial chromosomal DNA, leading to systematic overestimation of BAC DNA concentration.

5.2.2 Alkaline lysis based maxiprep

A second approach was tried, to test the hypothesis of an insufficient starting volume of bacterial cultures, using the protocol for isolation of BAC DNA from large-scale cultures (500ml), described in [207] Vol.1, Chapter 4.55. Spectrophotometric quantifications of the extracted DNA at 260nm gave reads in the order of 2-3 µg/µl. Given the poor results obtained with the column-based kit previously described, 1µl of each extracted DNA sample were run on gel electrophoresis to perform a further DNA quantification, using different concentration of HyperladderI as standard. Bands of intensities proportional to the calculated DNA concentration were expected. In order to increase the DNA yield, a second BAC DNA elution was performed for all the samples.

This method brought to low BAC DNA yield. Also in this case spectrophotometric quantifications, in the order of 2-3 µg/µl, were largely overestimating the real extraction efficiency. From the comparison with HyperladderI bands, real concentrations were approximately 100ng/µl for 201P10 and 47B11 and 30ng/µl for 217D13. Furthermore, it was possible to observe the presence of a considerable contamination, possibly short RNA fragments, fast-migrating on agarose gel. Longer run of the same gel, up to 1h 50min, evidenced the lost of these small sized fragments. A similar pattern of low DNA yield and presence of contaminants was hence observed in two independent maxipreps, challenging the reliability of this method for extraction of low-copies large insert clones.

In order to confirm the discrepancies observed in the BAC DNA extraction yield for the 6 clones,, a fluorescence-based quantification was performed, using the Quant-iT™ PicoGreen® System, specific for double-strand DNA quantification (Section 2.8.5.2). Two standard curve were prepared for data calibration: the high range standard curve, created with 4 serial dilutions, with DNA concentrations from 1µg/ml to 5ng/ml, and one blank, gave a R^2 value of 1; the low range standard curve, with 4 points with DNA concentrations from 25ng/ml to 125pg/ml, and one blank, gave a r^2 value of 0.99. Results of the subsequent samples quantification are listed in **Table 5.3**.

Clone name	DNA concentration [ng/µl]
47B11	192
135L4	210
201P10	114
148L4	213
246K23	214
217D13	232

Table 5.3: BAC DNA concentration for the 6 samples extracted with standard maxiprep, accordingly to PicoGreen® measurements.

The PicoGreen® quantification suggested that the BAC DNA concentrations obtained were suitable for PCR amplification, given the low template concentration required. Nevertheless, the low DNA yield was problematic when these DNA samples were used for BAC-end sequencing, generating low-quality reads. Another critical issue was the presence of contaminants, probably of different nature: bacterial nucleic acids, leading to a dramatic overestimation of the BAC DNA concentration, and possibly ethanol or other solvents used in the DNA extraction procedure, which could have interfered with the BAC-end sequencing reaction.

One key point likely to have decreased dramatically the DNA yield is that the standard procedure to induce bacterial cells to increased synthesis of BAC DNA is based on the use of the antibiotic chloramphenicol: in non-lethal doses, chloramphenicol inhibits the bacterial protein synthesis, without affecting the synthesis of new BAC copies; this has the effect of decreasing the bacterial growth rate increasing at the same time the average number of BAC copies per cell. This procedure is particularly advantageous to extract low-copy constructs, as BACs. Nevertheless, the vector used for the CHORI-250 library preparation, pTARBAC2.1 (**Figure 2.2**), contains a chloramphenicol resistance cassette, used for selective screening of the BAC-positive clones. For this reason, the chloramphenicol-based induction system was not applicable.

Despite the low yield, the BAC DNA extracted with this procedure was used to check the approximative size of the BAC inserts. Following *Not*I restriction enzyme digest, 2µg of DNA per each clone were run on a PFGE gel showing insert sizes around 200kb, as expected.

5.2.3 Anion exchange chromatography

In order to increase the BAC DNA purity and yield, an alternative method was used, applying an anion exchange chromatography system specifically designed for the isolation of BAC constructs, the NucleoBond® BAC 100 (Macherey-Nagel, CAT N° 740579). BAC DNA extracted with this kit gave good yield, with an average of 40µg of BAC DNA per each clone. A quality control aimed at the identification of possible bacterial genomic DNA carryover was performed, using different restriction enzyme patterns. 400kb of sequence centred on the *DEFB2L* gene, retrieved from the rhesus macaque UCSC Genome Browser assembly, were analysed using NEBcutter V2.0 with the aim of identifying suitable cleavage patterns. The selection was made on restriction enzymes cutting between 4 and 20 times: BamHI (13), KpnI (8), *Pml*I (10) and *Pme*I (5). 2µg of BAC DNA were used per each digestion. Results of the restriction pattern are shown in **Figure 5.3**. In all cases, it was possible to observe a discrete banding pattern, despite the presence of a weak smear, that could be indicative of low levels of bacterial DNA contamination or an excessive amount of BAC DNA loaded per each lane.

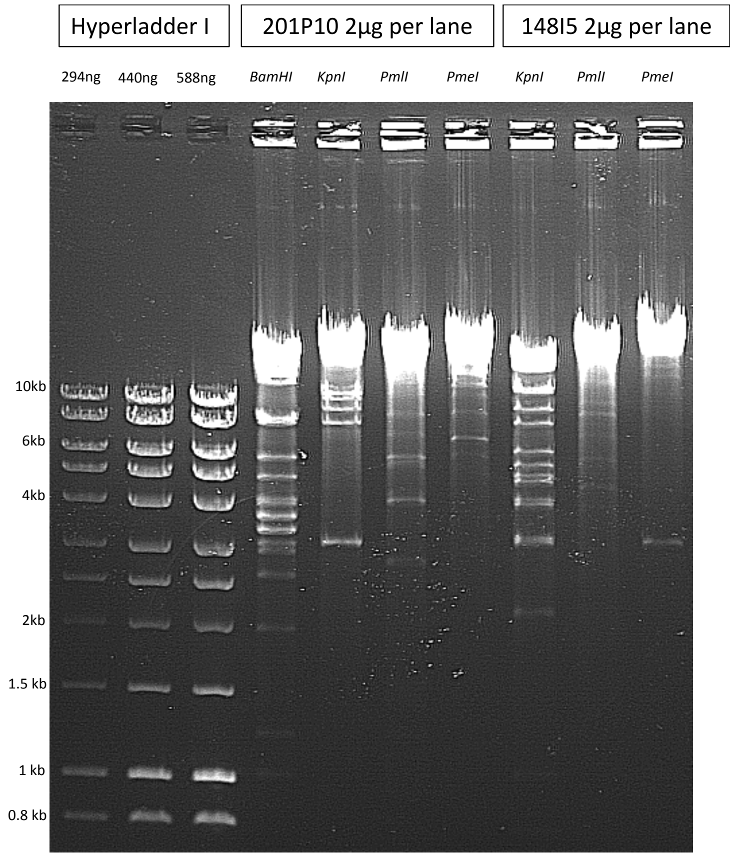


Figure 5.3: restriction patterns observed cutting the clones 201P10 and 148I5, extracted with NucleoBond® BAC 100, with 4 and 3 different restriction enzymes, respectively. 2µg of BAC DNA were used per each digestion. 294ng, 440ng and 588ng of HyperladderI were run in lanes 1,2 and 3 respectively.

5.2.4 Caesium chloride gradient

Given the BAC DNA purity and yield issues raised with the previously described extraction methods, an extraction based on alkaline lysis followed by DNA separation on caesium chloride centrifugation gradient was performed (Section 2.8.5.1). With this method, bacterial DNA contamination and BAC DNA (of smaller size), would have been separated and visible through ethidium bromide staining as two distinct rings after ultracentrifugation. As shown in **Figure 5.4**, this protocol worked perfectly for all the 6 clones, with no bacterial DNA or protein contaminants detected. Before proceeding to BAC DNA removal and cleaning up from the centrifugal unit, each vial has been checked under blue light, to confirm the absence of contaminants in concentrations not detectable with visible light. BAC DNA yield ranged between 20 and 40µg, lower than what obtained with the NucleoBond® BAC 100 system, but with higher purity.

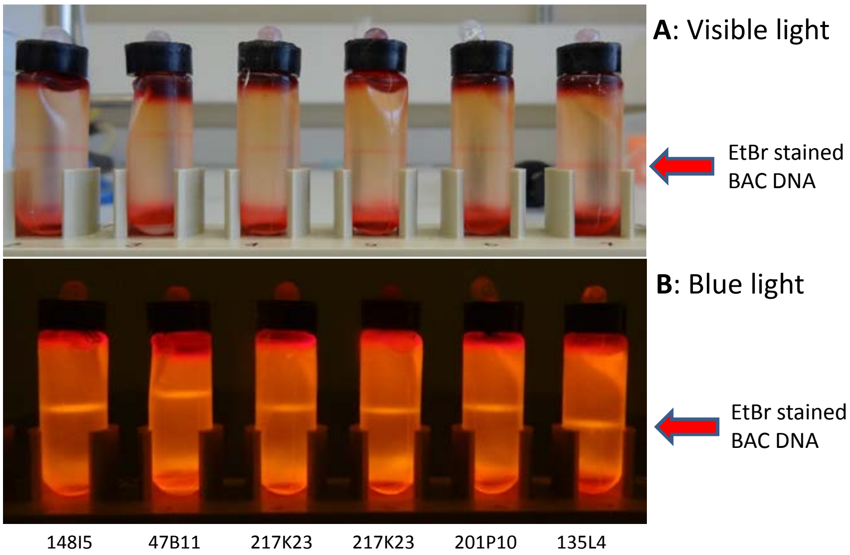


Figure 5.4: Panel **A**: BAC DNA separated with caesium chloride ultracentrifugation, as ethidium bromide-stained ring visible under direct light in the middle of the centrifugal unit. Panel **B**: same vials under blue light, showing the presence of just one DNA ring, indicative of the absence of protein carryover and bacterial DNA contamination.

In order to confirm the absence of bacterial genomic DNA contamination, clones 201P10 and 148I5 have been digested with the same restriction enzymes used in Section 4.2.3, for comparative purposes. Results of the electrophoresis run are shown in **Figure 5.5**.

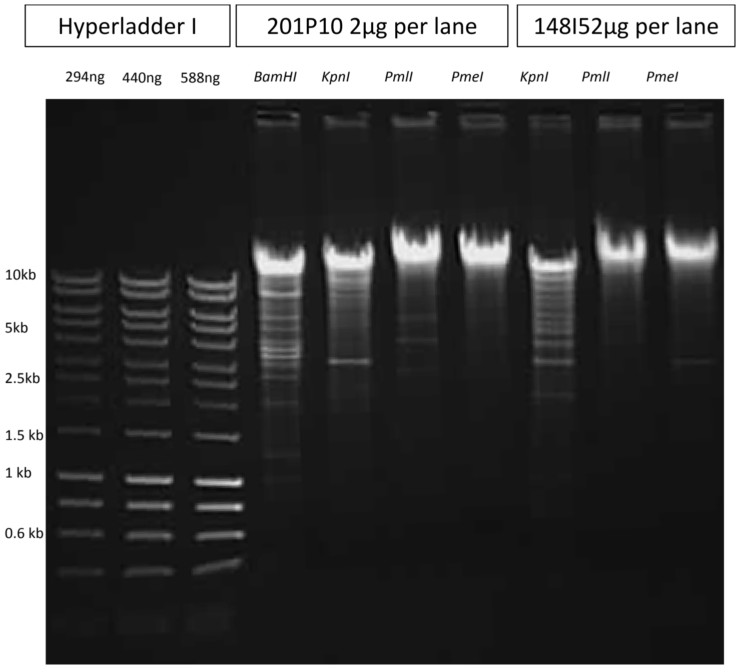


Figure 5.5: restriction patterns observed cutting the clones 201P10 and 148I5, extracted with caesium chloride gradient, with 4 and 3 different restriction enzymes, respectively. 2µg of BAC DNA were used per each digestion. 294ng, 440ng and 588ng of HyperladderI were run in lanes 1,2 and 3 respectively.

Given the good results obtained with the caesium chloride extraction of the BAC DNA for all the six clones available, these samples have been consequently used for the procedures of BAC-end sequencing described in the following sections.

5.3 BAC validation for the β-defensin region

Following BAC DNA extraction, a series of control PCRs have been performed on each clone, with the aim of confirming the presence of the *DEFB2L* gene and gathering information on the presence or absence of other β-defensin genes in the inserts. All the clones resulted positive for the PCR assays Left1 and Right1, used to design the hybridisation probes for the BAC library screening, confirming the accuracy of the clones selection phase. Two other PCR assays have been designed to validate the presence of *DEFB2L* exon 1 and exon 2 in all the clones, with positive results (**Figure 5.6**).

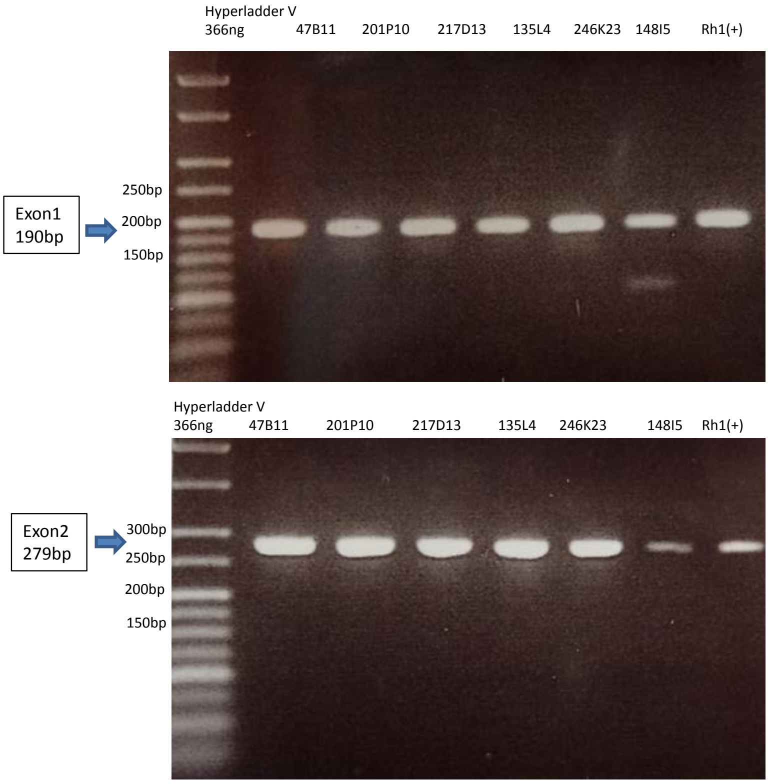


Figure 5.6: *DEFB2L* Exon1 and Exon2 PCR products for all the clones plus a genomic positive control (Rh1) run on 2% agarose gel. 366ng of Hyperladder V were loaded in lane 1.

After validation of the *DEFB2L* presence in all the clones considered, another set of PCR assays has been designed, to determine the presence of other β-defensin genes (For the complete list of the assays and primer pairs used refer to **Table 2.9**. All the assays used gave one specific PCR product of expected size. **Figure 5.7** shows the expected chromosomal localisation for all the assays used.

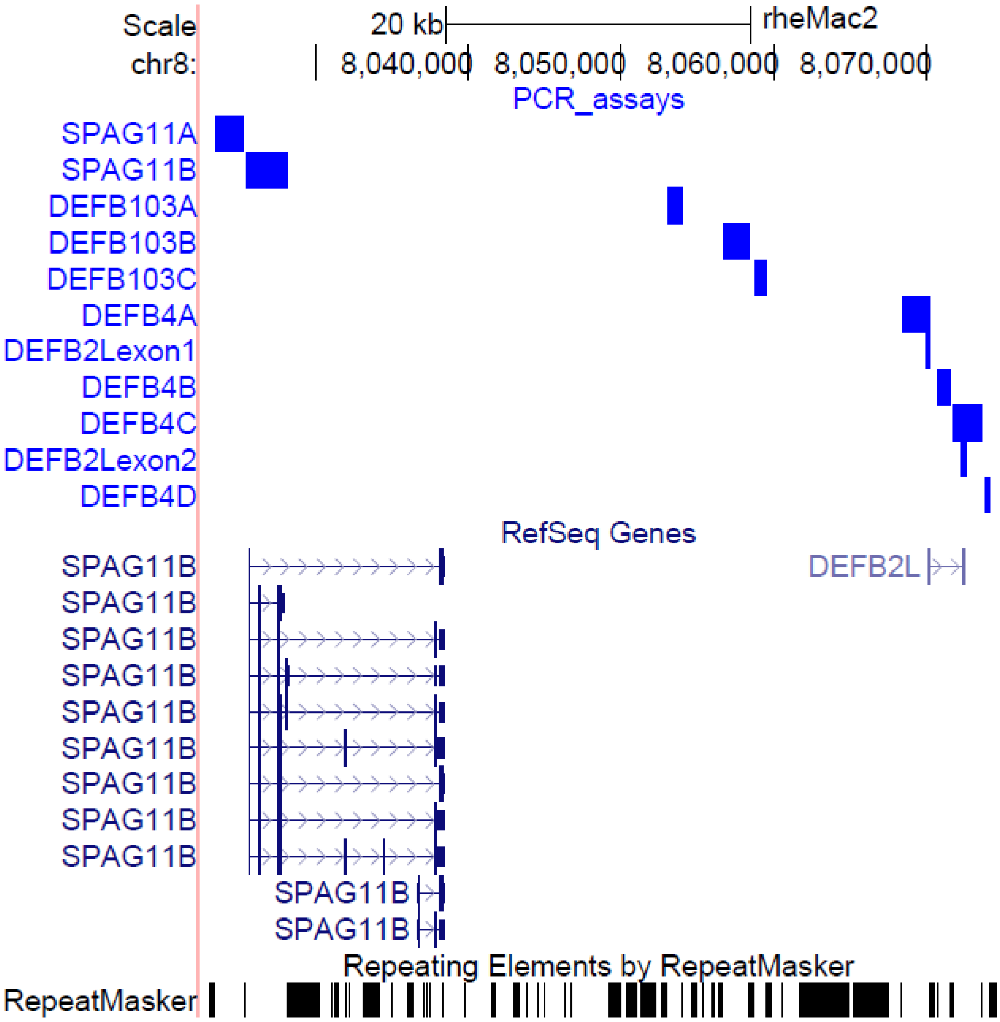


Figure 5.7: localisation of the PCR assays used to obtain informative data on the regions present in each clone insert.

Table 5.4 summarises the results for all the PCR assays performed on the six clones in analysis. From these results, it was possible to confirm that clones 47B11 and 246K23 span all the β-defensin cluster, whereas the others were likely to encompass *DEFB2L* gene and the genomic region centromeric to it.

Clone name	SPAG11 A	SPAG11 B	DEFB 103A	DEFB 103B	DEFB 103C	DEFB4A	DEFB4B	DEFB4C	DEFB4D
201P10	-	-	-	-	-	+	+	+	+
47B11	+	+	+	+	+	+	+	+	+
148I5	-	-	-	-	-	-	+	+	+
135L4	-	-	-	-	-	-	+	+	+
246K23	+	+	+	+	+	+	+	+	+
217D13	-	-	-	-	-	-	+	+	+
Rh1(+)	+	+	+	+	+	+	+	+	+

Table 5.4: Table showing the amplicons present in each of the clone in analysis. **-** indicates absence of PCR product; **+** indicates positive PCR amplification. Rh1 genomic DNA sample has been used as positive control.

5.4 BAC end sequencing

The following step towards a deeper characterisation of the BAC clones isolated has been the sequencing of the insert ends for each clone, in order to map them on the rhesus macaque genomic assembly, evaluating possible size discrepancies. T7 and SP6 promoters, flanking the insert sequence, have been used as priming sites for the sequencing reactions. All the resulting sequence reads have been mapped to the rhesus macaque reference sequence using the BLAT tool of UCSC Genome Browser. Custom tracks for each clone have been generated to calculate the derived insert size according to the reference sequence.

Through the combined data gathered from the control PCR assays, the PFGE for BAC sizing and the BAC-end sequencing, it has been possible to estimate some of the inserts boundaries and their size. **Table 5.5** presents a summary of the information gathered per each clone, with further details presented in the next sections. The relative location of the clones mapping to chromosome 8 is shown in **Figure 5.8**. It is importance to notice how the generation of reliable sequence data for these clones can help to close some of the rhesus macaque assembly gaps of the defensin region that are preventing more accurate analysis of the genomic structure of this complex region.

Clone name	SP6	T7	Insert coordinates	Size	Discrepancies
217D13	+	+	chr8:8072481-8198787	126.3 kb	Smaller size than PFGE
47B11	+	-	Telomeric to chr8:8092024	X	T7 end mapping to chr5 or chr11
201P10	+	+	chr8:8066253-8151005	84.7 kb	Smaller size than PFGE
135L4	+	-	Centromeric to chr8:8169989	X	T7 end mapping in a gap area
148I5	+	+	chr5:131821304-131997988	176.6 kb	Insert mapping on chr5
246K23	+	+	chr8:7957551-8141157	183.6 kb	No discrepancies

Table 5.5: summarising table showing the genomic regions present in each of the clone in analysis. **-** indicates absence of PCR product; **+** indicates positive PCR amplification. Rh1 genomic DNA sample has been used as positive control.

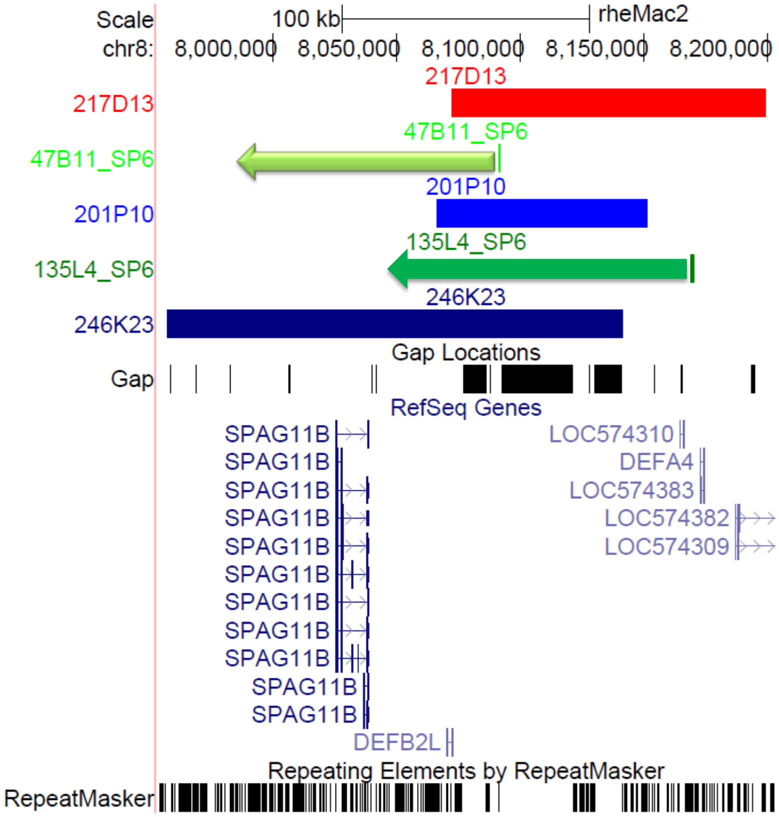


Figure 5.8: Custom UCSC Genome Browser tracks presenting the relative position of the clones mapping on chr8. For clones 47B11 and 135L4 just the SP6 sequence read positions are shown, as their T7-end sequencing gave discordant results; arrows of correspondent colours indicate the direction of the inserts derived from the PCR assays. The clone 148I5 is missing, as its insert boundaries mapped to chromosome 5.

5.4 217D13 insert mapping

For the clone 217D13, the T7 sequence read mapped to the expected position, at the *DEFB2L* start, as previously identified with the PCR assay 'DEFB4B'. SP6 read mapped to two positions, 51.6 kb far apart, in correspondence with the α-defensin locus. As presented in Section 3.3.2, the breakpoint analysis performed on rhesus macaque array CGH data evidenced this region to be variable, with some individuals presenting evidence of duplications. Hence this double hit may be due to the presence in the assembly of two partial copy number units, with the distal one being interrupted by sequence gaps.

According to the reference mapping, clone 217D13 covers the following region: chr8:8072481-8198787 (126.3 kb) or chr8:8072481-8147167 (74.6 kb). **Figure 5.9** presents the two possible sizes as UCSC Genome Browser custom tracks. The results of the PFGE run for this clone indicated an insert size of approximately 190kb, which may indicate an underestimation of the gaps size for the assembled genome.

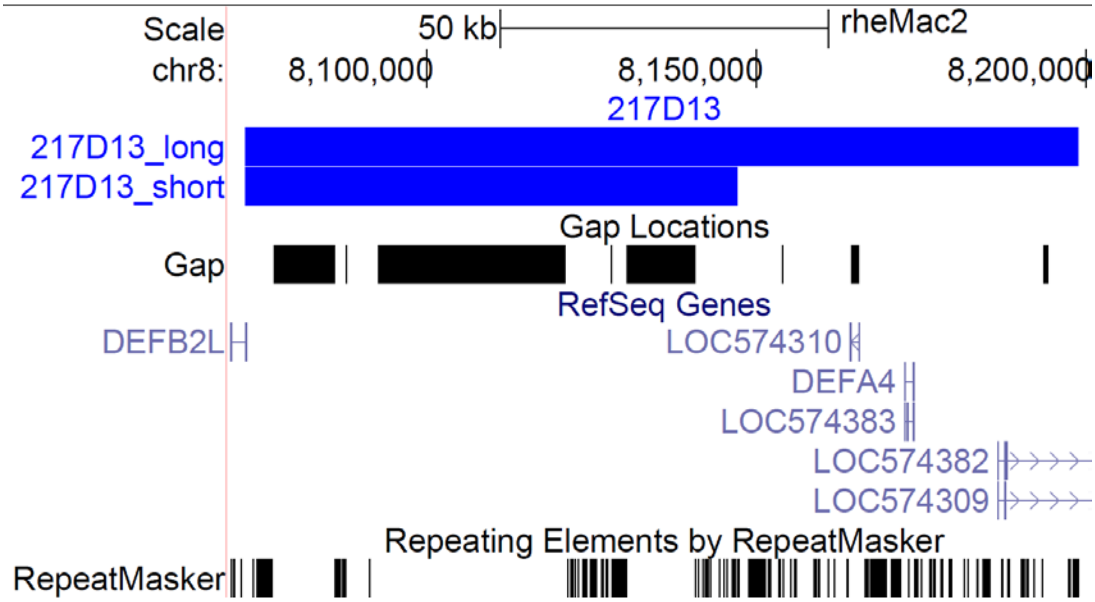


Figure 5.9: UCSC Genome Browser custom tracks (in blue) for the possible mappings of the clone 217D13. The upper track indicates the longer hit of 123.6 kb, while the lower blue track shows the shorter hit of 74.6 kb.

5.4.2 47B11 insert mapping

The sequencing reads for the clone 47B11 did not allow an estimate of the insert size, as the T7 read mapped to chromosome 11. Interestingly, this hit fell in a genic area coding for human salivary proline-rich protein, and orthologous to the family of taste receptors *Tas* in different mammal species, such as mouse, rat and cow (**Figure 5.10**). In rhesus macaque, the *TAS*2 family maps to two clusters, on chromosome 3 and chromosome 11, but possible duplications of this gene cluster on chr8p23 should be further investigated. Nevertheless, the T7 sequence read presents a high degree of identity with a cluster of SINE and LINE elements, present on the mentioned position on chr11 (94.3% identity) and on chr5 (94.9% identity). It is hence possible that the T7 sequence read maps in reality to a non-assembled repeated element on chromosome 8.

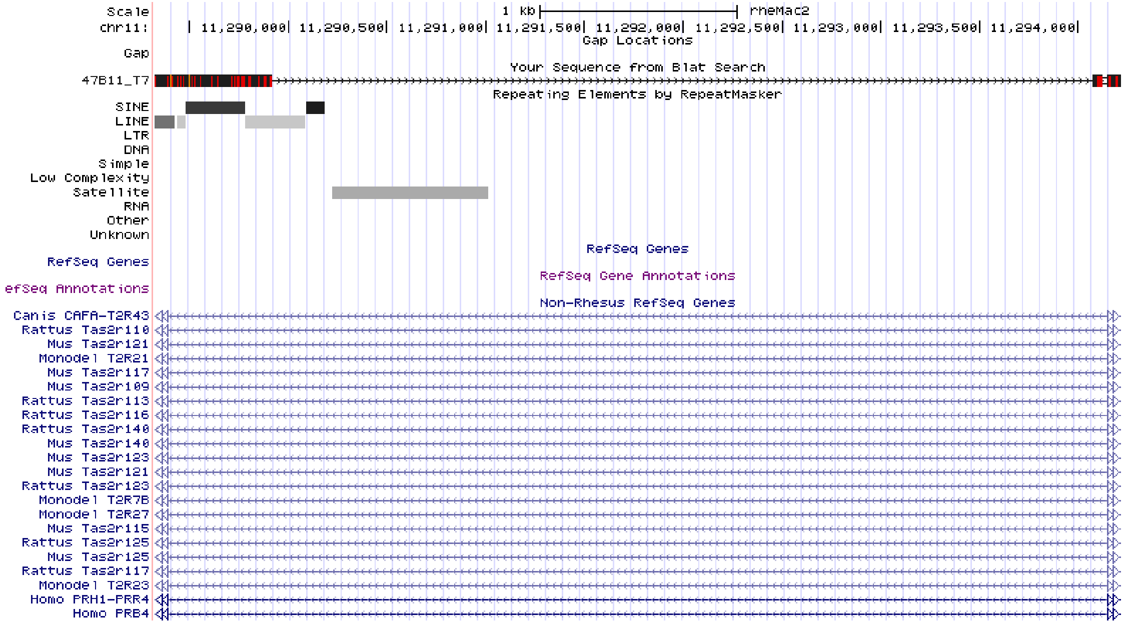


Figure 5.10: T7 sequence hit for the clone 47B11 on chromosome 11. This hit falls in a LINE/SINE cluster, indicated by the repeat masker track. Some of the non-rhesus reference sequences for the family of taste receptors genes are indicated in the lower part of the figure.

The SP6 read had two hits on chromosome 8, as presented in **Figure 5.11**; one distal to the annotated *DEFB2L* gene and the other 22.5 kb away, proximally. The PCR assays DEFB4B, DEFB4C and DEFB4D indicated the presence of at least one full copy of *DEFB2L* in 47B11 insert, and hence the proximal hit was considered to be more accurate.

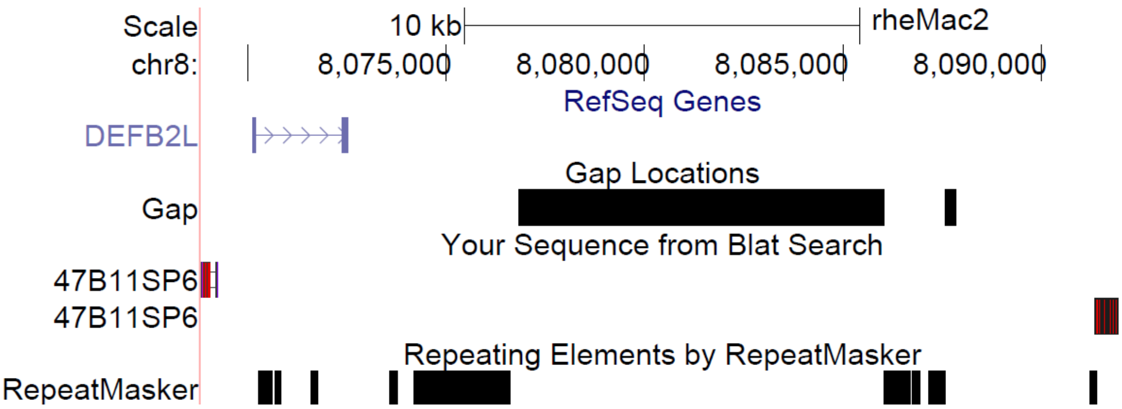


Figure 5.11: SP6 sequence hits for the clone 47B11.

5.4.3 201P10 insert mapping

The T7 read for the clone 201P10 gave two hits: the first at chr8:7091236-7100359 (94.9% identity), the second at chr8:8150102-8151005 (96.3% identity), both partially encompassing repeated elements. Considering that the SP6 read gave one hit on chr8:8066253-8067194, the first T7 hit would have given an unlikely insert size of 975 kb and was hence discarded. If the second T7 hit plus the SP6 read were considered as real insert boundaries, the total insert dimension was calculated to be 84.7kb (**Figure 5.12**). This size, estimated on the assembled rhesus macaque genome, did not correlate with the insert size of approximately 200kb visualised through PFGE for 201P10. These data may indicate an underestimation of the gaps size in the area centromeric to *DEFB2L*. Nevertheless, the PCR assays results for this clone showed perfect concordance with one of the insert boundaries (positive for DEFB4A, DEFB4B, DEFB4C and DEFB4D), falling at the telomeric side of *DEFB2L.*

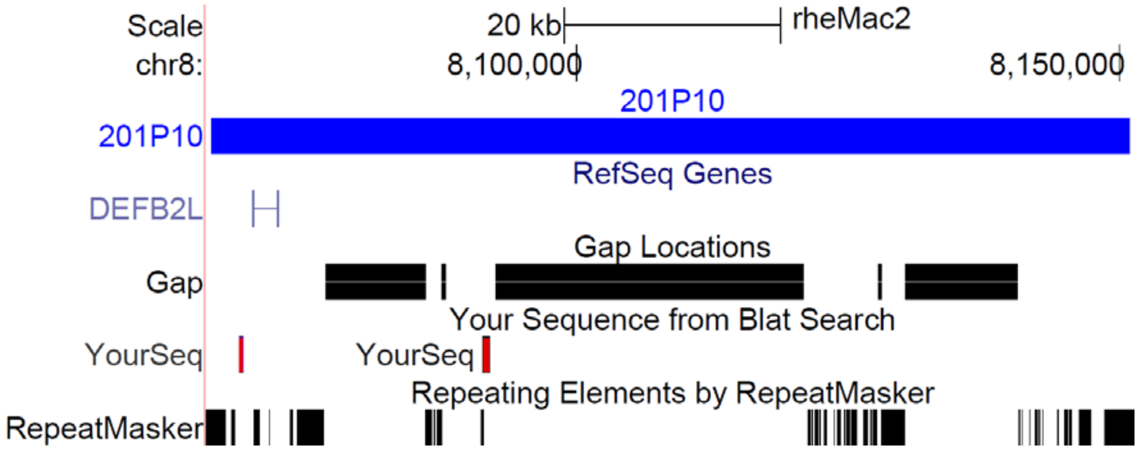


Figure 5.12: UCSC Genome Browser custom tracks (in blue) for the possible mapping of the clone 201P10. The total size of the insert according to its mapped ends is 84.7 kb.

5.4.4 135L4 insert mapping

The SP6 sequence read for the clone 135L4 gave one consistent hit on chr8:8168901-8169989, near the α-defensin cluster (**Figure 5.13**). Conversely, the T7 sequence read hit on a non consistent position on chromosome 8 (chr8:55491596-55491632) with 96.6% identity. As 135L4 showed indication of two *DEFB2L* copies, and the PCR assays DEFB4B, DEFB4C and DEFB4D gave positive results for this clone, it is possible that the T7 sequence read mapped to the area centromeric to the annotated *DEFB2L* gene, falling in a gap area with absence of sequence information. This data may be indicative of the presence of 'hidden' *DEFB2L* copies in the gap areas annotated, that could not be resolved with the shotgun sequencing approach used to generate the rhesus macaque assembly.

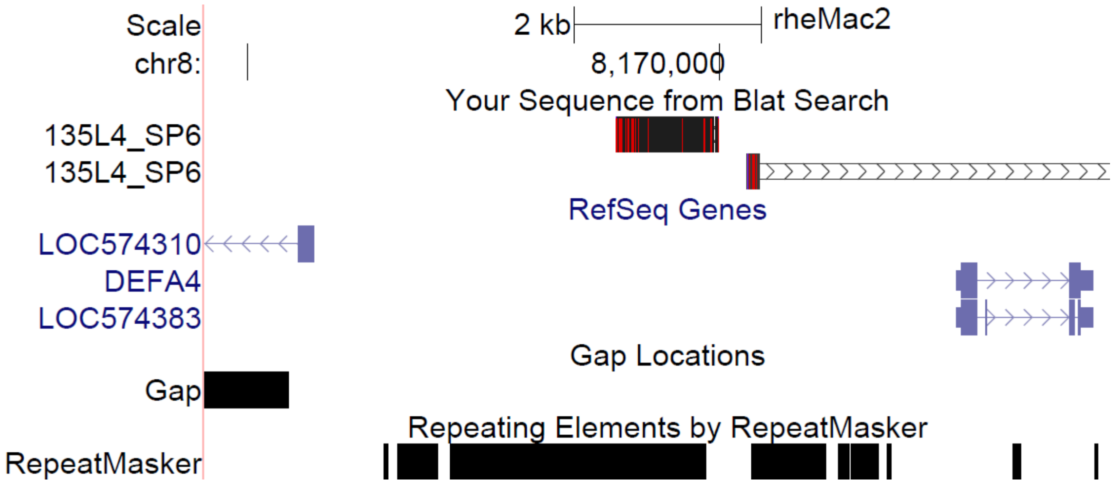


Figure 5.13: SP6 sequence read mapped on the α-defensin cluster on rhesus macaque chromosome 8.

5.4.5 148I5 insert mapping

The SP6 and T7 sequence reads for the clone 148I5 gave unexpected results, with the former hitting on chr5:131821304-131821883 (99.1% identity) and the latter on chr5:131997387-131997988 (99.9% identity); the insert size would hence be 176.6 kb (**Figure 5.14**), reasonably consistent with the PFGE results. Nevertheless, this clone presented two *DEFB2L* variants and gave positive results for the PCR assays DEFB4B, DEFB4C and DEFB4D. Further analysis should investigate if these findings are real, with the insert boundaries falling in a non-assembled area of chromosome 8 with high sequence identity to chromosome 5, or whether there had been a mistake in the clone selection process, with accidental cross contamination between clones.

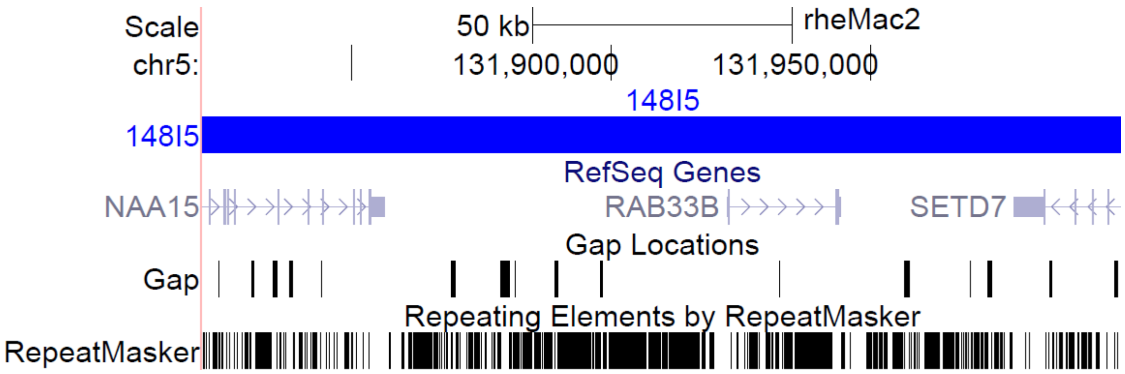


Figure 5.14: UCSC Genome Browser custom tracks (in blue) for the possible mapping of the clone148I5 on rhesus macaque chr5. The total size of the insert according to its mapped ends is 176.6 kb.

5.4.6 246K23 insert mapping

In this case, the BAC end sequencing for the clone 24K23 gave consistent results. SP6 and T7 sequencing reads produced respectively one hit, the former on chr8:8140658-8141157 and the latter on chr8:7957551-7958525. As presented in **Figure 5.15**, the insert size accordingly to the mapping on the assembled rhesus macaque genome was calculated to be 183.6 kb. The insert coordinates are also consistent with the positive PCR assays results for all the genomic areas screened from the β-defensin cluster.

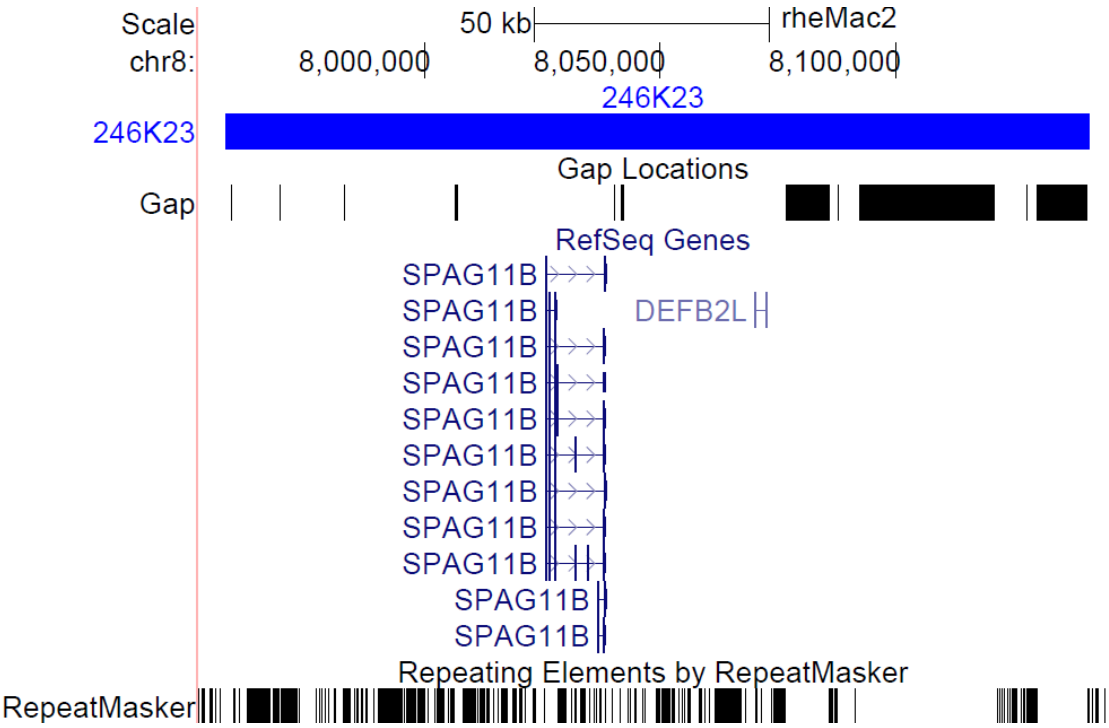


Figure 5.15: UCSC Genome Browser custom tracks (in blue) for the possible mapping of the clone 246K23 on rhesus macaque chr8. The total size of the insert according to its mapped ends is 183.6kb.

5.5 whole BAC sequencing

Three BAC clones have also been fully sequenced by GATC Biotech (Konstanz, Switzerland) using a Pacific Bioscience PacBio platform, based on single-molecule real-time sequencing. This technology, used for *de novo* genome sequencing, has the advantage of generating long sequence reads >3kb and up to 10kb. This allows reduction in the number of contigs generated and hence errors in the assembly phase, especially for complex genomic area that cannot be accurately resolved with sequencing technologies based on smaller reads. This sequencing platform was selected to retrieve high-quality sequence information on the genomic area centromeric to *DEFB2L* which presents extensive gaps in the rhesus macaque sequence assembly, and to get deeper sequence information on the β-defensin cluster. The data generated with the approaches presented in Sections 5.4 and 5.5 were used as selection criteria to select the most informative BACs for sequencing, as in this phase of the project the BAC-end sequencing protocols were still under optimisation.

The choice fell on clone 148I5, as it presented two *DEFB2L* copies with its insert encompassing the genomic area centromeric to *DEFB2L;* 201P10 was selected as it contains the reference variant of *DEFB2L*, coming from the other copy of chromosome 8 than 148L4, with its insert hence providing information on the genomic area centromeric to *DEFB2L* from the other chromosome copy. As last clone, 246K23 was selected, as it spans all of the β-defensin cluster, from *SPAG11* to *DEFB2L*.

It should be noticed that in the sample preparation performed to run single-molecule real time sequencing reaction the insert is not separated from the vector: hence, in case of concordance with the BAC-end reads mapped to the rhesus macaque genomic assembly, all the contigs were expected to be 13397 bp longer than the estimated insert size. Nevertheless, the identity percentage obtained through Megablast alignment between the pTARBAC2.1 complete deposited sequence and each contig could be used as rough estimate of the sequencing accuracy. Furthermore, as PacBio platform generates linear reads, the insert boundaries are not respected, and the contig read may start from any position.

5.5.1 148I5 BAC sequencing

The assembled contig for the clone 148I5 had a size of 191209 bp, hence 14.6 kb longer than the size estimated mapping the BAC-end reads on the rhesus macaque assembly (176.6kb). Megablast alignment with pTARBAC2.1 sequence highlighted a hit of 10.7kb, with 29 miscalled bases and 25 gaps. Hence the whole contig length was estimated to contain an average of 518 miscalled bases and 446 gaps, with an error percentage of 0.5%. The contig size was consistent with the PFGE results and with the BAC-end reads, with an error of 3.9kb.

In order to further investigate this genomic region, the contig sequence has been aligned with the published DNA sequence from the area spanning chr5:131821304-131997988, in order to evidence discrepancies between the two sequences (**Figure 5.16**). From the hit matrix is possible to evidence a great concordance in the sequencing data from 148I5 contig and the correspondent assembled region.

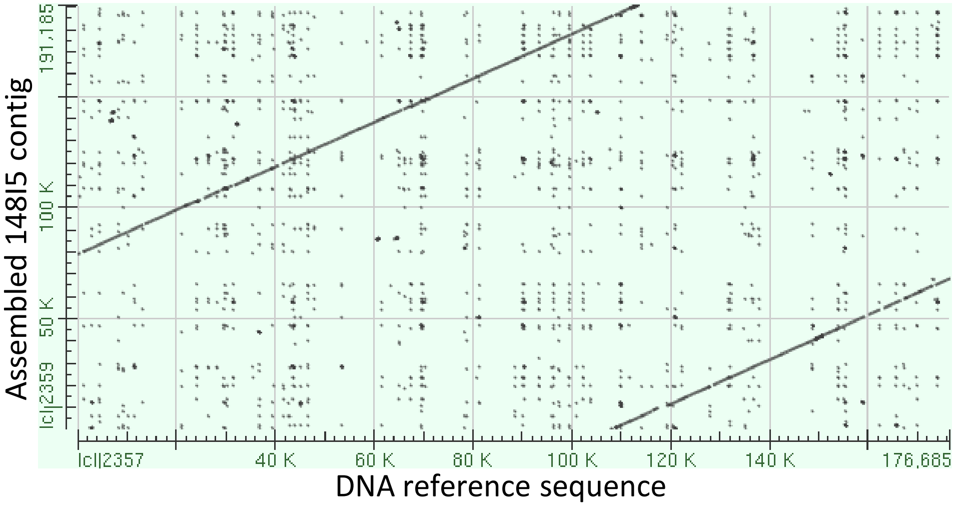


Figure 5.16: Hit matrix generated using the megablast Align tool showing the alignment of the DNA reference sequence correspondent to the insert mapped coordinates for the clone 148I5 (X axis) with the contig generated from the whole BAC sequencing (Y axis).

The hit matrix generated aligning the contig against itself did not evidenced any structural variation in the assembled sequence (**Figure 5.17**).

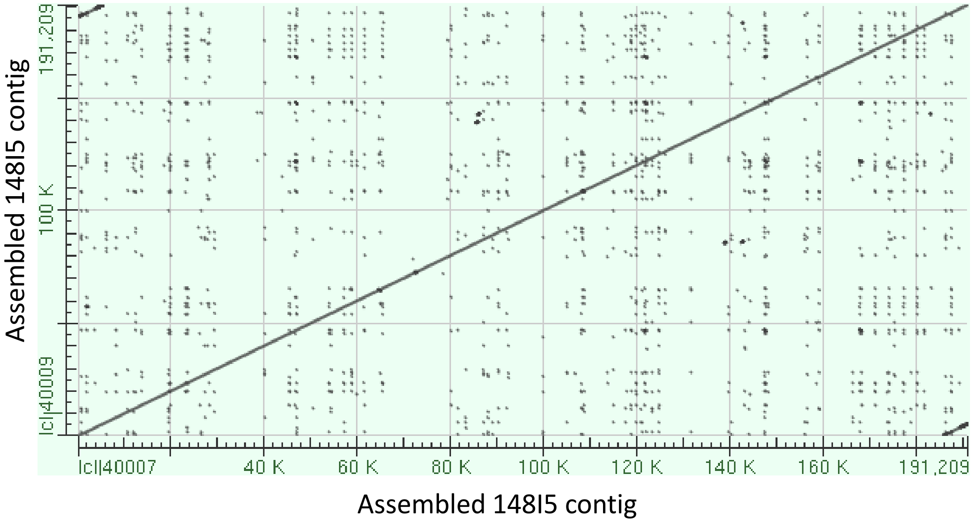


Figure 5.17: Hit matrix showing the alignment of the contig 148I5 against itself, generated using the megablast Align tool.

Any alignment performed with BLAST Align tool to identify the presence of *DEFB2L* gene or other genes of the β-defensin cluster in the contig were negative, demonstrating that there has been a mistake in the BAC screening process, followed by a sample cross-contamination prior to PCR procedures, that led to the unspecific and misleading amplification of chr8-specific genomic areas.

5.5.2 201P10 BAC sequencing

The assembled contig of 201P10 had a size of 96710 bp, compared with the 84.7 kb deduced from BAC-end mapping against the rhesus macaque genomic assembly. Megablast alignment with pTARBAC2.1 sequence highlighted a hit of 8089 bp, with 190 miscalled bases and 170 gaps. Hence the whole contig length was estimated to contain an average of 2271 miscalled bases and 2031 gaps, with an overall error percentage of 4.4%. The contig had hence a size of 88.6kb, with an error of 3.9kb compared with the estimated insert size from BAC-end sequencing. Interestingly, the first 5670 bp of the contig centromeric end did not have any correspondent hit on chromosome 8, whereas increasing the alignment size it was possible to observe a correct mapping from base 5671 onwards.

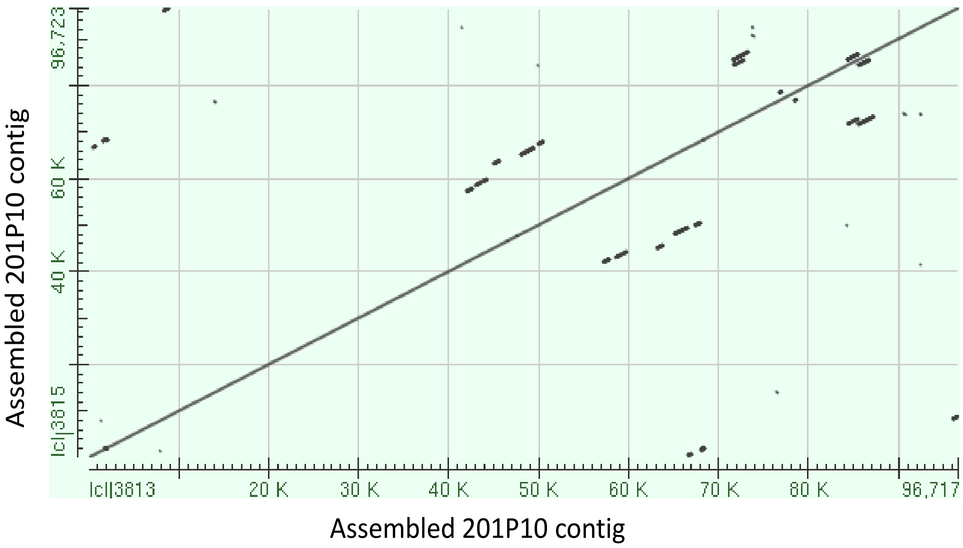


Figure 5.18: Hit matrix showing the alignment of the contig 201P10 against itself, generated using the megablast Align tool.

When aligned against the deposited sequence for *DEFB2L*, it was possible to observe the presence of the full size *DEFB2L* gene in the insert, plus a secondary hit correspondent to 700 bp of the *DEFB2L* intron, not encompassing any annotated repeated element (**Figure 5.19**) .

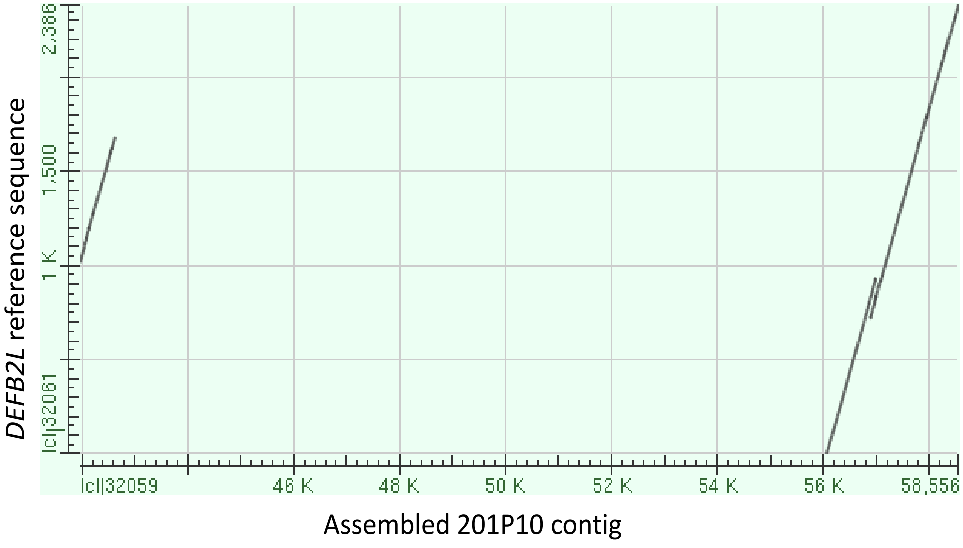
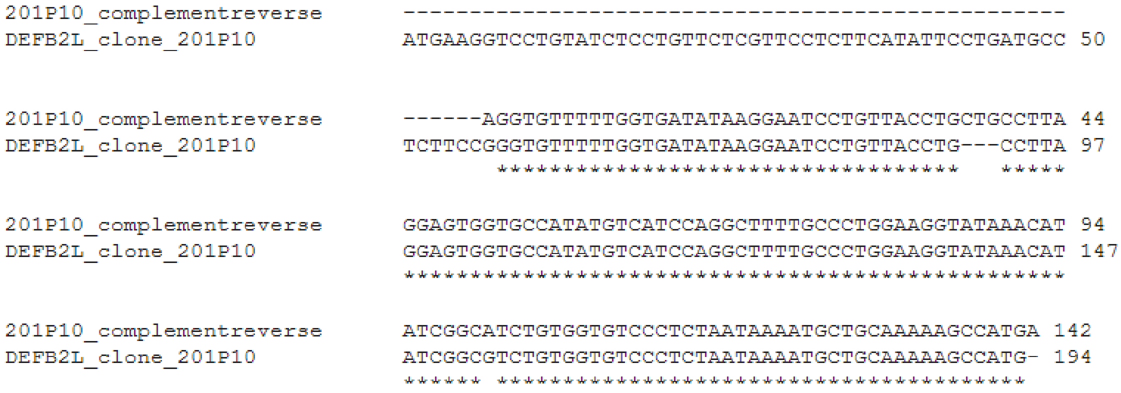


Figure 5.19: Hit matrix showing the alignment of the contig 201P10 (X axis) against the *DEFB2L* sequence (Y axis).

An alignment of the *DEFB2L* sequence from 201P10 contig was then performed against the reference *DEFB2L* sequence, that was the only one expected to be present in the clone, given the sequencing results previously gained on the region, and the BAC Variant *DEFB2L*. Interestingly, the alignment with the reference *DEFB2L* sequence gave the 97% of identity, with the presence of a triplet insertion and of a SNP in the second exon, compared with the reference. If not a sequencing artifact, these base pair changes would be non-synonymous substitutions, generating a cysteine insertion in position 31 of the mature protein, and a valine to isoleucine substitution in position 52 (also reported for the BD2 BAC Variant, as discussed in Section 6)

The alignment of the *DEFB2L* sequence retrieved with Sanger sequencing of 201P10 and the corresponding sequence from the 201P10 contig confirmed that the former sequence was identical to the reference, and the latter presented important base pair changes (**Figure 5.20**). As these changes were not observed in any other clone analysed with Sanger sequencing, it was concluded that the variant positions observed for the contig 201P10 may be sequence artifacts, given the error percentage of 4.4% in sequencing accuracy calculated for this contig.



```
201P10_complementreverse      --------------------------------------------------
DEFB2L_clone_201P10           ATGAAGGTCCTGTATCTCCTGTTCTCGTTCCTCTTCATATTCCTGATGCC 50


201P10_complementreverse      ------AGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCTGCCTTA 44
DEFB2L_clone_201P10           TCTTCCGGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTG---CCTTA 97
                                     ***********************************   *****

201P10_complementreverse      GGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAAGGTATAAACAT 94
DEFB2L_clone_201P10           GGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAAGGTATAAACAT 147
                              **************************************************

201P10_complementreverse      ATCGGCATCTGTGGTGTCCCTCTAATAAAATGCTGCAAAAAGCCATGA 142
DEFB2L_clone_201P10           ATCGGCGTCTGTGGTGTCCCTCTAATAAAATGCTGCAAAAAGCCATG- 194
                              ****** ****************************************
```

Figure 5.20 : Alignment of the *DEFB2L* sequence obtained from Sanger sequencing of 201P10 first and second *DEFB2L* exon, with the correspondent exon2 sequence from the 201P10 contig.

A further analysis has been performed to map the 201P10 contig against its corresponding coordinates in the rhesus macaque assembly. As visible from the hit matrix of **Figure 5.21**, just approximately 40% of the contig is present in the current assembly, with a remaining 60% of novel sequence matching to gaps in the reference genome.

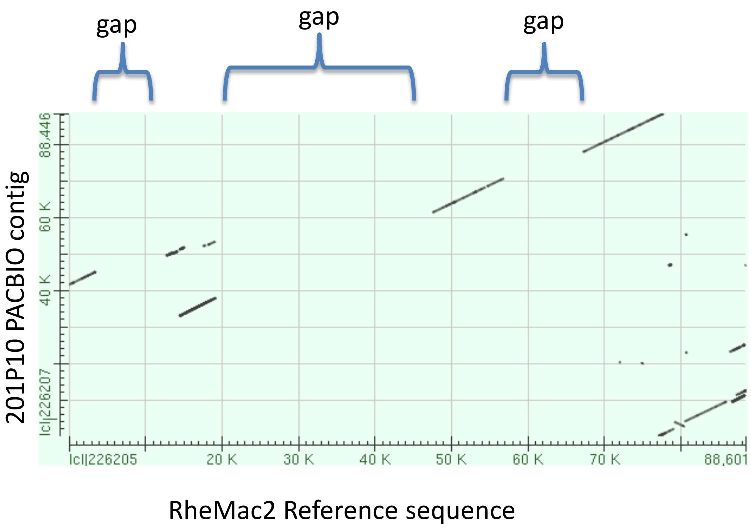


Figure 5.21: Hit matrix showing the alignment of the PACBIO-assembled contig 201P10 (Y axis) against the corresponding coordinates of the RheMac2 genome build (X axis). Gaps in the rhesus macaque assembly are indicated on top of the figure.

5.5.3 246K23 BAC sequencing

The assembled contig size for 246K23 was 169kb, compared with 183.6kb estimated from BAC-end mapping. Megablast alignment with pTARBAC2.1 sequence reported a hit of 10672 bp with extreme sequence fidelity, with 5 miscalled bases and 1 gap. The whole contig length was therefore estimated to contain an average of 79 miscalled bases and 16 gaps, with low error percentage. The contig had hence a size of 158.4kb, with an error of 25.2kb compared with the estimated insert size from BAC-end sequencing. A first alignment was performed between the contig and its correspondent DNA sequence annotated in the rhesus macaque assembly (**Figure 5.22**). This analysis unveiled 10% of novel genome sequence, correspondent to a 10kb gap in the published assembly.

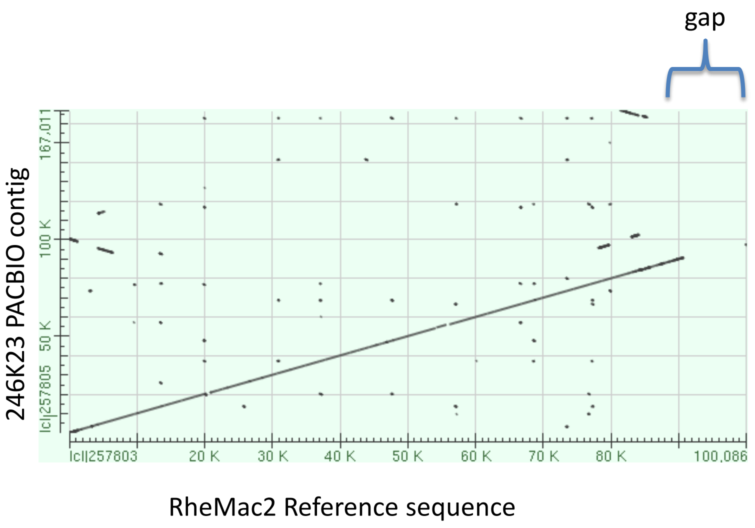


Figure 5.22: Hit matrix showing the alignment of the PACBIO-assembled contig 246K23 (Y axis) against the corresponding coordinates of the RheMac2 genome build (X axis). A 10kb gap in the rhesus macaque assembly is indicated on top of the figure.

In order to check the presence of other possible discrepancies in the *DEFB2L* sequence, an alignment has been performed using Megablast on the contig 246K23 against the reference sequence for *DEFB2L* (**Figure 5.23**). Results gave 100% of identity, showing the presence of a full *DEFB2L* copy in the contig, confirming the results obtained through Sanger sequencing for the same clone. Interestingly, the same secondary hit evidenced from clone 201P10, identifying the presence of 700 bp of sequence matching to part of the *DEFB2L* intron, is also visible from the 246K23 hit matrix plot (**Figure 5.23**), in a position corresponding to the approximate junction point between the two clones. Megablast alignment of the contigs 246K23 and 201P10 highlighted a hit of 29.5kb with 98% of identity between the two clones and 391 gaps (1%) (**Figure 5.24**). These discrepancies could be consistent with the intrinsic sequencing error percentage of the PACBIO sequencing platform, but also with natural sequence polymorphisms between the two copies of chromosome 8. At this stage of the analysis this latter hypothesis could not be eliminated.

Nevertheless, the merged contig sequence information from clones 201P10 and 246K23 may help to close the sequence gaps in the assembly, spanning a total of 217.5 kb non overlapping sequence.

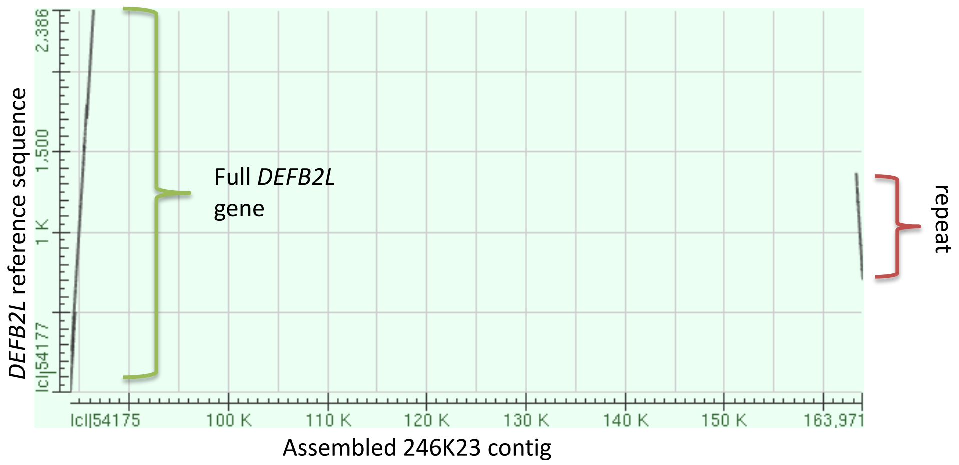


Figure 5.23: Hit matrix showing the alignment of the contig 246K23 (X axis) against the *DEFB2L* sequence (Y axis).

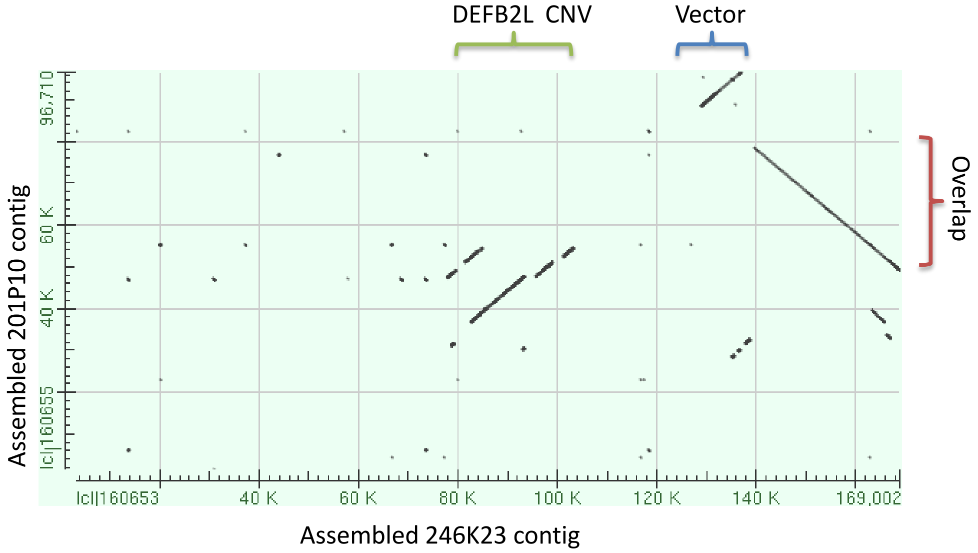


Figure 5.24: Hit matrix showing the alignment generated using the megablast Align tool of the contig 246K23 (X axis) against the contig 201P10 (Y axis). A hit of 29.5kb is produced, indicating the overlapping area between the two clones. Another hit of 8104bp indicates the expected alignment of the vector sequences.

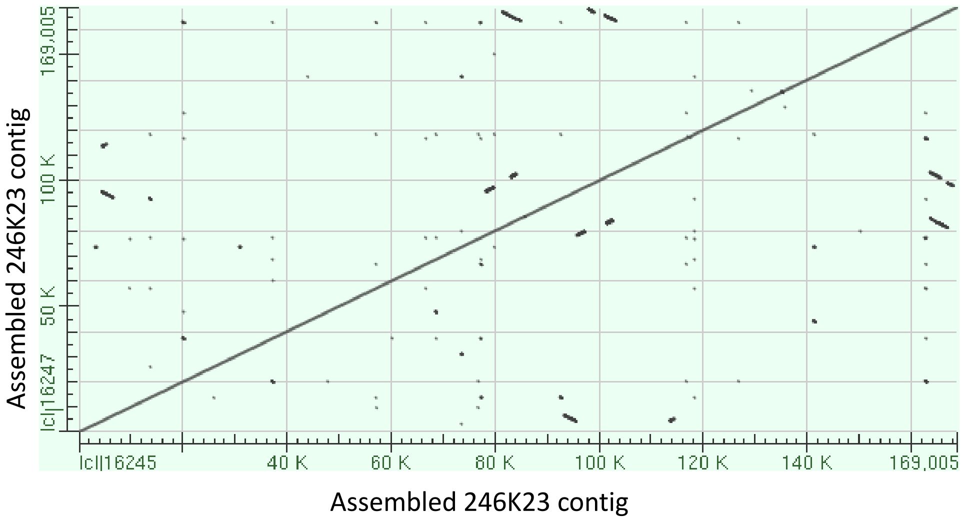


Figure 5.25: Hit matrix showing the alignment of the contig 246K23 against itself, generated using the megablast Align tool.

5.6 Analysis of BAC clones 65I2 and 243E20

Performing a MegaBLAST alignment of the *DEFB2L* sequence against the NCBI nucleotide collection database it was possible to identify the complete sequences of two other clones from the CHORI-250 BAC library: 65I2 (AC193549.4) and 243E20 (AC191454.4, working draft sequence, organised in two separate fragments).

5.6.1 Characterisation of 65I2

The clone 65I2 had a size of 167.3 kb and did not contain the pTARBAC2.1 vector sequence. The alignment of the clone against itself, using BLAST Align tool, revealed the presence of a duplication of 7257bp that reached the very end of the clone (**Figure 5.26**).

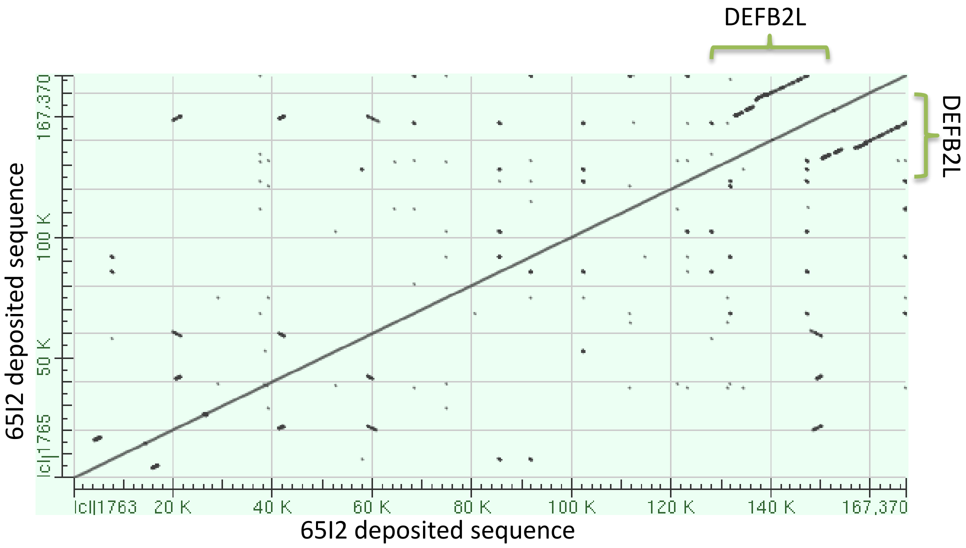


Figure 5.26: Hit matrix showing the alignment of the clone 65I2 against itself, generated using the megablast Align tool. A duplicated area of 7257bp is visible at the right top corner of the hit matrix.

A first alignment was performed to evaluate the presence of full length *DEFB2L* copies (**Figure 5.27**).

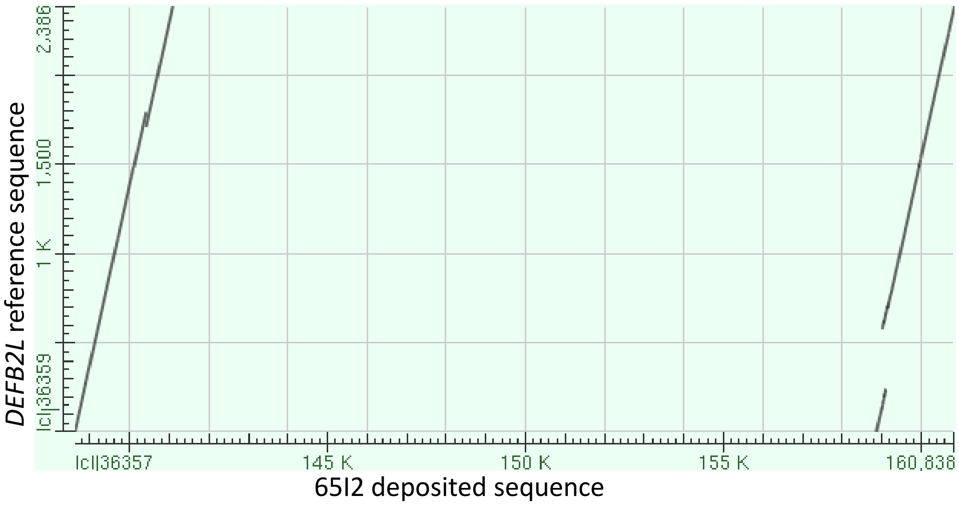


Figure 5.27: hit matrix showing the alignment of the clone 65I2 against the *DEFB2L* reference gene. Two gene copies are present, despite the presence of a gap in the alignment.

Two full length *DEFB2L* copies were evidenced, and the alignment coordinates were retrieved, in order to calculate the median distance between the two *DEFB2L* copies on the clone. The first *DEFB2L* copy aligned to the clone positions 138617-140413, and the second copy aligned to the positions 159009-160838. The average distance between the two copies was calculated to be 20.4 kb, a value that should correspond to the copy number repeat unit.

5.6.2 Characterisation of 243E20

The total length of the assembled clone 243E20 was 171.6 kb whereas the length estimated from the clone end-mapping to the rhesus macaque assembly was 196.3kb. Nevertheless, as visible from **Figure 5.29**, this size discrepancy may be due to the presence of a gap in the sequence assembly. The alignment of the clone against itself produced an unexpected result, highlighting a duplication of 25.3kb (**Figure 5.28**).

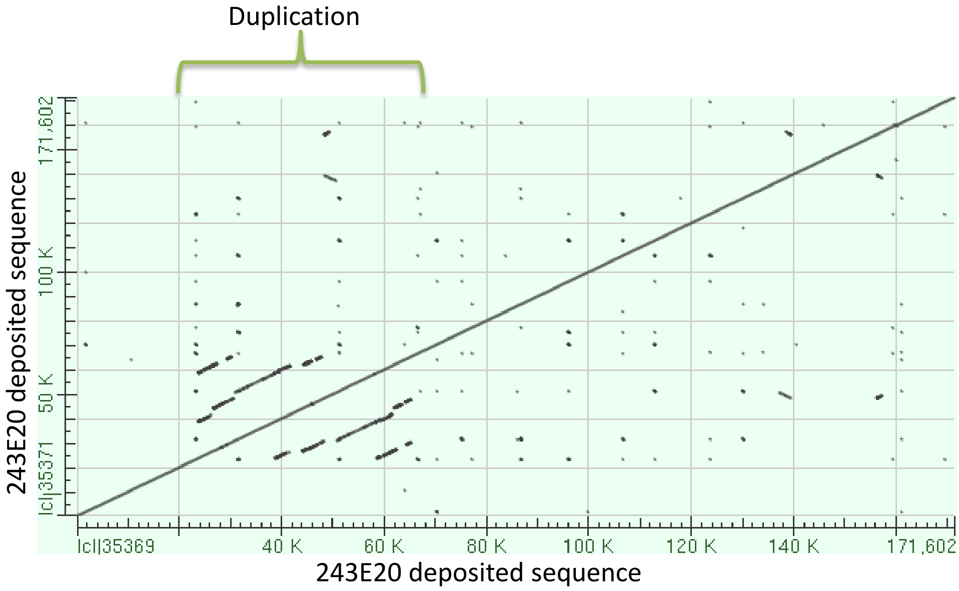


Figure 5.28: Hit matrix showing the alignment of the clone 243E20 against itself, generated using the megablast Align tool. A duplicated area of 25.3 kb is visible.

The duplication mapping on the rhesus macaque reference sequence highlighted an area distal to the *SPAG11* gene, as presented in **Figure 5.29**.

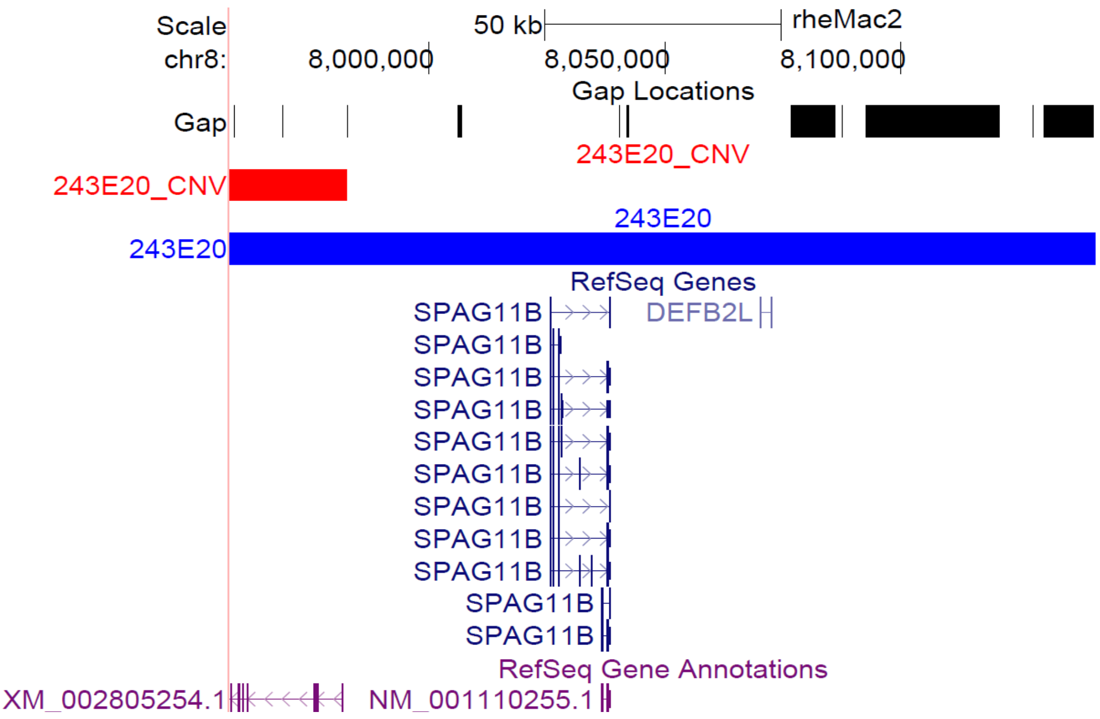


Figure 5.29: Chromosomal coordinates of the clone 243E20 mapped to the UCSC genome browser rhesus macaque assembly (blue track). The red track indicates the position of the duplication (chr8:7982576-7957681, 25.3kb) evidenced through sequence alignment of the clone against itself. The black track indicates the position of gaps in the assembly.

It should be noticed that the 25.3 kb duplication distal to *SPAG11* fell in correspondence of a predicted protein coding gene: XM_002805254.1, with unknown function. Further studies should be made in this direction, in order to confirm and possibly characterise this putative novel copy number.

The alignment of the clone against the *DEFB2L* reference sequence produced two hits, despite small discrepancies in the alignment, as shown in **Figure 5.30.**

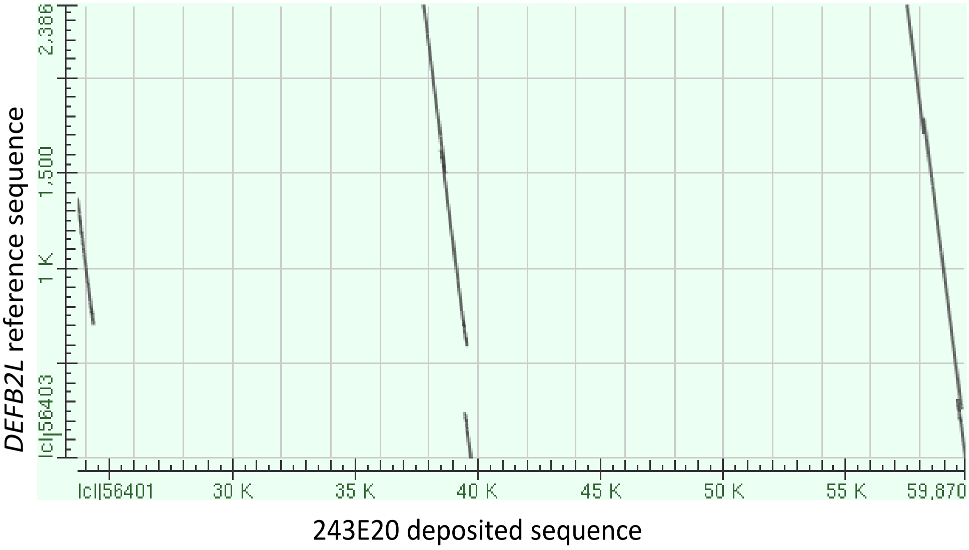


Figure 5.30: Hit matrix showing the alignment of the clone 243E20 against the *DEFB2L* reference gene. Two gene copies are present, despite discrepancies in the alignment.

Through the alignment, it was possible to manually calculate the derived position of the two *DEFB2L* copies on the mapped clone (DEFB2L_CNV1: chr8:8081298-8083674, DEFB2L_CNV2: chr8:8101454-8103378). Interestingly, the two copies fell in exact correspondence with the large sequence gap proximal to *DEFB2L* (**Figure 5.31**). The median distance between the two copies was 19930 bp; hence according to this clone the average copy number repeat unit should be 20 kb, in accordance with the results obtained from clone 65I2.

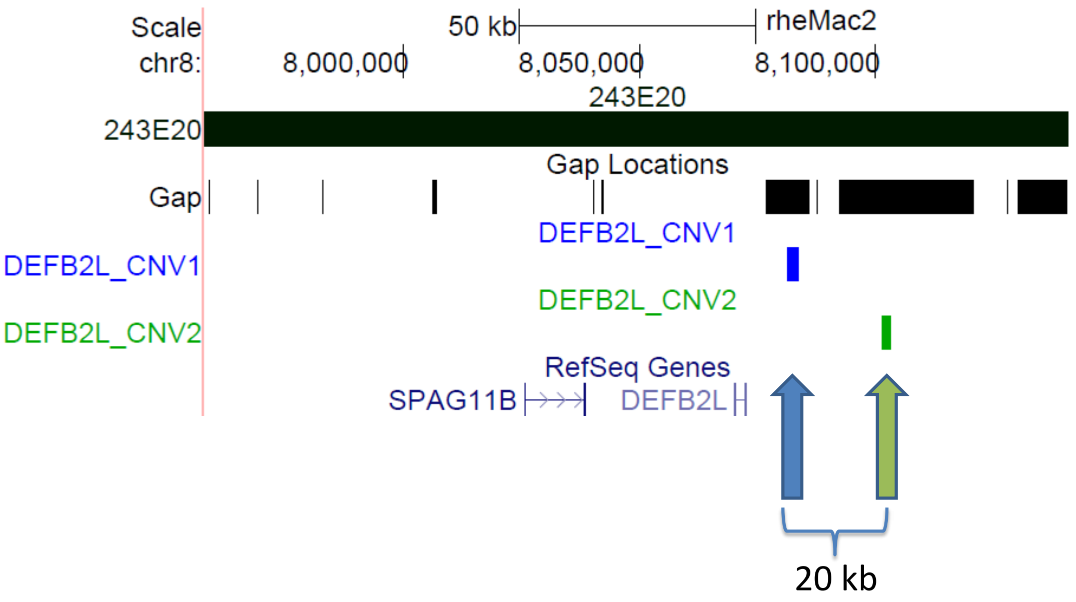


Figure 5.31: calculated coordinates of the two *DEFB2L* copies according to the clone relative position on the assembled reference sequence. The two copies are 20kb apart.

5.7 Cytogenetic mapping of the β-defensin cluster using BAC probes

After characterisation and mapping of the 6 BAC clones with the approaches previously described, the next phase of this project aimed at confirming the physical position of the rhesus macaque β-defensin cluster on chr8p23.1. In human, β-defensin genes are organised into three main clusters: 8p23.1, 20p13 and 20q11.1, with another probable small cluster on chromosome 6p12 [131], arisen as a consequence of different rounds of duplication and divergence of the β-defensin family members; hence, in order to confirm the specificity of the PCR-based assays described in Chapter 4 and to guarantee the copy number call accuracy, it was necessary to exclude the possibility of a *DEFB2L* duplication outside the expected position on chr8p23.1. With this aim, metaphase spreads were generated from lymphoblastoid rhesus macaque cell lines, to perform FISH experiments using fluorescently-labelled selected BAC clones as probes for the genomic area spanning *DEFB2L*. In order to identify chr8 among the other metaphase chromosome pairs, a commercial human chromosome 8 painting probe was used, as available orthologous to rhesus macaque chr8.

5.7.1 Probe generation

Accordingly to **Table 5.5** the clones 47B11 and 201P10 were selected for labelling, as they covered the genomic area distal and proximal to *DEFB2L.* 50ng of BAC DNA from each clone were amplified through a whole genome amplification (WGA) protocol and amplified fragments from 47B11 and 201P10 were then purified through Amicon centrifugal devices, high-recovery cellulose membrane-based cartridges used for DNA concentration and purification (**Figure 5.32**).

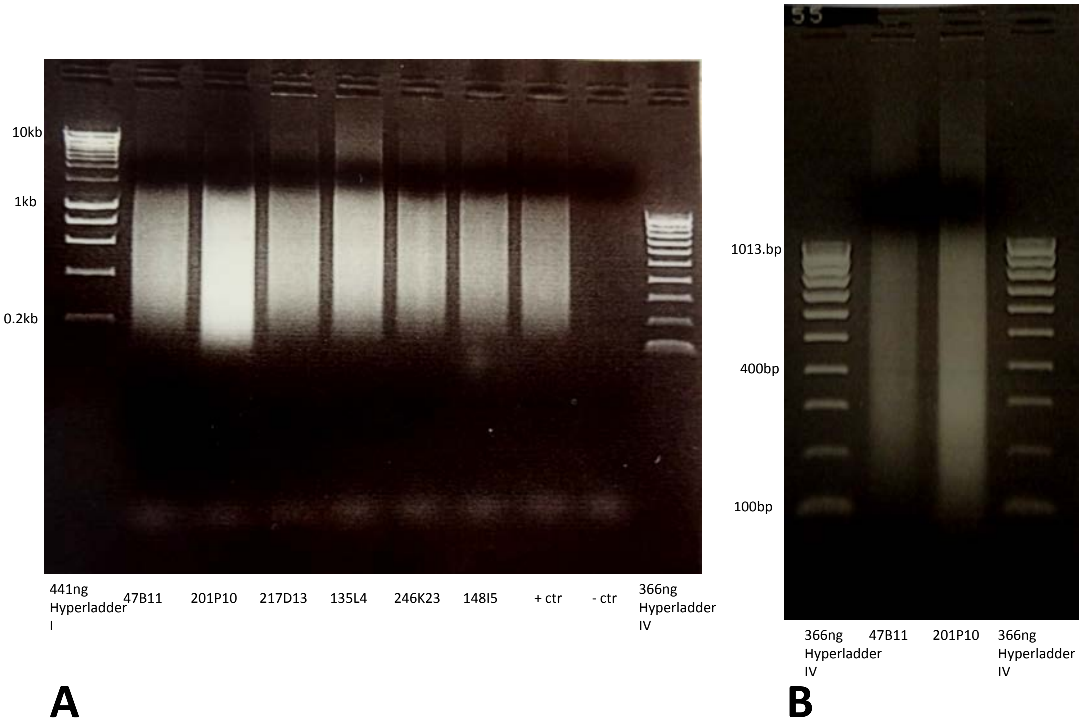


Figure 5.32: A: 8μl of WGA products for each BAC run on 1.5% agarose gel. **B:** 5μl of cleaned WGA product from the clones 47B11 and 201P10 run on a 1.5% agarose gel.

Each cleaned WGA amplified clone was separately labelled with biotin (500ng of template) and digoxygenin (1μg of template) and purified using a column-based kit. A test for the incorporation of the labelled nucleotides was performed by blotting 1μl of each probe on blotting membrane and applying a mixture of anti-digoxygenin and anti-biotin antibodies conjugated to alkaline phosphatase (AP).

5.7.2 Metaphase FISH

Following methapase spread preparation from 6 rhesus macaque lymphoblastoid cell lines, as described in Section 2.10.4, slides were hybridised with either digoxygenin-labelled 201P10 clone (emitting in the green field) and chromosome 8 paint (emitting in the red field) or digoxygenin-labelled 47B11 clone and chromosome 8 paint. With both probe combinations it was possible to observe a clear signal from the short arm of chromosome 8, identifying the β-defensin cluster (**Figure 5.33** and **Figure 5.34**). In **Figure 5.34** it is also possible to observe the nuclear localisation of the two copies of chromosome 8 in interphase nuclei. Nevertheless, different other secondary hybridisation sites were present, for both of the BAC probes, in all the slides analysed, requiring further investigations.

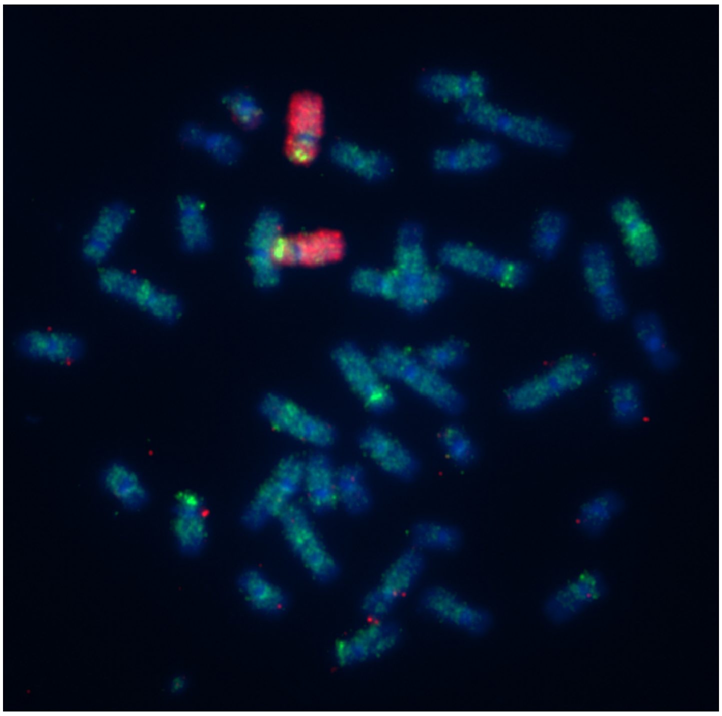


Figure 5.33: Example of metaphase FISH performed on metaphase chromosome spreads from the rhesus macaque lymphoblastoid cell line 2BX. Chromosomes are stained in DAPI (blue); In red is shown the emission signal from chromosome 8 painting; digoxygenin-labelled 201P10 clone emits in the green field and presented secondary hybridisation sites.

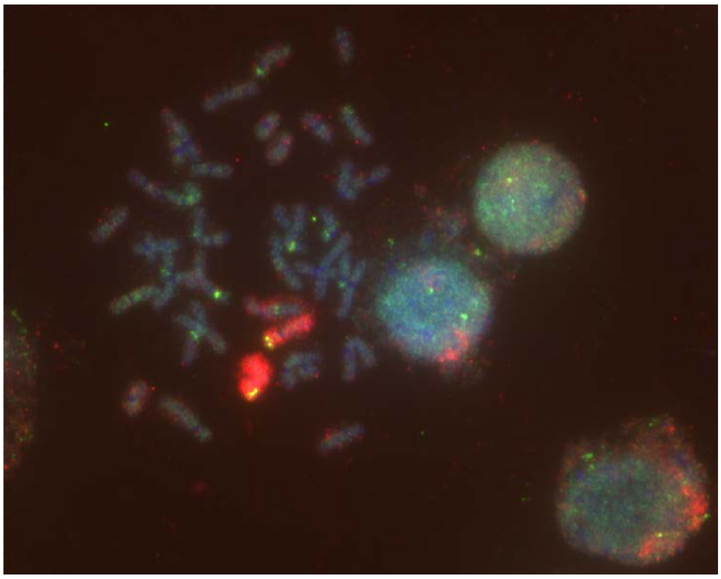


Figure 5.34: Example of FISH performed on metaphase spreads from the rhesus macaque lymphoblastoid cell line r00068. In red is shown the emission signal from chromosome 8 painting; digoxygenin-labelled 47B11 clone emits in the green field and presented secondary hybridisation sites. Chromosomes are stained with DAPI (blue). Three interphase nuclei are present in the field, two of which showing the localisation of chromosome 8.

The results obtained from this first set of FISH were not conclusive for a precise localisation of the β-defensin cluster. The presence of different hybridisation sites from two independent BACs could have implied a consistent grade of non-specific binding, or a real duplication signal. In particular, it was reasoned that the procedure of probe labelling was performed using the Klenow Fragment of the DNA polymerase I, which generates short fragments; hence labelled short products spanning repeated elements present in the BAC insert could have contributed to the generation of non-specific signal, as they were likely to bind to multiple chromosomal sites. In order to test this hypothesis, another set of FISH experiments was performed, adding 1μg of human placental DNA (Cot-1) per slide, in each hybridisation reaction, to help mask repeated elements and increase the probe specificity. This approach gave positive results, as shown in **Figure 5.35**, with just one specific signal detected from the hybridisation of digoxygenin-labelled 47B11 and from digoxygenin-labelled 201P10, co localising with the two chromosome 8 copies, at the telomeric end of the short arm.

With this approach it was possible to confirm the localisation of the β-defensin cluster at the expected position, on the short arm of chromosome 8, with no signs of inter-chromosomal duplications.

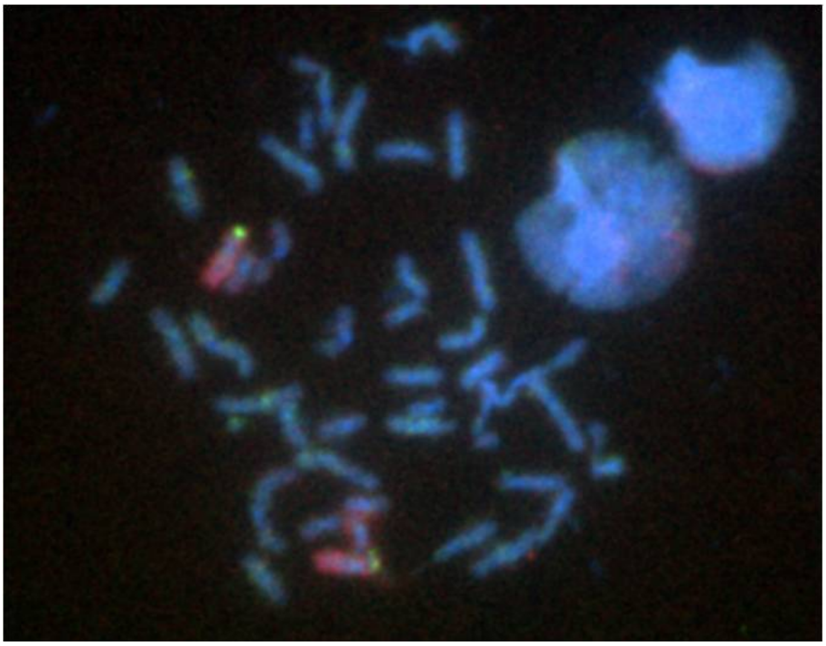


Figure 5.35: Example of metaphase FISH performed on metaphase spreads from the rhesus macaque lymphoblastoid cell line 2BZ. Chromosomes are DAPI-stained. In red is shown the emission signal from chromosome 8 painting; digoxygenin-labelled 47B11 clone emits in the green field, giving just two specific signals on the short arm of chromosome 8. Non-specific probe binding was blocked with 1μg of Cot-1 DNA per each slide.

5.8 Chapter summary

This chapter describes the procedures used for the screening of a rhesus macaque BAC library, with the aim of identifying clones positive for the gene *DEFB2L*. A total of 6 clones have been selected and different approaches and challenges for the extraction of high-yield and high-purity BAC DNA from those are here presented.

Three of the clones have been fully sequenced, in order to generate clone contigs with long reads, to try and fill the gaps in the sequence assembly for the rhesus macaque β-defensin cluster. In particular, the sequence information coming from clones 201P10 and 246K23, likely to have inserts with overlapping ends mapping to the same chromosome copy, could be used as new reference sequence for an area spanning 217.5 kb.

The sequence analysis of two clones deposited online, 65I2 (AC193549.4) and 243E20 (AC191454.4) evidenced the presence of two copies of DEFB2L in each clone. The calculation of the median distance between the two copies of the gene allowed the estimation of a copy number unit repeat size of 20 kb.

A successful cytogenetic mapping was performed to confirm the expected position of the β-defensin cluster just at chr8p, and no inter-chromosomal duplications involving *DEFB2L* were observed.

Characterisation of DEFB2L variants

This section presents the characterisation of a novel variant of *DEFB2L* and a comparative analysis of another five variants (named *DEFB2L1-5*) predicted from sequencing of cDNA clones amplified from rhesus macaque gastric tissue by Dr Mike Hornsby and Prof Charles Bevins (University of California Davis, US) [208] and deposited online on UCSC Genome Browser (EU090139-EU090143).

As there are no reported sequence variations for human hBD2, the discovery of putative BD2 variants in rhesus macaque raised the possibility of an alternative evolutionary pattern undertaken after the separation of Old World Monkeys from the human lineage. For this reason, a deeper characterisation of the variants identified was necessary, to understand whether they were accumulating synonymous mutations under neutral selection or whether they may determine functional changes with phenotypic effects.

6.1 *DEFB2L* sequencing from BAC clones

The BAC clones described in Chapter 5 were used as template to sequence the *DEFB2L* coding region, with the aim of identifying possible variants in the *DEFB2L* coding sequence. For each clone, the first and the second exon of *DEFB2L* were sequenced from both directions, using the same primers validated in section 5.4 from the PCR assays 'Exon1' and 'Exon2'.

6.1.1 *DEFB2L* exon 1 sequencing

The sequence reads obtained from the first *DEFB2L* exon, sequencing Exon1 PCR product from both ends, were consistent for all the clones and did not highlight variant positions compared with the rhesus macaque assembly. This finding may be consistent with the fact that exon1 codes for the signal peptide of BD2 protein, subjected to cleavage before the release of the mature protein form. It is probable that, in order to maintain its signalling activity, this area may be strongly conserved, as previously reported [141].

6.1.2 *DEFB2L* exon 2 sequencing

Conversely, the sequence reads retrieved from the second *DEFB2L* exon, sequencing Exon2 PCR product from both ends, revealed the presence of variant positions for 3 clones out of six, compared with the annotated *DEFB2L* reference. These variant sites were called consistently among the clones 135L4, 148I5 and 217D13, with the same base pair changes detected in reads from both directions. An example of chromatographic profile showing a variable site is presented in **Figure 6.1**. It should be noticed that BACs provide haploid information, as their insert will have come from one chromosome copy or the other. Hence variant positions, when confirmed not to be sequence artefacts, are reflective of sequence changes between two paralogue copies. In the case in analysis, the two sequences were separated subtracting the correspondent single base pair call from the other chromosomal copy (as shown in the lower chromatogram of **Figure 6.1**).

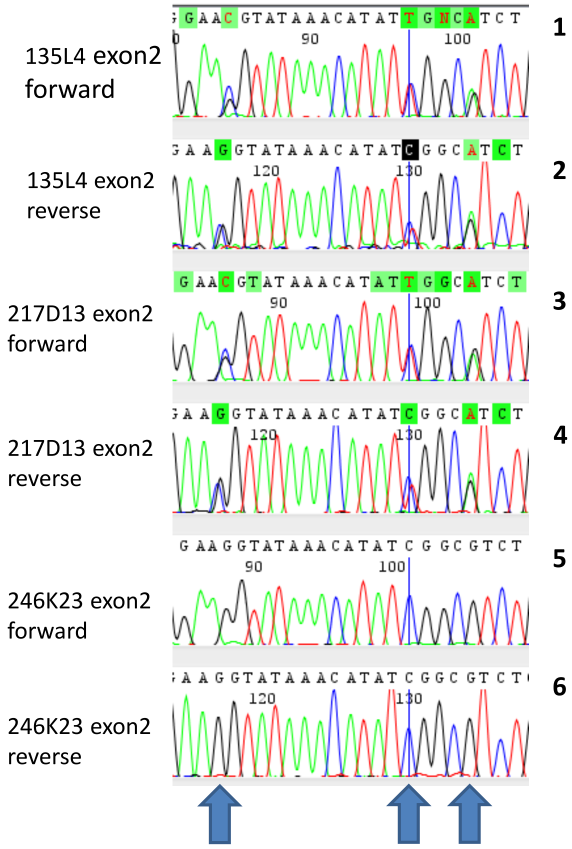


Figure 6.1: Chromatograms showing three variable sites (blue arrows) found in the second exon of *DEFB2L* from the clone 135L4 (**1** and **2**) and 217D13 (**3** and **4**) with consistent base pair calling from both sequencing directions. Chromatograms **5** and **6** show the correspondent sequence from the clone 201P10, that does not show variable sites.

These results brought to infer a novel sequence variant of *DEFB2L* that in the rest of the project will be called 'BAC variant'.

However, it is necessary to report that in the course of this work it was not possible to determine unequivocally the phase of the *DEFB2L* copies hypothesised from BAC sequencing, as the presence of variants with intermediate changes between the reference *DEFB2L* and the BAC variant could not be excluded. The remaining part of this study will address the characterisation of the most extreme case, in which all the paralogous sequence changes are attributed to a different sequence variant, awaiting for further studies to confirm this working assumption.

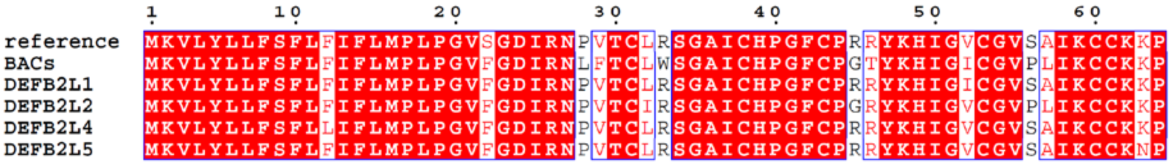
6.2. Protein variants alignment and structural analysis

For all the *DEFB2L* DNA sequences variants, coding sequence files were retrieved through UCSC Genome Browser; for the BAC Variant, the coding sequence was generated through pairwise alignment with the deposited *DEFB2L* mRNA sequence (NM_001128851.1). Percentage of identity with the rhesus macaque reference sequence and number of nucleotide substitutions are indicated in **Table 6.1**.

BD2 variant name	Percentage of identity	Number of nucleotide substitutions compared with the reference sequence	Number of amino acid changed compared with the reference
Reference	100%	0	0
BAC Variant	94%	12	9
DEFB2L1	99%	2	2
DEFB2L2	96%	7	5
DEFB2L3	100%	0	0
DEFB2L4	98%	3	3
DEFB2L5	99%	2	2

Table 6.1: Percentage of identity and number of nucleotide substitutions obtained through pairwise alignment of all rhesus macaque *DEFB2L* coding sequence variants with the annotated rhesus macaque mRNA reference.

From this alignment, it was possible to observe that the variant DEFB2L3 was identical to the reference, and was hence excluded from the analysis. All the resulting output files were translated using ExPASy Translate tool (http://web.expasy.org/translate/) [209] and aligned with ClustalO (http://www.ebi.ac.uk/Tools/msa/clustalo/) [210]. Graphic representations of the levels of conservation of the different protein sequences are shown in **Figure 6.2**. It should be noticed that amino acids between position 1 and 27 constitute the BD2 signal peptide, cleaved before the release of the mature protein form, which contains 37 residues.

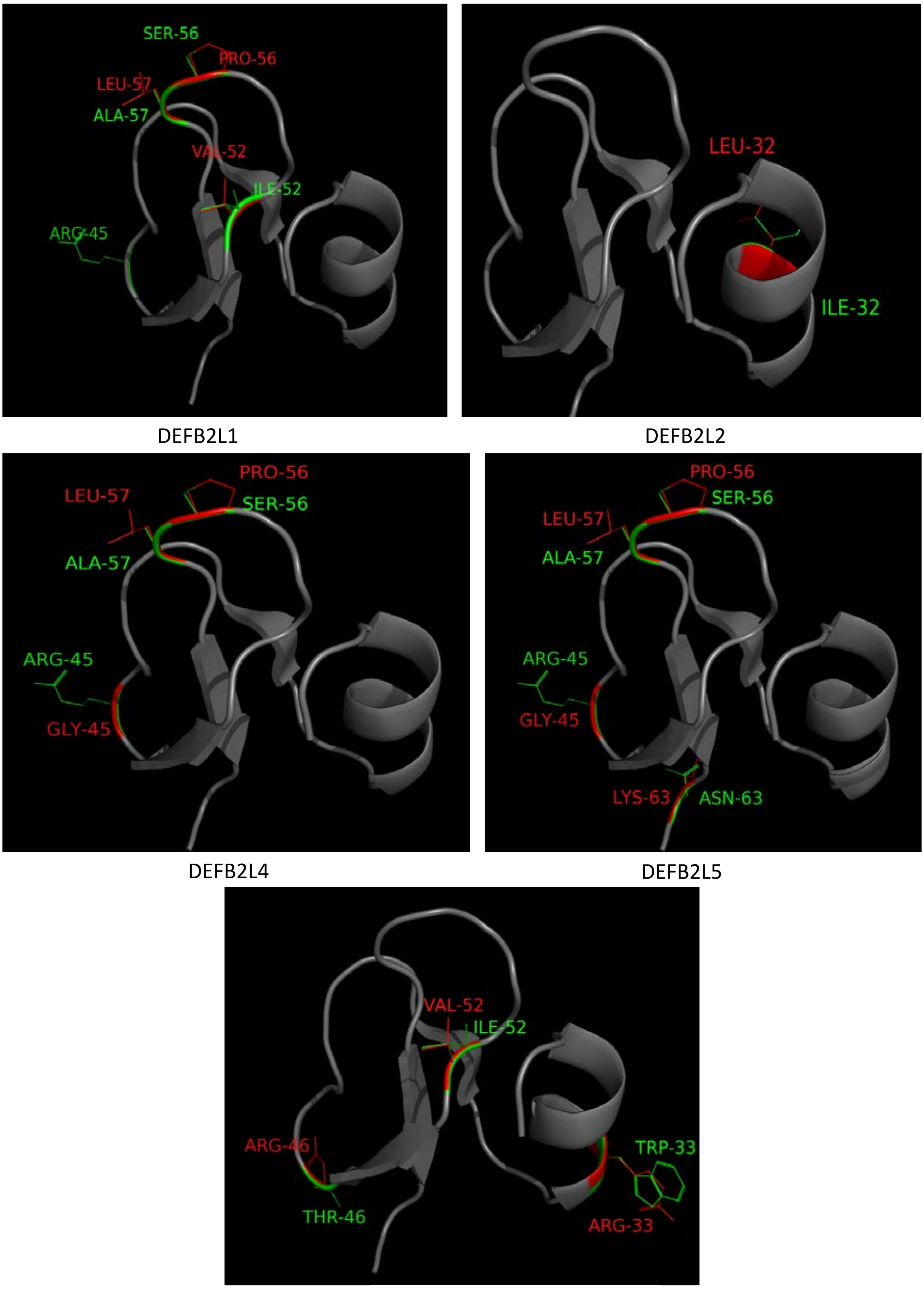


```
            1         10        20        30        40        50        60
reference   MKVLYLLFSFLFIFLMPLPGVSGDIRNPVTCLRSGAICHPGFCPRRYKHIGVCGVSAIKCCKKP
BACs        MKVLYLLFSFLFIFLMPLPGVFGDIRNLFTCLWSGAICHPGFCPGTYKHIGICGVPLIKCCKKP
DEFB2L1     MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPRRYKHIGICGVSAIKCCKKP
DEFB2L2     MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCIRSGAICHPGFCPGRYKHIGVCGVPLIKCCKKP
DEFB2L4     MKVLYLLFSFLLIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPRRYKHIGVCGVSAIKCCKKP
DEFB2L5     MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPRRYKHIGVCGVSAIKCCKNP
```

Figure 6.2: Protein sequence alignment for all the reported rhesus macaque BD2 variants. Amino acid consensus is shown in red; amino acid positions in red font indicate non-synonymous substitutions between two amino acids with similar properties; amino acid positions in black font indicate non-synonymous substitutions, for amino acids with different properties

From the protein alignment it was possible to observe the presence of 12 non-synonymous substitutions across the six variants, four of which with changes to amino acids with different properties. Positions 45 (Gly→Arg), 52 (Val→Ile), 56 (Pro→Ser) and 57 (Leu→Ala) showed amino acid substitutions in more than one variant, suggesting the presence of positions more prone to variation.

Structure models of each variant were generated in PyMOL, to visualise the position of the amino acid changes on the tertiary structure of BD2 (**Figure 6.3**). All the models were based upon the deposited PDB file (1E4Q) for the human hBD2 structure. It should be noticed that the HBD2 structure was resolved with solution NMR on the mature peptide form and hence information on the spatial conformation of the signal peptide is missing. For this reason, Phe→Leu change in position 12 for DEFB2L4 and Ser→Phe changes in all the variants are not present in **Figure 6.3**.



DEFB2L1

DEFB2L2

DEFB2L4

DEFB2L5

BAC Variant

Figure 6.3: Protein models generated for each BD2 variant identified. Residues present in the reference sequence are shown in red; residues specific for the variant in analysis are shown in green.

As shown in **Figure 6.3**, most of the amino acid changes fell on the loop connecting two anti-parallel β-strands. It is possible that these changes are permitted because of the loop flexibility and hence do not impair the protein folding, despite the small dimension of the mature peptide (37 residues). Nevertheless, the most dramatic change is observed for the BAC Variant, with an Arg→Trp change falling on the α-helix secondary structure of BD2. Interestingly, there are evidences of the role of hBD2 α-helix in anchoring to pathogens cell walls [162]. This non-synonymous substitution causes a shift from a polar, hydrophilic residue to a non-polar hydrophobic residue with high steric hindrance, likely to have an effect on the anchoring kinetics of the BD2 BAC Variant. Moreover, as this variant has been identified with concordant Sanger sequencing readings in two independent clones, it is unlikely to be a sequencing artefact or a spontaneous mutation affecting the BAC insert.

In order to evaluate overall changes in charge and hydropathicity, all the protein sequences in analysis were tested with ExPASy ProtParam tool, a computational approach to estimate chemical-physic properties of a given protein from its primary sequence. Just the residues constituting the mature peptide were tested (**Table 6.2**).

Protein variant	**Isoelectric point**	**Grand average of hydropathicity (GRAVY)**
DEFB2L	9.56	0.114
BAC Variant	8.9	0.573
DEFB2L1	9.56	0.122
DEFB2L2	9.36	0.276
DEFB2L4	9.56	0.114
DEFB2L5	9.4	0.124

Table 6.2: Isoelectric point and grand average of hydropathicity calculated on the base of the primary structure of the mature peptide for each protein variant.

As hypothesised from the position of the amino acid changes observed for the protein BAC Variant, its calculated isoelectric point and especially its hydropathicity showed a clear change when compared with the other variants. For this reason, a graphic representation of the electrostatic charge on the surface of the protein was generated using PyMol software (**Figure 6.4**).

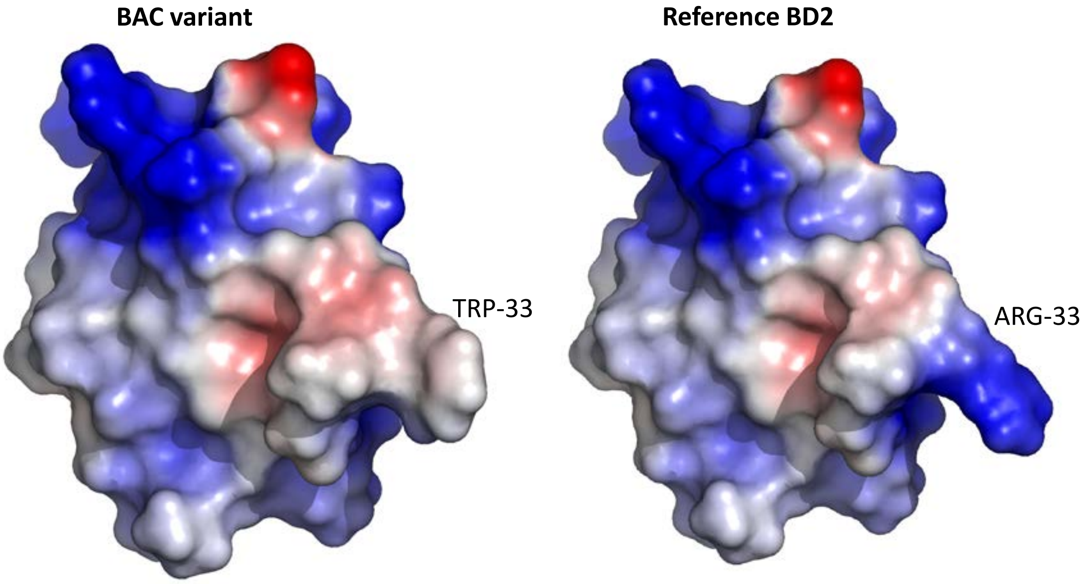


Figure 6.4: Surface graphic representation of the calculated electrostatic charge of the BAC Variant protein. Colours in blue scale indicate positively charged areas of the protein surface. Colours in red scale indicate negatively charged areas.

From **Figure 6.4** it was possible to observe a neat charge change in the anchoring area of BD2, together with a conformational change due to the different steric hindrances of arginine and tryptophan. Although a functional characterisation of the two variants was not performed, these data indicate the possibility of differences in their anchoring properties, with suspected changes in the affinity for different pathogen classes.

6.2 Tests of selection on *DEFB2L* coding sequence variants

Given the changes in the amino acid sequence reported in the previous section for the *DEFB2L* variants, with the BAC Variant presenting clear physico-chemical changes compared with the calculated values for the reference protein, the subsequent step was testing the hypothesis of positive selection on *DEFB2L*. For this aim, the Ka/Ks ratio test was chosen, being an indicator of selective pressure acting on a protein-coding gene, as in the case in the analysis.

Ka/Ks ratio is a test based on the ratio between synonymous and non-synonymous substitutions in pairwise alignments of protein coding genes. Ka/Ks ratio is an indicator of the type of selection acting on a protein coding gene: Ka/Ks=1 indicates an equal proportion of non-synonymous and synonymous substitution, suggesting neutral evolution; Ka/Ks<1 indicates an excess of synonymous substitutions over non-synonymous, indicative of purifying selection; Ka/Ks>1 indicates a higher proportion of synonymous substitutions over non-synonymous, with the presence of more amino acidic changes than expected by chance, suggestive of diversifying selection.

Pairwise alignments of each DEFB2L coding sequence variants were performed against the reference coding sequence, and Ka_Ks Calculator software was used to compute the ratio, applying a Model Averaging (MA) maximum likelihood method implemented by the creators of the software. Results are presented in **Table 6.3**.

Sequence	Ka	Ks	Ka/Ks	P-Value (Fisher)	S-Sites	N-Sites	Substitutions	Divergence Time
BAC	0.08	0.04	1.93121	0.47	47.21	144.79	12	0.07
DEFB2L1	0.01	0.00	49.9896	0.63	54.37	137.63	2	0.01
DEFB2L2	0.05	0.02	2.76584	0.17	60.40	131.60	7	0.04
DEFB2L4	0.01	0.00	48.3336	0.59	49.13	142.87	2	0.01
DEFB2L5	0.01	0.00	47.5475	0.61	48.90	143.10	2	0.01

Table 6.3: Ka/Ks ratio calculated for all the DEFB2L coding sequence variants in analysis (192 bp long). All the coding sequences were pairwise-aligned with *DEFB2L* reference coding sequence before running the test. Ka: Non-synonymous substitution rate. Ks: Synonymous substitution rate. Ka/Ks: Selective strength. P-Value(Fisher): The value computed by Fisher exact test. S-Sites: Synonymous sites. N-Sites: Non-synonymous sites. Substitutions: Substitutions between sequences. Divergence-Time: calculated divergence time between sequences; it is a weighted average of synonymous substitution rate and non-synonymous substitution rate.

Ka/Ks results are indicative of diversifying selections for all the variants in analysis but not significant, as evidenced by the Fisher's exact Test P value. The total absence of synonymous sites for DEFB2L1, DEFB2L4 and DEFB2L5 placed problems in the calculation of the Ka/Ks ratio, as shown from their unrealistic values in the fourth column. The presence of more substitutions for the BAC Variant and DEFB2L2 determined a more reasonable computing of Ka/Ks, but without reaching a significant threshold. A possible reason to explain these results could lie in the small number of substitutions between the two variants which could be due to recent divergence or to the short sequence in analysis (192bp). In this case, it is possible that a combination of the two factors did not confer sufficient power to the analysis.

6.3. Phylogenetic analysis of BD2 protein orthologs

Another important aspect to be considered was the evolutionary relationship among rhesus macaque BD2 protein variants and their orthologues in other mammalian species. With this aim a BLASTProtein search was made, querying the non-redundant protein collection database, to retrieve all the putative orthologue proteins to BD2 (coded by the *DEFB2L* gene). An E value of 5e^{-04} was selected as lower threshold, as hits with lower E value presented a drop in the primary structure homology (<48%) that would have challenged the assumption of real orthology. A selection of representative species was made on the filtered hits, with particular attention to all the deposited primate sequences. With this approach, the following species were selected: mouse, rat, pig, horse, marmoset, gorilla, chimpanzee, gibbon, orangutan, *Homo sapiens* and, for rhesus macaque, all the known BD2 variants were included (the annotated reference, the BAC Variant and the 4 variants identified through subcloning). An alignment file for all the BD2 protein orthologues retrieved was generated using Clustal Omega, and the output file was then processed through MEGA6, generating a neighbour-joining tree with distance correction (**Figure 6.5**), to obtain a first overall estimate of the phylogenetic relations. One of the human paralogue to BD2, DEFB103, was used as outlier group to root the tree, considered the low sequence conservation among β-defensin paralogues.

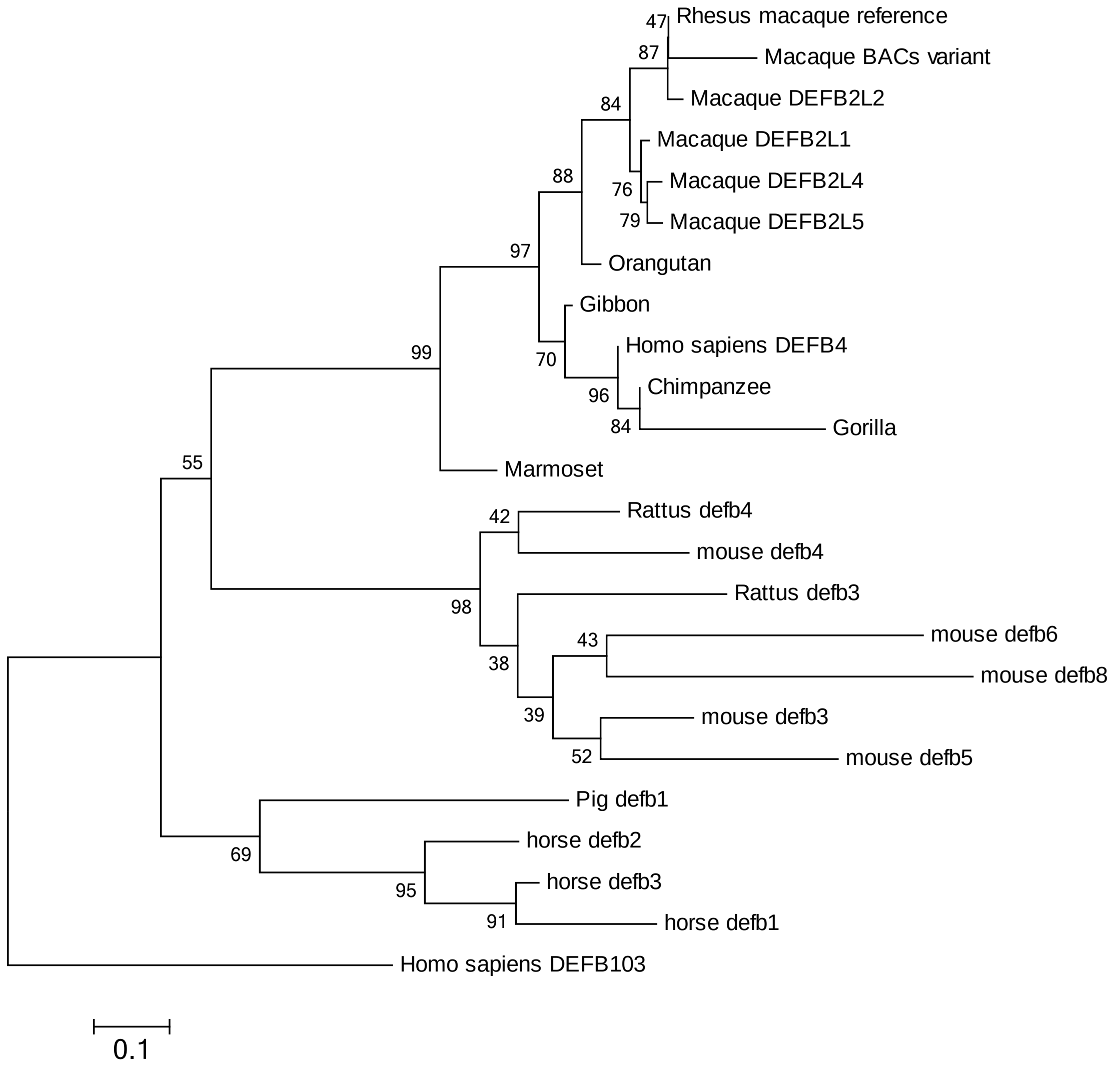


Figure 6.5: Neighbour-joining phylogenetic tree generated on 17 BD2 protein orthologs, 5 rhesus macaque BD2 known variants and the paralogous human protein DEFB103, used to root the tree. The optimal tree with the sum of branch length = 5.85 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Dayhoff matrix based method and are in the units of the number of amino acid substitutions per site.

Despite the small size of the proteins aligned (64 amino acids), it was possible to observe how the neighbour-joining tree generated reflected the separation among the species considered, with an accurate branching kept also within the primates order (refer to Appendix 1: Primates phylogeny). The only error is posed by the separation gibbon-orangutan, as the former, being a small ape, was expected to have a higher sequence divergence, whereas the latter should have clustered with the other great apes. Murine DEFB proteins presented variable genetic distances, data consistent with the complex pattern of rapid divergence that characterise the evolution of murine defensins [140], posing problems in the identification of real BD2 orthologues. In order to identify real BD2 orthologs outside the primate order, a further filtering was performed on the putative orthologs: on the base of the genetic distance values calculated (**Figure 6.5**), just the protein sequence with the lowest genetic distance was kept for each species: horse DEFB3, mouse DEFB3, rat DEFB4 and all the primate protein sequences. A maximum likelihood tree based on this reduced dataset was generated using MEGA6, performing 500 bootstrap replications (**Figure 6.6**).

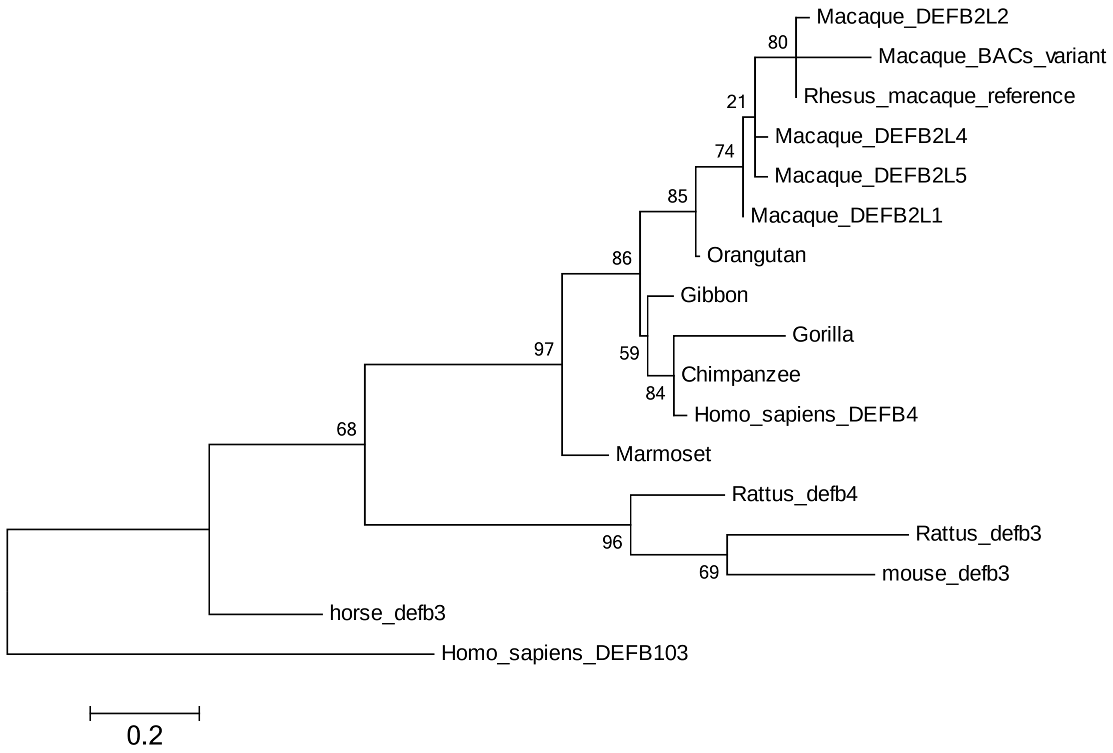


Figure 6.6: Maximum-likelihood tree generated on the putative BD2 paralogues selected accordingly to the lowest genetic distance. The tree with the highest log likelihood (-849.65) is shown, computed with 500 bootstrap repetitions. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The human paralogue sequence DEFB103 is used for rooting the tree.

. A further test was made generating a tree with identical parameters without DEFB103, and no differences were observed in the branching order for each node (**Figure 6.7**). Given these results, this protein dataset has been selected to perform further selection tests.

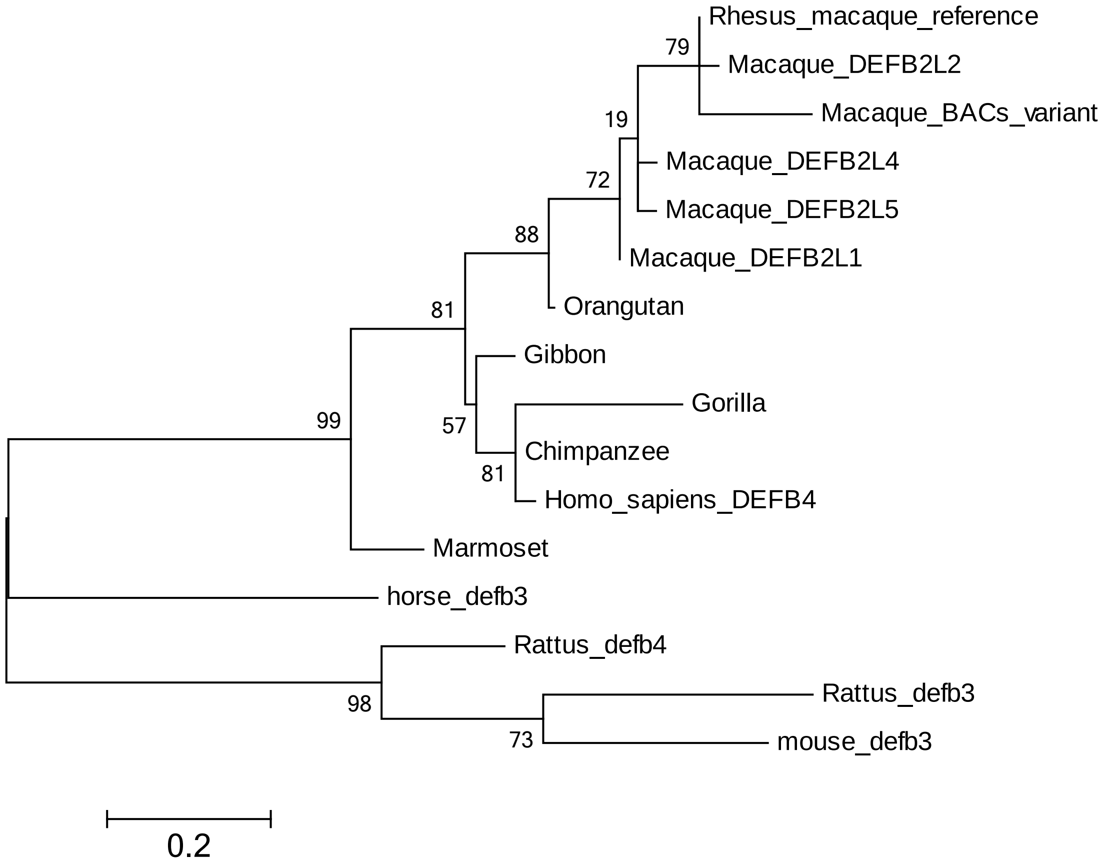


Figure 6.7: Maximum-likelihood tree generated on the putative BD2 paralogues selected accordingly to the lowest genetic distance. The unrooted tree with the highest log likelihood (-734.38) is shown, computed with 500 bootstrap repetitions. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site.

6.3.1 Ka/Ks ratio on BD2 orthologues

In order to find signs of selection among BD2 orthologues, an inter-species Ka/Ks analysis was performed. Two approaches were tested: the coding sequences of all the putative BD2 orthologues were pairwise aligned either with the human *DEFB4* reference (**Table 6.4**) or with rhesus *DEFB2L* reference (**Table 6.5**).

Sequence	**Ka**	**Ks**	**Ka/Ks**	**P-Value (Fisher)**	**S-Sites**	**N-Sites**	**Substitu-tions**	**Divergence Time**
rhesus	0.11	0.06	1.85	0.39	47.25	144.75	18	0.10
chimp	0.01	0.03	0.24	0.23	45.35	146.65	2	0.01
gibbon	0.03	0.09	0.34	0.12	42.88	149.12	8	0.05
gorilla	0.89	1.35	0.66	0.00	43.65	133.35	124	1.00
horse	0.40	0.73	0.55	0.00	50.92	141.08	64	0.48
orangutan	0.07	0.12	0.57	0.29	40.05	151.95	14	0.08
marmoset	0.12	0.45	0.26	0.00	46.38	145.62	29	0.20

Table 6.4: Ka/Ks ratio calculated for 7 DEFB2L coding sequence orthologues. All the coding sequences were pairwise-aligned with human *DEFB4* coding sequence before running the test Ka: Non-synonymous substitution rate. Ks: Synonymous substitution rate. Ka/Ks: Selective strength. P-Value(Fisher): The value computed by Fisher exact test. S-Sites: Synonymous sites. N-Sites: Non-synonymous sites. Substitutions: Substitutions between sequences. Divergence-Time: calculated divergence time between sequences; it is a weighted average of synonymous substitution rate and non-synonymous substitution rate.

Results of the Ka/Ks ratio using human *DEFB4* for all the pairwise alignments showed weak indication of diversifying selection just for the rhesus macaque *DEFB2L* coding sequence. All the other BD2 orthologues considered presented a Ka/Ks ratio <1, suggestive of purifying selection, with low frequency of non-synonymous substitutions. It should be noted that the gorilla coding sequence reference (ENSGGOG00000027908) is 165 base pairs long, compared with a length of 195bp in rhesus macaque, hence the high number of substitutions and the maximum divergence time observed for gorilla are just highlighting the presence of a splicing variant and are not truly indicative of high divergence. Nevertheless, just the values calculated for horse BD2 orthologues passed the Fisher's exact test, because of the higher number of substitutions present. The Ka/Ks values for mouse defb3 and rat defb3 failed to be computed, likely because of excessive sequence divergence against human *DEFB4*.

Sequence	Ka	Ks	Ka/Ks	P-Value (Fisher)	S-Sites	N-Sites	Substitu-tions	Divergence Time
human	0.11	0.06	1.85	0.39	47.25	144.75	18	0.10
chimp	0.11	0.08	1.33	0.77	46.69	145.31	18	0.10
gibbon	0.10	0.02	4.00	0.11	51.59	140.41	14	0.08
gorilla	0.87	1.43	0.61	0.00	41.92	135.08	121	1.00
horse	0.47	0.67	0.70	0.08	47.77	144.23	68	0.52
orangutan	0.05	0.04	1.28	0.65	47.32	144.68	9	0.05
marmoset	0.18	0.25	0.72	0.27	49.61	142.39	32	0.20

Table 6.5: Ka/Ks ratio calculated for 7 DEFB2L coding sequence orthologues. All the coding sequences were pairwise-aligned with rhesus macaque reference *DEFB2L* coding sequence before running the test.

Results of the Ka/Ks ratio applying rhesus *DEFB2L* reference for all the pairwise alignments showed slightly different results compared with **Table 6.4**. This finding is possibly reflective of the higher rate of non-synonymous substitution of rhesus *DEFB2L*, evidenced through the sequence alignment with the other orthologues. Also in this case, just horse BD2 orthologue presented a significant P-value, as gorilla values reflected a deletion in the coding sequence. Nonetheless, the Ka/Ks value of 4 reported for the gibbon may be a real signal of diversifying selection; the rise of lineage-specific non-synonymous substitutions in gibbon DB2 may also explain the wrong rooting in all the maximum likelihood trees generated, despite further studies on this Primate should be necessary to confirm this hypothesis.

6.4 Chapter summary

This chapter presents the characterisation of a novel rhesus macaque *DEFB2L* variant, indentified through Sanger sequencing on two independent BAC clones with inserts spanning the rhesus macaque β-defensin cluster. As other *DEFB2L* variants have been validated from an independent research, the hypothesis of selective pressure acting on the diversification of *DEFB2L* in rhesus macaque was tested.

The 'BAC Variant' identified presented a total of 12 base pair changes compared with the reference, of which 9 non-synonymous substitutions. The creation of protein models of the 3D spatial conformation of the protein variants highlighted the presence of amino acid changes in specific positions, mainly clustered to flexible loops connecting two antiparallel β-sheets of BD2. Interestingly, the BAC Variant presented a significant amino acid change (Arg33→Trp33) in correspondence of an α-helix reported to be the anchoring point of BD2 to pathogens cell walls, to exert its antibacterial activity. *In Silico* calculations of the chemical-physic parameters of the BAC Variant protein highlighted hydropathicity changes; as BD2 is a secreted peptide, these data were suggestive of a potential effect on its activity kinetics and specificity. Ka/Ks ratio tests aimed at the identification of possible selective forces driving the evolution of these variants was not conclusive, possibly for a combination of short coding sequence length (192 base pairs), with small number of substitutions, and recent divergence. The dataset used may have been inadequate for the statistical power needed for the test.

A phylogenetic approach was tried to visualise the relations among BD2 orthologues in other species and in the rhesus macaque BD2 variants identified. Maximum likelihood trees based on protein sequence data overall reflected the speciation order of the species considered. An exception was constituted by the separation gibbon-orangutan that did not follow the expected branching pattern. Ka/Ks ratio was used as indicator of selective pressure across the BD2 orthologues in the analysis, but failed to give consistent results on the forces driving the diversification of BD2.

DISCUSSION

Over the last decade, as the real extent of copy number variable regions in the human genome was progressively unravelled, an increasing attention from the scientific community was placed in the understanding of the importance of structural variations in shaping human adaptation and pathology [19 82] (among others). It hence became crucial to get insights on the copy number variation pattern in non-human primates, to address questions about CNV formation and maintenance under an evolutionary perspective. It became clear that sequence gain and loss events could alter the gene complement and expression of an organism, bringing to phenotypic variations subjected to selective pressures, as altered environmental conditions or infectious diseases. In agreement with this observation, different immunity-related genes were found to be CNV in humans [2 19 29 82] (among others) and in other non-human primates [48 50 79 80], constituting significant examples of how CNVs can contribute to rapid adaptation.

Among the CNV immune-related genes, β-defensins, with a diploid copy number in humans ranging between 1 and 12 copies, constitute a paradigm case of a gene family that evolved to gain diverse functions in different species [131 146]. Moreover, β-defensin proteins possess a key role in linking innate and adaptive immunity [175], and associations were found between high β-defensin copy numbers and susceptibility to the autoimmune disease psoriasis [121] and to HIV infection, with poor response to antiretroviral therapy [211]. Evidence of β-defensin copy number variations were also found in chimpanzee [35] and they showed signals of variability also in other non-human primates, in different cross-species genome-wide array-CGH studies [50 79 158]. Nevertheless, the low probe coverage in genomic regions containing segmental duplications, the poor quality of the genomic assemblies for non-human primates in paralogous regions and the increasing level of sequence divergence for primates with a longer divergence time from human, posed notable problems for an accurate comparative analysis of the genomic organisation of the β-defensin cluster in different primates. In order to try and overcome these limitations, this project focused on the characterisation and copy number typing of the rhesus macaque β-defensin region. Among the other species, rhesus macaque was chosen as model organism for two main reasons: the first is that it is the most widespread non-human primate, with wide distribution and large population size, showing a great adaptation to different environments and diets; moreover, it often shares the same environments with human, hence facing with them the same environmental challenges. The second reason is that, given 25 millions of years of divergence time between human and rhesus macaque, CNVs shared between the two lineages are likely to have arisen as consequence of independent events [79] and can provide information on the different evolutionary forces driving CNV formation and maintenance.

At the start of this project little was known about the organisation of the β-defensin cluster in *Macaca mulatta* species. In the first published rhesus macaque assembly (January 2006 MGCS Merged 1.0/RheMac2) just the gene *DEFB2L* was annotated, despite sequence alignment of the seven human β-defensins present on the chr8p23.1 cluster revealed their likely presence in the same position and with high degree of sequence homology in the rhesus macaque β-defensin cluster.

Before the start of this project, Lee AS *et al*. detected a signal from array CGH data indicative of CNV at the rhesus macaque β-defensin locus [79]. Nevertheless, because of the low probe coverage due to the presence of large segmental duplications, they could not identify the boundaries of the variable area. They tried to confirm the array CGH signal for *DEFB2L* using quantitative PCR but, according to the CNV boundaries here described, they designed primers out of the CNV, bringing to a discordant validation of the CNV. Notwithstanding, from the array CGH data they observed a noticeable difference in the copy number repeat size between human (>5 times bigger) and rhesus macaque. Authors attributed the discrepancy to problems in the assembly, with sequence gaps in both human and macaque reference sequences and to the presence of large segmental duplication. These aspects truly impeded a more accurate sizing of this CNV but in reality the signal observed was more authentic than was previously thought: previous analysis [29] identified a copy number repeat unit of approximately 240 kb for the human β-defensin cluster, and, from the characterisation of large insert clones, this project provided evidences of a copy number repeat unit of 20 kb in rhesus macaque, confirming the large size discrepancy.

The first part of this study focused on the identification of CNV signals in the rhesus macaque β-defensin cluster and on the characterisation of the copy number breakpoints, with the highest probe coverage allowed by the published genomic assembly quality for the area in analysis. It was possible to detect clear signs of variation localised on the β-defensin area, despite the breakpoint analysis identified just the telomeric boundary of the CNV repeat unit, due to presence of a sequence gap in the genomic area proximal to *DEFB2L.* Nevertheless, these findings highlighted that just the *DEFB2L* gene falls in the CNV repeat unit, whereas the other 6 defensin genes distal to *DEFB2L* belong to a non-CNV area. This represents a clear structural difference compared with the organisation of the β-defensin cluster in human, with 7 defensins being part of the same CNV repeat unit of approximately 240 kb. The same breakpoint analysis performed on the other genes screened on the same array-CGH chip, not included in this dissertation (Table 2.3), did not support the presence of structural variations. Given that these genes were selected on the basis of previous CNV indications reported from other studies [50 79], these negative results were unexpected. It is hence possible that the hidden Markov model filtering applied for breakpoint identification was too strict. It should be considered that the model was trained on larger deletions and that Korbel JO *et al.* detected CNVs with median size of 15kb and average size of 85kb applying the BreakPtr algorithm to their high-density array CGH data [202]. Hence, it is possible that small sized copy numbers or small variations in copy number compared to the reference may not be detected with this approach. Further analyses and the design and application of dedicated algorithms on the raw data generated from the other CNV candidate regions shall be performed, in order to get more accurate information on the copy number state of these genomic regions of biological relevance. Nevertheless, the signal intensity values obtained from the array CGH demonstrated a good level of accuracy. The dosage positive controls on chromosome X clustered around the expected log intensity values and it was possible to separate males from females just on the basis of a 'one copy' difference. Moreover, the average log intensity values for the probes falling into the copy number repeat unit of 20 kb (chr8:8068451-8076651) showed a good concordance with PRT8-14 (r^2=0.78) copy number calling and with digital droplet PCR assays (r^2=0.86, if sample Rh6 is excluded).

Another interesting aspect highlighted from the array CGH data was the detection of a signal increase associated with the α-defensin cluster, for three of the samples in analysis. A previous study confirmed the presence of extensive copy number variation associated with *DEFA1/DEFA3* in human and high copy number of *DEFA1* in great apes [212]; moreover, signals of copy number variation were detected in correspondence of the *DEFA5* gene, from array CGH data on rhesus macaque [79]. Further investigations should be necessary in this direction, to confirm the presence of α-defensin copy number variation in Old World monkeys.

Nevertheless, given the negative effects of the lack of probe coverage in the area proximal to *DEFB2L* on the identification of the copy number boundaries, a different experimental strategy should have been applied, in the light of the positive results obtained in a later phase of the project through the β-defensin region analysis using large insert clones. The generation of 217.5kb of contiguous sequence information spanning the β-defensin cluster should have been used to cover the gaps present on the rhesus macaque genomic assembly, prior to the probe design for the array CGH. This would have allowed a better probe coverage, with a consequent higher resolution for the area in analysis.

Using PCR-based methods, it was possible to confirm that the *DEFB2L* locus is copy number variable in rhesus macaque, as observed on a cohort of 88 samples, with a diploid copy number ranging between 2 and 11. These results are extremely interesting, given a comparable β-defensin copy number distribution in humans, between 1 and 12. Considering the great difference in the CNV repeat unit between humans and rhesus macaque (240kb vs 20 kb), it is possible to hypothesise that the absence of rhesus macaque individuals with copy numbers >12 is not due to physical constraints in the genomic architecture of chr8p23.1.

A more probable explanation could be that higher levels of BD2 expression, associated with higher copy numbers, would have a deleterious effect in both lineages, possibly as a consequence of excessive immune system activation. If these two CNV clusters arose as a consequence of independent events [79], this similar CNV distribution observed could be due to convergent selection acting on the number of copies, based on functional constraints. Considering that rhesus monkeys often co-inhabit human environments, also the hypothesis of a similar pathogenic pressure shaping β-defensin copy number distribution of the two lineages should be taken into account and further investigated.

The paralogue ratio test was confirmed to be an accurate method for CNV typing, showing a good level of concordance among replicates, also in case of low-quality DNA samples. The only drawback of this technique was the difficulty associated in the design of one single primer pair for the test and the reference region, on repeated elements. The possible presence of non assembled repeats in the rhesus macaque genome, together with the lack of SNPs and length polymorphisms information for this assembly, made the design of a suitable PRT candidate extremely challenging. The use of a bioinformatic approach for PRT primers selection is hence strongly suggested to reduce the time needed for the optimisation of a suitable assay.

Digital droplet PCR was confirmed to be a new alternative method for copy number typing with efficiency comparable to PRT, despite higher standard deviation values observed among replicates. Nevertheless, the use of two different primer pairs, for the test and for the diploid reference region, makes the areas of interest easily targeted. For this reason, ddPCR could be a valid alternative to type copy numbers in genomic regions where the design of suitable PRT assays is not possible.

The rhesus macaque CHORI-250 BAC library screening represented a very informative approach for the characterisation of the β-defensin region. Through single-molecule real time sequencing (Pacific Bioscience PacBio Platform) it was possible to generate two *de novo* contigs mapping to the same chromosome copy, with overlapping ends spanning the β-defensin region, for a total of 217.5 kb. It should be noticed that this sequencing technique has the advantage of generating long reads (from 3kb to 10 kb) but presented a higher percentage of error compared with Sanger sequencing, for miss-called bases and gaps insertion. This was the case for one of the contigs generated for this project that presented a triplet insertion and a one non-synonymous substitution in the second exon of *DEFB2L*. These variations were not confirmed through Sanger sequencing and are likely to represent sequencing artefacts, but without an alternative confirmation method and other BAC clones with supporting information, a novel variant might have been erroneously called. This aspect should be taken into account when using this sequencing approach for *de novo* assemblies that require high-sequence fidelity.

Two of the BAC clones mapping to the β-defensin region, 201P10 and 47B11, were successfully used as fluorescent probes to confirm the localisation of the β-defensin cluster in rhesus macaque through a cytogenetic approach. Results of metaphase FISH confirmed the localisation of the cluster on the short arm of chromosome 8. A higher-resolution cytogenetic approach was also attempted, with the aim of determining unequivocally the β-defensin diploid copy number on DNA fibres generated from rhesus macaque lymphoblastoid cell lines. These lines would have become copy number standards to normalise all the PCR-based assays described herein. Any attempt done using BAC clone probes for fibres localisation together with four probes spanning the *DEFB2L* locus for copy number counting, failed to give reliable results. Nevertheless, fibre-FISH still represents an elegant approach to confirm and normalise copy number typing assays based on PCR amplification, as their reaction kinetics may introduce systematic bias in the copy number call. Optimistically, this approach should be applied in each copy number typing project that cannot rely on positive copy number controls previously identified.

From Sanger sequencing of *DEFB2L* coding sequence it was possible to identify two variants of the same gene from independent clones: one correspondent to the reference coding sequence (NM_001128851.1) and one not previously described that was called 'BAC Variant'. As there are no coding sequence variants reported for human *DEFB4*, these findings catalysed the last part of this project towards the characterisation of this variant, together with other five sequence variants identified by [208] and now annotated on the rhesus macaque assembly (EU090139-EU090143). Four of these variants presented non-synonymous substitutions compared with the reference coding sequence, raising the possibility of an alternative evolutionary pattern undertaken by BD2 proteins after the separation of Old World Monkeys from the human lineage. In particular, the BAC Variant presented 9 non-synonymous substitutions that were calculated to determine important changes in the hydropathic properties of the protein. This shift in the chemical-physical properties of the BAC Variant protein was driven by an Arg→Trp substitution falling on the α-helix secondary structure of BD2. Interestingly, there are evidences of the role of human hBD2 α-helix in anchoring to pathogens cell walls [162]. This non-synonymous substitution causes a shift from a polar, hydrophilic residue to a non-polar hydrophobic residue with high steric hindrance, likely to have an effect on the anchoring kinetics of the BD2 BAC Variant. This aspect was speculated to determine a fine tuning in the affinity of the different BD2 variants for different classes of pathogens. This would create a repertoire of antimicrobial peptides with fine differences, shuffled in the rhesus macaque population through copy number change. Moreover, the key position of the β-defensin cluster in a copy number hotspot [47] would maintain the tendency to recurrent duplications and deletions through non-allelic homologous recombination; this would contribute to the adaptation process to different pathogenic pressures, with the generation and expression of different combination of antimicrobial peptide variants. In order to prove this hypothesis, a functional study would have been required, to test the different protein variants for changes in their immune signalling activity and in their antibacterial properties. Nevertheless, the formation of the right order of disulphide-bridges between the six-cysteine motif characteristic of defensins is a key step for the correct folding of the BD2 proteins. Unfortunately normal bacterial systems of protein expression do not guarantee the production of just the desired BD2 variant with the correct disulphide-bond pairing. A laborious method described to purify human α-defensin is based on monitoring and changing the reduced/oxidised state of the defensin peptide in a series of steps, with the aim of guiding the protein folding [213], which was beyond the scopes of this project.

The Ka/Ks selection test performed on all the *DEFB2L* coding sequence variants, in pairwise alignments with the reference coding sequence, was suggestive in all cases of diversifying selection, despite none of the variants presented a significant Fisher's exact test p value. This result can be explained by the fact that the amount of statistical power derived from paralogous sequences of small size is dramatically smaller than that obtained from the analysis of polymorphic sites among different individuals. Hence it was possible to obtain just indications of the selection forces acting on these variants.

Phylogenetic analysis of putative BD2 orthologs in different species, with the generation of maximum likelihood trees based on protein sequence data, overall reflected the speciation order of the species considered. The only discrepancy observed regarded a non expected rooting of gibbon and orangutan, with the former clustering with great apes and the latter presenting a higher divergence than expected.

A Ka/Ks selection test performed on all the putative *DEFB2L* coding sequences orthologues in 10 different species, in pairwise alignment with either human or rhesus macaque reference coding sequences, suggested signs of positive selections just in macaque, with all the other BD2 orthologues presenting indication of purifying selection.

Interestingly, the results of the Ka/Ks test for gibbon, in pairwise alignment with rhesus macaque, revealed possible signs of diversifying selection. The rise of lineage-specific non-synonymous substitutions in gibbon BD2 may also explain the wrong rooting reported from the phylogenetic trees. In order to test for this hypothesis, further studies may be conducted on this primate, to look for evidences of diversification and copy number variation at the β-defensin cluster. Again, the small size of the coding sequences did not provide enough statistical power to pass the Fisher's exact test. The results of this analysis should be hence considered just suggestive of positive selection on rhesus macaque BD2 coding sequence.

This study represents a further step towards a deeper understanding of the mechanisms driving the evolution of the β-defensin region; here is presented a case where the same area prone to copy number variation evolved to present a different copy number unit structure in two different lineages, still converging in the same copy number distribution, possibly for the effect of similar constraints on functional level. This is the first reported case of a β-defensin whose sequence is not fixed in the rhesus macaque population, opening a time-window on the intermediate stage between gene duplication and divergence, with fixation of the new paralogue in the population, as happened with the β-defensin paralogues of chr8p23.1.

BIBLIOGRAPHY

1. Bridges CB. The bar" gene" a duplication. Science 1936;**83**(2148):210-11

2. O'Neill GJ, Yang SY, Dupont BO. Two HLA-linked loci controlling the fourth component of human complement. Proceedings of the National Academy of Sciences 1978;**75**(10):5165-69

3. Isenman D, Young J. The molecular basis for the difference in immune hemolysis activity of the Chido and Rodgers isotypes of human complement component C4. The Journal of Immunology 1984;**132**(6):3019-27

4. Giles C. Antigenic determinants of human C4, Rodgers and Chido. Experimental and clinical immunogenetics 1987;**5**(2-3):99-114

5. Howard BH, Sakamoto K. Alu interspersed repeats: selfish DNA or a functional gene family? The New Biologist 1990;**2**(9):759

6. Ostertag EM, Kazazian Jr HH. Biology of mammalian L1 retrotransposons. Annual review of genetics 2001;**35**(1):501-38

7. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. Nature genetics 2004;**36**(9):949-51

8. Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. Science 2004;**305**(5683):525-28

9. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature Reviews Genetics 2006;**7**(2):85-97

10. Bailey JA, Eichler EE. Primate segmental duplications: crucibles of evolution, diversity and disease. Nature Reviews Genetics 2006;**7**(7):552-64

11. Hurles ME, Dermitzakis ET, Tyler-Smith C. The functional impact of structural variation in humans. Trends in Genetics 2008;**24**(5):238-45

12. Perry GH, Dominy NJ, Claw KG, et al. Diet and the evolution of human amylase gene copy number variation. Nature genetics 2007;**39**(10):1256-60

13. DeVries S, Gray JW, Pinkel D, et al. Comparative genomic hybridization. Current Protocols in Human Genetics 2001:4.6. 1-4.6. 18

14. Sharp AJ, Locke DP, McGrath SD, et al. Segmental duplications and copy-number variation in the human genome. American journal of human genetics 2005;**77**(1):78

15. Neill NJ, Torchia BS, Bejjani BA, et al. Comparative analysis of copy number detection by whole-genome BAC and oligonucleotide array CGH. Molecular cytogenetics 2010;**3**(11)

16. Park H, Kim JI, Ju YS, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nature genetics 2010;**42**(5):400-05

17. Pinto D, Darvishi K, Shi X, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nature biotechnology 2011;**29**(6):512-20

18. Tuzun E, Sharp AJ, Bailey JA, et al. Fine-scale structural variation of the human genome. Nature genetics 2005;**37**(7):727-32

19. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. nature 2006;**444**(7118):444-54

20. Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science 2007;**318**(5849):420-26

21. Carvalho CMB, Lupski JR. Copy number variation at the breakpoint region of isochromosome 17q. Genome research 2008;**18**(11):1724-32

22. Weaver S, Dube S, Mir A, et al. Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. Methods 2010;**50**(4):271-76

23. Ståhlberg A, Åman P, Ridell B, et al. Quantitative real-time PCR method for detection of B-lymphocyte monoclonality by comparison of κ and λ immunoglobulin light chain expression. Clinical Chemistry 2003;**49**(1):51-59

24. Guescini M, Sisti D, Rocchi MB, et al. A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition. BMC bioinformatics 2008;**9**(1):326

25. Kontanis EJ, Reed FA. Evaluation of Real-Time PCR Amplification Efficiencies to Detect PCR Inhibitors. Journal of forensic sciences 2006;**51**(4):795-804

26. Marubini E, Verderio P, Raggi CC, et al. Statistical diagnostics emerging from external quality control of real-time PCR. International Journal of Biological Markers 2004;**19**(2):141-46

27. Chen QX, Book M, Fang XM, et al. Screening of copy number polymorphisms in human β-defensin genes using modified real-time quantitative PCR. Journal of immunological methods 2006;**308**(1):231-40

28. Groth M, Szafranski K, Taudien S, et al. High-resolution mapping of the 8p23. 1 beta-defensin cluster reveals strictly concordant copy number variation of all genes. Human mutation 2008;**29**(10):1247-54

29. Hollox EJ, Armour JAL, Barber JCK. Extensive normal copy number variation of a β-defensin antimicrobial-gene cluster. The American Journal of Human Genetics 2003;**73**(3):591-600

30. Armour JAL, Sismani C, Patsalis PC, et al. Measurement of locus copy number by hybridisation with amplifiable probes. Nucleic Acids Research 2000;**28**(2):605-09

31. Hollox EJ, Davies J, Griesenbach U, et al. Beta-defensin genomic copy number is not a modifier locus for cystic fibrosis. Journal of negative results in biomedicine 2005;**4**(1):9

32. Sellner LN, Taylor GR. MLPA and MAPH: new techniques for detection of gene deletions. Human mutation 2004;**23**(5):413-19

33. Schouten JP, McElgunn CJ, Waaijer R, et al. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic acids research 2002;**30**(12):e57-e57

34. Deutsch S, Choudhury U, Merla G, et al. Detection of aneuploidies by paralogous sequence quantification. Journal of medical genetics 2004;**41**(12):908-15

35. Armour JAL, Palla R, Zeeuwen PLJM, et al. Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. Nucleic acids research 2007;**35**(3):e19-e19

36. Sykes P, Neoh S, Brisco M, et al. Quantitation of targets for PCR by use of limiting dilution. Biotechniques 1992;**13**(3):444-49

37. Hindson BJ, Ness KD, Masquelier DA, et al. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Analytical chemistry 2011;**83**(22):8604-10

38. Pinheiro LB, Coleman VA, Hindson CM, et al. Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. Analytical chemistry 2011;**84**(2):1003-11

39. Henrich TJ, Gallien S, Li JZ, et al. Low-level detection and quantitation of cellular HIV-1 DNA and 2-LTR circles using droplet digital PCR. Journal of virological methods 2012

40. Alkan C, Kidd JM, Marques-Bonet T, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nature genetics 2009;**41**(10):1061-67

41. Cantsilieris S, Baird PN, White SJ. Molecular methods for genotyping complex copy number polymorphisms. Genomics 2012

42. Schipper J, Chanson JS, Chiozza F, et al. The status of the world's land and marine mammals: diversity, threat, and knowledge. Science 2008;**322**(5899):225-30
43. Goodman M. The genomic record of Humankind's evolutionary roots. American journal of human genetics 1999;**64**(1):31
44. Guy J, Spalluto C, McMurray A, et al. Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q. Human molecular genetics 2000;**9**(13):2029-42
45. Horvath JE, Viggiano L, Loftus BJ, et al. Molecular structure and evolution of an alpha satellite/non-alpha satellite junction at 16p11. Human molecular genetics 2000;**9**(1):113-23
46. Garrigan D, Hammer MF. Reconstructing human origins in the genomic era. Nature Reviews Genetics 2006;**7**(9):669-80
47. Perry GH, Tchinda J, McGrath SD, et al. Hotspots for copy number variation in chimpanzees and humans. Proceedings of the National Academy of Sciences 2006;**103**(21):8006-11
48. Perry GH, Yang F, Marques-Bonet T, et al. Copy number variation and evolution in humans and chimpanzees. Genome research 2008;**18**(11):1698-710
49. Dumas L, Kim YH, Karimpour-Fard A, et al. Gene copy number variation spanning 60 million years of human and primate evolution. Genome Research 2007;**17**(9):1266-77
50. Gökçümen Ö, Lee C. Copy number variants (CNVs) in primate species using array-based comparative genomic hybridization. Methods 2009;**49**(1):18-25
51. Yi S, Ellsworth DL, Li W-H. Slow molecular clocks in Old World monkeys, apes, and humans. Molecular Biology and Evolution 2002;**19**(12):2191-98
52. Elango N, Thomas JW, Soojin VY. Variable molecular clocks in hominoids. Proceedings of the National Academy of Sciences of the United States of America 2006;**103**(5):1370-75
53. Hahn MW, Demuth JP, Han SG. Accelerated rate of gene gain and loss in primates. Genetics 2007;**177**(3):1941-49
54. Bailey JA, Yavor AM, Massa HF, et al. Segmental duplications: organization and impact within the current human genome project assembly. Genome Research 2001;**11**(6):1005-17
55. Lynch M, O'Hely M, Walsh B, et al. The probability of preservation of a newly arisen gene duplicate. Genetics 2001;**159**(4):1789-804
56. Sudmant PH, Huddleston J, Catacchio CR, et al. Evolution and diversity of copy number variation in the great ape lineage. Genome research 2013
57. Lynch M. The origins of genome architecture. 2007
58. Kumar RA, KaraMohamed S, Sudi J, et al. Recurrent 16p11. 2 microdeletions in autism. Human molecular genetics 2008;**17**(4):628-38
59. Locke DP, Segraves R, Carbone L, et al. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome research 2003;**13**(3):347-57
60. Goidts V, Armengol L, Schempp W, et al. Identification of large-scale human-specific copy number differences by inter-species array comparative genomic hybridization. Human genetics 2006;**119**(1):185-98
61. Wilson GM, Flibotte S, Missirlis PI, et al. Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. Genome research 2006;**16**(2):173-81
62. Go Y, Niimura Y. Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. Molecular biology and evolution 2008;**25**(9):1897-907
63. Jacobs GH. Evolution of colour vision in mammals. Philosophical Transactions of the Royal Society B: Biological Sciences 2009;**364**(1531):2957-67

64. Rouquier S, Taviaux S, Trask BJ, et al. Distribution of olfactory receptor genes in the human genome. Nature genetics 1998;**18**(3):243-50

65. Gilad Y, Wiebe V, Przeworski M, et al. Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. PLoS biology 2004;**2**(1):e5

66. Zhang J, Zhang Y-p, Rosenberg HF. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nature genetics 2002;**30**(4):411-15

67. Beintema JJ. The primary structure of langur (Presbytis entellus) pancreatic ribonuclease: adaptive features in digestive enzymes in mammals. Molecular biology and evolution 1990;**7**(5):470-77

68. Barnard EA. Biological function of pancreatic ribonuclease. 1969

69. Libonati M, Sorrentino S. Degradation of double-stranded RNA by mammalian pancreatic-type ribonucleases. Methods in Enzymology 2001;**341**:234-48

70. Guyton AC, Hall J. Textbook of Medical Physiology. 1996. Philadelphia: Saunders Company

71. Zhang J. Parallel functional changes in the digestive RNases of ruminants and colobines by divergent amino acid substitutions. Molecular biology and evolution 2003;**20**(8):1310-17

72. Lebenthal E. Role of salivary amylase in gastric and intestinal digestion of starch. Digestive Diseases and Sciences 1987;**32**(10):1155-57

73. Groot PC, Bleeker MJ, Pronk JC, et al. The human α-amylase multigene family consists of haplotypes with variable numbers of genes. Genomics 1989;**5**(1):29-42

74. Lambert JE. Competition, predation, and the evolutionary significance of the cercopithecine cheek pouch: the case of Cercopithecus and Lophocebus. American journal of physical anthropology 2005;**126**(2):183-92

75. Vandepoele K, Van Roy N, Staes K, et al. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. Molecular biology and evolution 2005;**22**(11):2265-74

76. Popesco MC, MacLaren EJ, Hopkins J, et al. Human lineage–specific amplification, selection, and neuronal expression of DUF1220 domains. Science 2006;**313**(5791):1304-07

77. Szatmari P, Paterson AD, Zwaigenbaum L, et al. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nature genetics 2007;**39**(3):319-28

78. Bramble DM, Lieberman DE. Endurance running and the evolution of Homo. Nature 2004;**432**(7015):345-52

79. Lee AS, Gutiérrez-Arcelus M, Perry GH, et al. Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. Human molecular genetics 2008;**17**(8):1127-36

80. Gibbs RA, Rogers J, Katze MG, et al. Evolutionary and biomedical insights from the rhesus macaque genome. science 2007;**316**(5822):222-34

81. Gokcumen O, Babb PL, Iskow RC, et al. Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. Genome biology 2011;**12**(5):R52

82. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. Nature 2009;**464**(7289):704-12

83. Tuzun E, Bailey JA, Eichler EE. Recent segmental duplications in the working draft assembly of the brown Norway rat. Genome research 2004;**14**(4):493-506

84. Elsik CG, Tellam RL, Worley KC. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 2009;**324**(5926):522-28

85. Nicholas TJ, Cheng Z, Ventura M, et al. The genomic architecture of segmental duplications and associated copy number variants in dogs. Genome research 2009;**19**(3):491-99

86. She X, Cheng Z, Zöllner S, et al. Mouse segmental duplication and copy number variation. Nature genetics 2008;**40**(7):909-14

87. Wu C-I, Li W-H. Evidence for higher rates of nucleotide substitution in rodents than in man. Proceedings of the National Academy of Sciences 1985;**82**(6):1741-45

88. Li W-H, Tanimura M. The molecular clock runs more slowly in man than in apes and monkeys. Nature 1987;**326**(6108):93-96

89. The evolution of human segmental duplications and the core duplicon hypothesis. Cold Spring Harbor symposia on quantitative biology; 2009. Cold Spring Harbor Laboratory Press.

90. Jiang Z, Tang H, Ventura M, et al. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nature genetics 2007;**39**(11):1361-68

91. Lupski JR. Genomic rearrangements and sporadic disease. Nature genetics 2007;**39**:S43-S47

92. Mefford HC, Eichler EE. Duplication hotspots, rare genomic disorders, and common disease. Current opinion in genetics & development 2009;**19**(3):196-204

93. Juyal RC, Figuera LE, Hauge X, et al. Molecular analyses of 17p11. 2 deletions in 62 Smith-Magenis syndrome patients. American journal of human genetics 1996;**58**(5):998

94. Kumar S, Hedges SB. A molecular timescale for vertebrate evolution. Nature 1998;**392**(6679):917-20

95. Worobey M, Telfer P, Souquière S, et al. Island biogeography reveals the deep history of SIV. Science 2010;**329**(5998):1487-87

96. Silvestri G, Paiardini M, Pandrea I, et al. Understanding the benign nature of SIV infection in natural hosts. Journal of Clinical Investigation 2007;**117**(11):3148-54

97. Shaw CJ, Lupski JR. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Human molecular genetics 2004;**13**(suppl 1):R57-R64

98. Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. Trends in genetics: TIG 2002;**18**(2):74

99. Bailey JA, Gu Z, Clark RA, et al. Recent segmental duplications in the human genome. Science 2002;**297**(5583):1003-07

100. Stankiewicz P, Shaw CJ, Withers M, et al. Serial segmental duplications during primate evolution result in complex human genome architecture. Genome research 2004;**14**(11):2209-20

101. Lupski JR, Stankiewicz P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS genetics 2005;**1**(6):e49

102. Garcia-Lopez G, Flores-Espinosa P, Zaga-Clavellina V. Tissue-specific human beta-defensins(HBD) 1, HBD 2, and HBD 3 secretion from human extra-placental membranes stimulated with Escherichia coli. Reproductive Biology and Endocrinology 2010;**8**:146-46

103. Reiter LT, Murakami T, Koeuth T, et al. A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. Nature genetics 1996;**12**(3):288-97

104. Bi W, Park SS, Shaw CJ, et al. Reciprocal crossovers and a positional preference for strand exchange in recombination events resulting in deletion or duplication of chromosome 17p11. 2. The American Journal of Human Genetics 2003;**73**(6):1302-15

105. Jeffreys AJ, Neumann R. The rise and fall of a human recombination hot spot. Nature genetics 2009;**41**(5):625-29

106. Lieber MR, Ma Y, Pannicke U, et al. Mechanism and regulation of human non-homologous DNA end-joining. Nature reviews Molecular cell biology 2003;**4**(9):712-20

107. Schwarz K, Ma Y, Pannicke U, et al. Human severe combined immune deficiency and DNA repair. Bioessays 2003;**25**(11):1061-70

108. Lieber MR, Lu H, Gu J, et al. Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system. Cell research 2007;**18**(1):125-33

109. Lieber MR. The mechanism of human nonhomologous DNA end joining. Journal of Biological Chemistry 2008;**283**(1):1-5

110. Lee JA, Carvalho C, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 2007;**131**(7):1235-47

111. Goodier JL, Kazazian HH. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 2008;**135**(1):23-35

112. Babushok DV, Kazazian Jr HH. Progress in understanding the biology of the human mutagen LINE-1. Human mutation 2007;**28**(6):527-39

113. Zhang F, Gu W, Hurles ME, et al. Copy number variation in human health, disease, and evolution. Annual review of genomics and human genetics 2009;**10**:451-81

114. Nozawa M, Kawahara Y, Nei M. Genomic drift and copy number variation of sensory receptor genes in humans. Proceedings of the National Academy of Sciences 2007;**104**(51):20421-26

115. Hasin Y, Olender T, Khen M, et al. High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. PLoS genetics 2008;**4**(11):e1000249

116. Poliseno L, Salmena L, Zhang J, et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 2010;**465**(7301):1033-38

117. Lower KM, Hughes JR, De Gobbi M, et al. Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. Proceedings of the National Academy of Sciences 2009;**106**(51):21771-76

118. Lee JA, Lupski JR. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. Neuron 2006;**52**(1):103

119. Sebat J, Lakshmi B, Malhotra D, et al. Strong association of de novo copy number mutations with autism. Science 2007;**316**(5823):445-49

120. Nakajima T, Kaur G, Mehra N, et al. HIV-1/AIDS susceptibility and copy number variation in< i> CCL3L1</i>, a gene encoding a natural ligand for HIV-1 co-receptor CCR5. Cytogenetic and genome research 2008;**123**(1-4):156-60

121. Hollox EJ, Huffmeier U, Zeeuwen PLJM, et al. Psoriasis is associated with increased β-defensin genomic copy number. Nature genetics 2007;**40**(1):23-25

122. de Cid R, Riveira-Munoz E, Zeeuwen PLJM, et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nature genetics 2009;**41**(2):211-15

123. Ferreira MA, O'Donovan MC, Meng YA, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nature genetics 2008;**40**(9):1056-58

124. Williams HJ, Craddock N, Russo G, et al. Most genome-wide significant susceptibility loci for schizophrenia and bipolar disorder reported to date cross-traditional diagnostic boundaries. Human molecular genetics 2011;**20**(2):387-91

125. Nathans J, Piantanida TP, Eddy RL, et al. Molecular genetics of inherited variation in human color vision. Science 1986;**232**(4747):203-10

126. Lifton RP, Dluhy RG, Powers M, et al. A chimaeric llβ-hydroxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and human hypertension. 1992

127. Velagaleti GVN, Bien-Willner GA, Northup JK, et al. Position effects due to chromosome breakpoints that map∼ 900 kb upstream and∼ 1.3 Mb downstream of SOX9 in two patients with campomelic dysplasia. American journal of human genetics 2005;**76**(4):652

128. Kurotaki N, Shen JJ, Touyama M, et al. Phenotypic consequences of genetic variation at hemizygous alleles: Sotos syndrome is a contiguous gene syndrome incorporating coagulation factor twelve (FXII) deficiency. Genetics in Medicine 2005;**7**(7):479-83

129. Albers CA, Paul DS, Schulze H, et al. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nature genetics 2012;**44**(4):435-39

130. Yan J, Bi W, Lupski JR. Penetrance of Craniofacial Anomalies in Mouse Models of Smith-Magenis Syndrome Is Modified by Genomic Sequence Surrounding< i> Rai1</i>: Not All Null Alleles Are Alike. The American Journal of Human Genetics 2007;**80**(3):518-25

131. Ganz T. Defensins: antimicrobial peptides of innate immunity. Nature Reviews Immunology 2003;**3**(9):710-20

132. Bulet P, Stocklin R. Insect antimicrobial peptides: structures, properties and gene regulation. Protein and peptide letters 2005;**12**(1):3-11

133. Castro MS, Fontes W. Plant defense and antimicrobial peptides. Protein and Peptide Letters 2005;**12**(1):11-16

134. Thomma BP, Cammue BP, Thevissen K. Plant defensins. Planta 2002;**216**(2):193-202

135. Lehrer RI. Primate defensins. Nature Reviews Microbiology 2004;**2**(9):727-38

136. Xiao Y, Hughes AL, Ando J, et al. A genome-wide screen identifies a single β-defensin gene cluster in the chicken: implications for the origin and evolution of mammalian defensins. Bmc Genomics 2004;**5**(1):56

137. Zhao C, Nguyen T, Liu L, et al. Gallinacin-3, an inducible epithelial β-defensin in the chicken. Infection and immunity 2001;**69**(4):2684-91

138. Zou J, Mercier C, Koussounadis A, et al. Discovery of multiple beta-defensin like homologues in teleost fish. Molecular immunology 2007;**44**(4):638-47

139. Liu L, Zhao C, Heng HHQ, et al. The human β-defensin-1 and α-defensins are encoded by adjacent genes: two peptide families with differing disulfide topology share a common ancestry. Genomics 1997;**43**(3):316-20

140. Maxwell AI, Morrison GM, Dorin JR. Rapid sequence divergence in mammalian beta-defensins by adaptive evolution. Molecular immunology 2003;**40**(7):413

141. Semple CAM, Rolfe M, Dorin JR. Duplication and selection in the evolution of primate b-defensin genes. Genome Biol 2003;**4**(5):R31

142. Glazko GV, Nei M. Estimation of divergence times for major lineages of primate species. Molecular biology and evolution 2003;**20**(3):424-34

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

144. Yang D, Biragyn A, Hoover DM, et al. Multiple Roles of Antimicrobial Defensins, Cathelicidins, and Eosinophil-Derived Neurotoxin in Host Defense*. Annu Rev Immunol 2004;**22**:181-215

145. Selsted ME, Ouellette AJ. Mammalian defensins in the antimicrobial immune response. Nature immunology 2005;**6**(6):551-57

146. Lehrer RI, Ganz T. Defensins of vertebrate animals. Current opinion in immunology 2002;**14**(1):96-102

147. Klüver E, Schulz-Maronde S, Scheid S, et al. Structure-activity relation of human β-defensin 3: influence of disulfide bonds and cysteine substitution on antimicrobial activity and cytotoxicity. Biochemistry 2005;**44**(28):9804-16

148. Yenugu S, Hamil K, Birse C, et al. Antibacterial properties of the sperm-binding proteins and peptides of human epididymis 2 (HE2) family; salt sensitivity, structural dependence and their interaction with outer and cytoplasmic membranes of Escherichia coli. Biochem J 2003;**372**:473-83

149. Zanich A, Pascall JC, Jones R. Secreted epididymal glycoprotein 2D6 that binds to the sperm's plasma membrane is a member of the β-defensin superfamily of pore-forming glycopeptides. Biology of reproduction 2003;**69**(6):1831-42

150. Alekseeva L, Huet D, Féménia F, et al. Inducible expression of beta defensins by human respiratory epithelial cells exposed to Aspergillus fumigatus organisms. BMC microbiology 2009;**9**(1):33

151. Kao CY, Chen Y, Zhao YH, et al. ORFeome-based search of airway epithelial cell-specific novel human β-defensin genes. American journal of respiratory cell and molecular biology 2003;**29**(1):71-80

152. Otri AM, Mohammed I, Al-Aqaba MA, et al. Variable Expression of Human beta Defensins 3 and 9 at the Human Ocular Surface in Infectious Keratitis. Investigative Ophthalmology & Visual Science 2012;**53**(2):757-61

153. Giglio S, Broman KW, Matsumoto N, et al. Olfactory receptor–gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. The American Journal of Human Genetics 2001;**68**(4):874-83

154. Terminal deletion in chromosome region 8p23. 1-8pter in a child with features of velo-cardio-facial syndrome. Annales de génétique; 1995.

155. Barber J. Directly transmitted unbalanced chromosome abnormalities and euchromatic variants. Journal of medical genetics 2005;**42**(8):609-29

156. Barber JC, Maloney VK, Huang S, et al. 8p23. 1 duplication syndrome; a novel genomic condition with unexpected complexity revealed by array CGH. European Journal of Human Genetics 2007;**16**(1):18-27

157. Hardwick RJ, Machado LR, Zuccherato LW, et al. A worldwide analysis of beta-defensin copy number variation suggests recent selection of a high-expressing DEFB103 gene copy in East Asia. Human mutation 2011;**32**(7):743-50

158. Marques-Bonet T, Kidd JM, Ventura M, et al. A burst of segmental duplications in the genome of the African great ape ancestor. Nature 2009;**457**(7231):877-81

159. Joly S, Maze C, McCray PB, et al. Human β-defensins 2 and 3 demonstrate strain-selective activity against oral microorganisms. Journal of clinical microbiology 2004;**42**(3):1024-29

160. Bauer F, Schweimer K, Klüver E, et al. Structure determination of human and murine β-defensins reveals structural conservation in the absence of significant sequence similarity. Protein Science 2001;**10**(12):2470-79

161. Schibli DJ, Hunter HN, Aseyev V, et al. The solution structures of the human β-defensins lead to a better understanding of the potent bactericidal activity of HBD3 against Staphylococcus aureus. Journal of Biological Chemistry 2002;**277**(10):8279-89

162. Taylor K, Barran PE, Dorin JR. Structure–activity relationships in β-defensin peptides. Peptide Science 2007;**90**(1):1-7

163. Semple CAM, Maxwell A, Gautier P, et al. The complexity of selection at the major primate β-defensin locus. BMC evolutionary biology 2005;**5**(1):32

164. Vylkova S, Nayyar N, Li W, et al. Human β-defensins kill Candida albicans in an energy-dependent and salt-sensitive manner without causing membrane disruption. Antimicrobial agents and chemotherapy 2007;**51**(1):154-61

165. Harder J, Bartels J, Christophers E, et al. Isolation and characterization of human β-defensin-3, a novel human inducible peptide antibiotic. Journal of Biological Chemistry 2001;**276**(8):5707-13

166. Bals R, Wang X, Wu Z, et al. Human beta-defensin 2 is a salt-sensitive peptide antibiotic expressed in human lung. Journal of Clinical Investigation 1998;**102**(5):874

167. García JRC, Krause A, Schulz S, et al. Human β-defensin 4: a novel inducible peptide with a specific salt-sensitive spectrum of antimicrobial activity. The FASEB Journal 2001;**15**(10):1819-21

168. Goldman MJ, Anderson GM, Stolzenberg ED, et al. Human β-defensin-1 is a salt-sensitive antibiotic in lung that is inactivated in cystic fibrosis. Cell 1997;**88**(4):553-60

169. Singh PK, Jia HP, Wiles K, et al. Production of β-defensins by human airway epithelia. Proceedings of the National Academy of Sciences 1998;**95**(25):14961-66

170. Sørensen OE, Thapa DR, Rosenthal A, et al. Differential regulation of β-defensin expression in human skin by microbial stimuli. The Journal of Immunology 2005;**174**(8):4870-79

171. Doss M, White MR, Tecle T, et al. Human defensins and LL-37 in mucosal immunity. Journal of leukocyte biology 2010;**87**(1):79-92

172. Harder J, Schröder JM. Psoriatic scales: a promising source for the isolation of human skin-derived antimicrobial proteins. Journal of leukocyte biology 2005;**77**(4):476-86

173. Nomura I, Goleva E, Howell MD, et al. Cytokine milieu of atopic dermatitis, as compared to psoriasis, skin prevents induction of innate immune response genes. The Journal of Immunology 2003;**171**(6):3262-69

174. Froy O. Microreview: Regulation of mammalian defensin expression by Toll-like receptor-dependent and independent signalling pathways. Cellular microbiology 2005;**7**(10):1387-97

175. Yang D, Chertov O, Bykovskaia SN, et al. β-defensins: linking innate and adaptive immunity through dendritic and T cell CCR6. Science 1999;**286**(5439):525-28

176. Steinman L. A brief history of TH17, the first major revision in the TH1/TH2 hypothesis of T cell–mediated tissue damage. Nature medicine 2007;**13**(2):139-45

177. Matusevicius D, Kivisäkk P, He B, et al. Interleukin-17 mRNA expression in blood and CSF mononuclear cells is augmented in multiple sclerosis. Multiple Sclerosis 1999;**5**(2):101-04

178. Aarvak T, Chabaud M, Miossec P, et al. IL-17 is produced by some proinflammatory Th1/Th0 cells but not by Th2 cells. The Journal of Immunology 1999;**162**(3):1246-51

179. Albanesi C, Scarponi C, Cavani A, et al. Interleukin-17 is produced by both Th1 and Th2 lymphocytes, and modulates interferon-γ-and interleukin-4-induced activation of human keratinocytes. Journal of investigative dermatology 2000;**115**(1):81-87

180. Pandrea I, Apetrei C. Where the wild things are: Pathogenesis of SIV infection in African nonhuman primate hosts. Current HIV/AIDS Reports 2010;**7**(1):28-36

181. El Hed A, Khaitan A, Kozhaya L, et al. Susceptibility of human Th17 cells to human immunodeficiency virus and their perturbation during infection. Journal of Infectious Diseases 2010;**201**(6):843-54

182. Reynolds JM, Martinez GJ, Chung Y, et al. Toll-like receptor 4 signaling in T cells promotes autoimmune inflammation. Proceedings of the National Academy of Sciences 2012;**109**(32):13064-69

183. King AE, Paltoo A, Kelly RW, et al. Expression of natural antimicrobials by human placenta and fetal membranes. Placenta 2007;**28**(2):161-69

184. Baroni A, Donnarumma G, Paoletti I, et al. Antimicrobial human beta-defensin-2 stimulates migration, proliferation and tube formation of human umbilical vein endothelial cells. Peptides 2009;**30**(2):267-72

185. Conejo-Garcia JR, Benencia F, Courreges MC, et al. Tumor-infiltrating dendritic cell precursors recruited by a β-defensin contribute to vasculogenesis under the influence of Vegf-A. Nature medicine 2004;**10**(9):950-58

186. Chiodoni C, Paglia P, Stoppacciaro A, et al. Dendritic cells infiltrating tumors cotransduced with granulocyte/macrophage colony-stimulating factor (GM-CSF) and CD40 ligand genes take up and present endogenous tumor-associated antigens, and prime naive mice for a cytotoxic T lymphocyte response. The Journal of experimental medicine 1999;**190**(1):125-34

187. Gabrilovich D, Ishida T, Oyama T, et al. Vascular endothelial growth factor inhibits the development of dendritic cells and dramatically affects the differentiation of multiple hematopoietic lineages in vivo. Blood 1998;**92**(11):4150-66

188. Tumour escape from immune surveillance through dendritic cell inactivation. Seminars in cancer biology; 2002. Elsevier.

189. Hanahan D. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. cell 1996;**86**:353-64

190. Sun L, Finnegan CM, Kish-Catalone T, et al. Human β-defensins suppress human immunodeficiency virus infection: potential role in mucosal protection. Journal of virology 2005;**79**(22):14318-29

191. Quiñones-Mateu ME, Lederman MM, Feng Z, et al. Human epithelial [beta]-defensins 2 and 3 inhibit HIV-1 replication. Aids 2003;**17**(16):F39-F48

192. Weinberg A, Quinones-Mateu ME, Lederman MM. Role of Human β-defensins in HIV Infection. Advances in Dental Research 2006;**19**(1):42-48

193. Leikina E, Delanoe-Ayari H, Melikov K, et al. Carbohydrate-binding molecules inhibit viral fusion and entry by crosslinking membrane glycoproteins. Nature immunology 2005;**6**(10):995-1001

194. Huang L, Ching CB, Jiang R, et al. Production of bioactive human beta-defensin 5 and 6 in< i> Escherichia coli</i> by soluble fusion expression. Protein expression and purification 2008;**61**(2):168-74

195. Klotman ME, Chang TL. Defensins in innate antiviral immunity. Nature Reviews Immunology 2006;**6**(6):447-56

196. Rodríguez-Jiménez FJ, Krause A, Schulz S, et al. Distribution of new human β-defensin genes clustered on chromosome 20 in functionally different segments of epididymis. Genomics 2003;**81**(2):175-83

197. Yenugu S, Hamil KG, Radhakrishnan Y, et al. The androgen-regulated epididymal sperm-binding protein, human β-defensin 118 (DEFB118)(formerly ESC42), is an antimicrobial β-defensin. Endocrinology 2004;**145**(7):3165-73

198. Tollner TL, Yudin AI, Treece CA, et al. Macaque sperm coating protein DEFB126 facilitates sperm penetration of cervical mucus†. Human reproduction 2008;**23**(11):2523-34

199. Candille SI, Kaelin CB, Cattanach BM, et al. A β-defensin mutation causes black coat color in domestic dogs. Science 2007;**318**(5855):1418-23

200. Swope VB, Jameson JA, McFarland KL, et al. Defining MC1R regulation in human melanocytes by its agonist α-melanocortin and antagonists agouti signaling protein and β-defensin 3. Journal of Investigative Dermatology 2012;**132**(9):2255-62

201. Maaser C, Kannengiesser K, Kucharzik T. Role of the melanocortin system in inflammation. Annals of the New York Academy of Sciences 2006;**1072**(1):123-34

202. Korbel JO, Urban AE, Grubert F, et al. Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. Proceedings of the National Academy of Sciences 2007;**104**(24):10110-15

203. Wilson JF, Erlandsson R. Sexing of Human and Other Primate DMA. Biol Chem 1998;**379**:1287-88

204. Zhang Z, Li J, Zhao X-Q, et al. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics, Proteomics & Bioinformatics 2006;**4**(4):259-63

205. Bosch N, Cáceres M, Cardone MF, et al. Characterization and evolution of the novel gene family FAM90A in primates originated by multiple duplication and rearrangement events. Human molecular genetics 2007;**16**(21):2572-82

206. Shimmin L, Chang B-J, Li W-H. Male-driven evolution of DNA sequences. Trends in Genetics 1993;**9**(7):233

207. Sambrook J. Molecular cloning: a laboratory manual (3-volume set).

208. Hornsby MJ, Huff JL, Kays RJ, et al. < i> Helicobacter pylori</i> Induces an Antimicrobial Response in Rhesus Macaques in a< i> cag</i> Pathogenicity Island-Dependent Manner. Gastroenterology 2008;**134**(4):1049-57
209. Gasteiger E, Gattiker A, Hoogland C, et al. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic acids research 2003;**31**(13):3784-88
210. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology 2011;**7**(1)
211. Hardwick RJ, Amogne W, Mugusi S, et al. β-defensin Genomic Copy Number Is Associated With HIV Load and Immune Reconstitution in Sub-Saharan Africans. Journal of Infectious Diseases 2012;**206**(7):1012-19
212. Aldred PM, Hollox EJ, Armour JA. Copy number polymorphism and expression level variation of the human α-defensin genes DEFA1 and DEFA3. Human molecular genetics 2005;**14**(14):2045-52
213. Pazgier M, Lubkowski J. Expression and purification of recombinant human α-defensins in< i> Escherichia coli</i>. Protein expression and purification 2006;**49**(1):1-8
214. Nishihara H, Terai Y, Okada N. Characterization of novel Alu-and tRNA-related SINEs from the tree shrew and evolutionary implications of their origins. Molecular biology and evolution 2002;**19**(11):1964-72
215. Tavaré S, Marshall CR, Will O, et al. Using the fossil record to estimate the age of the last common ancestor of extant primates. Nature 2002;**416**(6882):726-29
216. Perelman P, Johnson WE, Roos C, et al. A molecular phylogeny of living primates. PLoS genetics 2011;**7**(3):e1001342
217. Prado-Martinez J, Sudmant PH, Kidd JM, et al. Great ape genetic diversity and population history. Nature 2013;**499**(7459):471-75

APPENDIX 1 Primates phylogeny

Living mammals are classified in approximately 20 Orders, depending on the taxonomy applied, one of which is the Primates. Primates fall in the sub-class of placental mammals (Eutheria), that can be further classified on the base of gene sequences, indels, and retroposons insertion data, in the clade of Euarchontoglires, the only Eutherian order to have *Alu*-like SINEs derived from the signal recognition particle ribonucleoprotein complex (7SL RNA) [214]. The internal classification of the Primates Order remains a highly disputed matter, for the absence of strong morphological features that can identify unequivocally their hierarchy.

The oldest known fossil primates date to the Eocene period (56-33.9 MYA) but statistical modelling of species preservation and genetic data can help to estimate when the most recent common ancestor of all primates may have lived. These models are generated combining different parameters, as the number of extant species, the average generation time and the number of the fossil species found in each stratigraphic interval, to correct for the number of species not preserved in the evolution of a certain Order. This also gives an estimate of the proportion of species that existed in an interval that were found as fossils [215]. Output data from this computational branching process suggested a mid-Cretaceous divergence of Primates from other placental mammals (90 MYA). A phylogenetic tree illustrating the branching of the Order of Primates is presented in **Figure 9.1**.

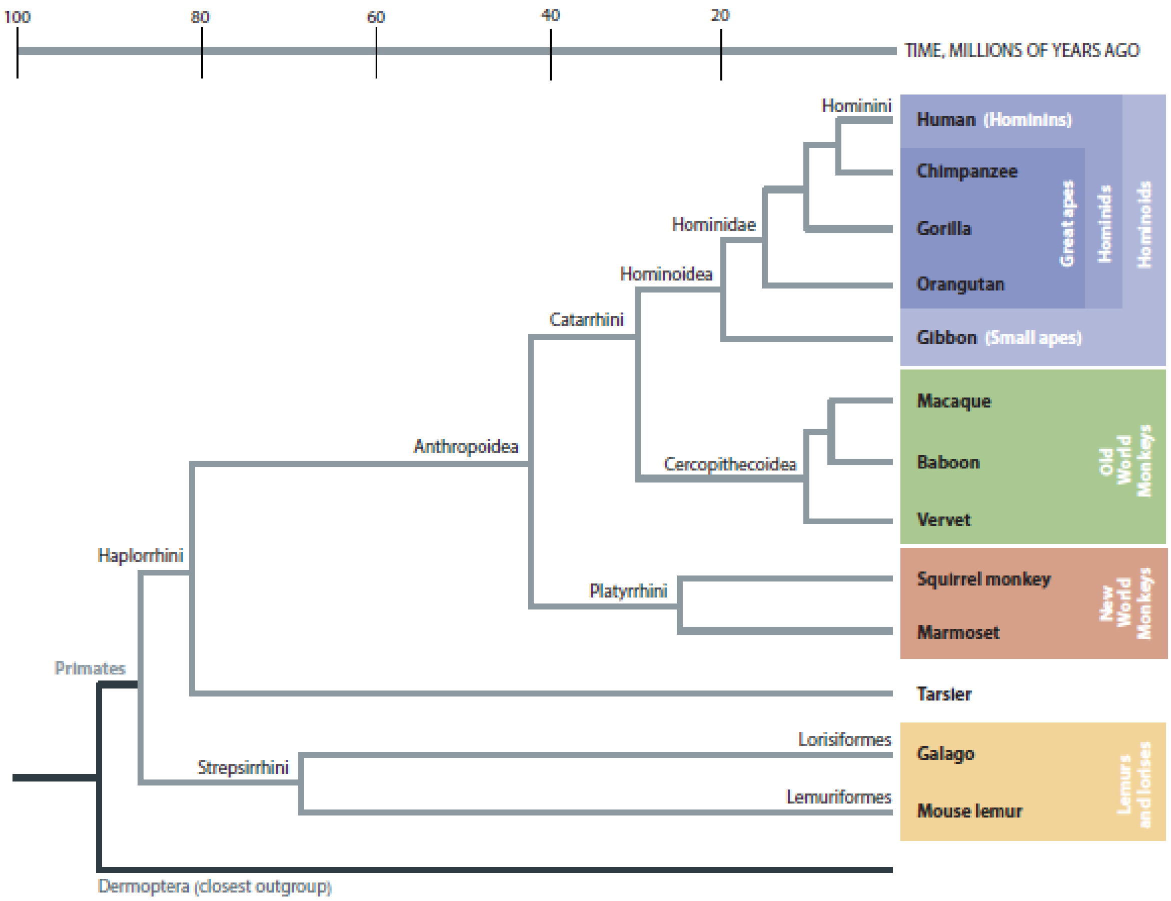


Figure 9.1: Phylogeny tree for primate species that have undergone complete genome sequencing. The branching order is shown on a timescale inferred from molecular data [216]. Picture taken from Human Evolutionary Genetics second edition 2013, chapter 7.

With the advent of next-generation sequencing techniques, it has been possible to retrieve genomic information on many Primates samples, allowing to conduct more sophisticate analysis to reconstruct the population history of the Primate Order.

The most comprehensive study performed up to date focussed on detecting and quantifying diversity among the extant Primate Family of Hominidae (humans and Great Apes) [217]. Authors identified 84 millions of fixed substitutions and 88.8 millions of high quality segregating sites that could be used to infer the population history of the Hominidae family, presented in

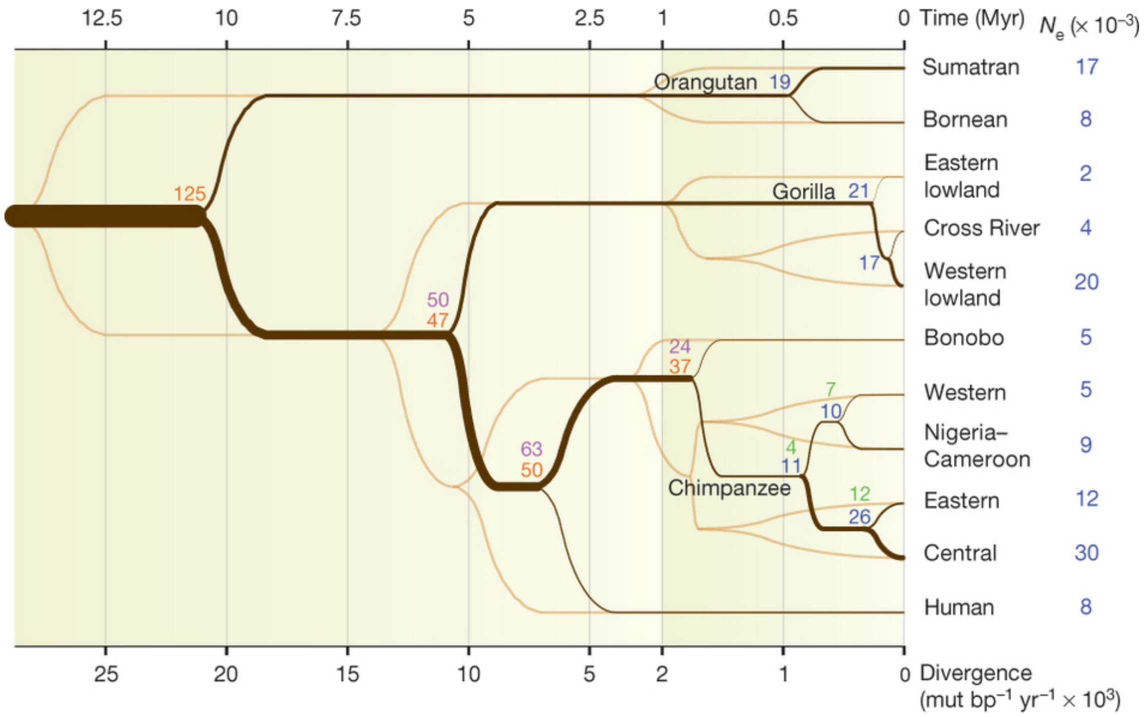


Figure 9.2: Inferred population history for the Hominidae family. The figure shows population splits and effective population sizes (N_e) characterising the evolution of Great Apes. Split times (dark brown) and divergence times (light brown) are plotted as a function of divergence (d) on the bottom and time on top. Time is estimated using a single mutation rate (μ) of 1×10^{-9} mut bp^{-1} $year^{-1}$. The same mutation rate is used to estimate the ancestral and current effective population sizes. Different methods have been used to estimate N_e (COALHMM, ILS COALHMM, PSMC and ABC), coloured respectively in orange, purple, blue and green. The terminal N_e corresponds to the effective population size after the last split event [217].

APPENDIX 2: BUFFERS

10X PCR Mix

10X PCR mix contained final concentrations of 50mM Tris-HCl (pH8.8@25°C), 12mM ammonium sulphate, 5mM magnesium chloride ($MgCl_2$), 125µg/ml BSA, 7.4mM 2-mercaptoethanol and 1.1mM of each dNTP (Promega).

10X Low dNTPs PCR Mix

10X Low dNTPs PCR mix was another buffer containing final concentrations of 50mM Tris-HCl (pH8.8@25°C), 12.5mM ammonium sulphate, 1.4mM magnesium chloride, 125µg/ml BSA, 7.5mM 2-mercaptoethanol and 200µM of each dNTP (Promega).

11.1X PCR mix

11.1 X PCR mix contained final concentrations of 45mM Tris-HCl(pH8.8), 11 mM ammonium sulphate, 4.5 mM magnesium chloride, 6.7 mM 2-mercaptoethanol, 4.4 µM EDTA, 1 mM of each dNTP (Promega), 113µg/ml BSA{{3 Jeffreys, A.J. 1990;}}.

TE buffer

10 mM Tris-HCl (pH 8@25°C)

1 mM EDTA

5X TBE (pH 8.3)

Tris Base 0.45M, Boric acid 0.4M, EDTA 0.1M

5X Oligonucleotide Labelling Buffer (OLB)

0.5 mM Tris-HCl pH 8

0.5mM $MgCl_2$

190 mM β-mercaptoethanol

0.01 mM dGTP

0.01 mM dATP

0.01 mM dTTP

10 µg random hexamers (CAT N° 27-2166.01, Pharmacia)

Oligonucleotide Stop Solution (OSS)

0.02M NaCl

0.2M Tris pH 7.5@25°C

2mM EDTA

0.25% SDS

1µM dCTP

20X SSC

3M sodium chloride

300 mM trisodium citrate

Adjusted to pH7@25 °C with HCl

50X Denhardt's solution

1% Ficoll 400

1% polyvinylpyhrolidone

1% BSA (Fraction V)

Filtered to sterilyse

Buffers for incorporation of labels test:

Buffer 1 (pH 7.5@25°C):

100 mM Tris-HCl

15 mM NaCl

Buffer 2 (pH 7.5@25°C):

0.5% w/v Blocking reagent (CAT N° NIP:552, Amersham Life Science) in Buffer 1

Buffer 3 (pH 9.5@25°C):

100 mM Tris-HCl

100 mM NaCl

50 mM $MgCl_2$

Solutions for caesium chloride gradient based BAC DNA extraction:

Solution 1

50 mM glucose

10 mM EDTA

25 mM Tris-HCl (pH 8@25°C)

Solution 2

0.2 M NaOH

1% W/V SDS

Solution 3

3 M potassium acetate

11.5% glacial acetic acid

APPENDIX 3: Sequences

APPENDIX 3.1 *DEFB2L* genomic variants

DEFB2L reference sequence:

```
>rheMac2_dna range=chr8:8070118-8072503 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGAAGGTCCTGTATCTCCTGTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGAGATGGGCCAGGGAAATGGGAGGTTGGAAAAATGGAAGAATCGGGTAGAGGCTGTCC
ATATCCTCTCTGATTTCCCCTAATTCTCCCTCTGTTTCTCTCTCCCTCCCTCCTCTATCTCTCTCTCTCCTGCCTCTGTCTCTCTTCTTCCATCTCTCTCTCTCCTGACCCTTGCTCT
CTCCTGCCTCTCTCTCTCCTGCCTCTCTCTCTCTGCCTCTCTCTCTCTCTCTCCTGACTCTCTCTCCTACATCTCTCTCTCCTGACTGTGTCTCCTGCCCTTTGCTCTCTCCCTGTCT
TTCTCGTTTCTCTTTCCCTACGCGCTCTCTTCTACCTCTCTCTCTTTCCTGCACCCTCTCCCTCTATCTCTCTATTTCCCGTCTCTCCTTGTCTCCCTCTATCCCTCTCTCTCCTGAC
TGTCTCTCTATCTCCCTCTCCTGTCTCTGTCTCCTGTGTCTGTCTGTCTCCCTCTATTTTTATCTCTCCTGACCCTTGCTCTCTCCCTCTCTCACTCCTGTCCCCTCTCTCTCTATCG
CTCTCTCCTGCCGCTCTCTCTCTCCCCACTCCCTCCTGCCTCTTTCTCTGTTTCCCTCTCTCTCTCTATTTTCCCTGTCTCTCTCTCTCCCTCTGTCCCCCAGAGCTGGTCTTTCTCT
TTCTTCTACACACACTAATAGACAGAGTAGACCGTATGCGTTACGTAATTGAACCAAGCATTGGTTCAATATAGAAGTTTGACAACTCGATGGACACCTCACTCTCTCTTCTGAGCCA
ATATGAAGGAGCCCAGTAGCTTGTAAATCTCATCTCCTCACTGCTTTCCATGCTACAACTGCTAAGACTGTGGTTGAAACCTGTTAGGTGACTTTTTAAATAAAAGTCAGAAATTTTG
ATTTTATCTAAAGAAAGTAGTCCAGAATGTCATTTTCTAAATTTTTATATTGAAAGGGTAGATACTGCAACCTAGAAAATTCCATATAATCTTAAGGCCCAGCCTATACTGTAGGAGC
TACTGCTGCAGACACTCTGCCCCAGGACTTTTCCGATCGGAGGCCCTGAGAACAGTCCCTGCCACGAGGCCACCGCAGGTTCACAGTCCAGAGACGGCCCATGGAAAGCAACTTTTAA
CCGGGACATCTAACCTTCCATTTCTCCTTGATATTATGGAAATAAAATAAAAACCATGAAAGGATAAAAGAGGGAGAGTGGAAGGGAAGGATAGAGAGAGGGAAAAAGAAAATTTGAG
ACTAAAACCTAAAACAATTAATCACATAGATATTATATTGTGAAAGCATCATTTTACCAATTTTATTTATGAGTCCCGGGTGTTTTGAGAAGAACGGGGCTCTGAGTGGCACCAGAGA
CCTCAAATTTTCCAACACCTAGAACAGTATCATGAAGGAAGGCAGGGAAGTAGGGAGGCAGGGAGGGAGGCAGGGAGGCAGGCAGGGACGGAGGGAGGGAGGGAGGCCGGGAGGGACT
GAGGCAGGGAGGGAGAAAGGCAGGGAGGCAGGCAGGGAGGGAAGAAGGGAGGGAGGCAGGGAGGGAAAAAGGCAGGGAAGCAGGGAGGGAGGGAGAGAGGCAGGGAGGCAGGGAGGGA
GGGACAGAGACAGGGAAAGATAAAAAAAAGAAGAATGAGGTTGAAACCAGGACTTAGATATTAGAAACAAGCCATTACAAAAGTTTATTTCTGTGGTTAATTGTGGTTTTCAACTGTA
AGCTATATGGTGTTAATTTCACATTGAACAATTTCTGTGAGTTGTATCTTTTTATCCCATCTCAGATCAAATACTTAACAGACTAAATGATTTGAAAAGCAAAAAGTTACTGGCTTGT
GTGTTAAAACGGAGGTAGGGTGGCTTTGATCTTATCTACTTGTGGTGGAGCTGAAGTCACAAGAGAACATTACCGAGCTCCTACCAGACCCCACCCGGAGGCCCCAGTCACTCAGGAG
AGATCAGGGTCTTTCACAATCAGATTCGACAAAAACTAACATCCCCCAAACCACAGCAGTGCCAGTTTCCATGTCAGAAACTTCCATCCAAATGATTGACTCGCGTCTCATGGAAAAG
CCCTGGCTTCAGAAAGAAGTCCACTGCAGATTTATTCAAGGCAATACAGACACAGCATTTGTGTTTTGCAACATGAGTTTTCAGTTCTAACACGCTGTTTGCTCTTTGTGTGTTTTTT
TCCTTTTAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAAGGTATAAACATATCGGCGTCTGTGGTGTCCC
TCTAATAAAATGCTGCAAAAAGCCAT
```

DEFB2L exon1 reference:

```
>DEFB2L_exon1_reference
ATGAAGGTCCTGTATCTCCTGTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCC
```

DEFB2L exon1 BAC variant:

```
>DEFB2L_exon1bac_variant
ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTCATGCCTCTTCC
```

DEFB2L exon2 reference:

```
>DEFB2L_exon2_reference
AGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAAGGTATAAACATATCGGCGTCTGTGGTGTCCCTCTAATAAAATGCTGC
AAAAAGCCATGA
```

DEFB2L exon2 BAC variant:

```
>DEFB2L_exon2_bac_variant
AGGTGTTTTTGGTGATATAAGGAATCTTTTTACCTGCCTTTGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAACGTATAAACATATTGGCATCTGTGGTGTCCCTCTAATAAAATGCTGCA
AAAAGCCATGA
```

BAC 65I2 first CNV

```
>CNV1trimmed
agtatcatgatgtttgcctctttctaaaggtgacagcgacatatacattttaaaaacataattattaatttagacaaagcacataaaggctttatttcccgttccatttctctgtatgctttcttcaccaggaggaaatagttttagtgtcgggaatgaatgagtctgc
cccttgattccagcctgttcaacacacagggagacaaactcctgacaaagggaatgactctctggtgactgagctgcagccctgggttcatgttcgtttagcagttccaatggcacctgacccagccctctctttgcataccccatcagaaactccctttttatggac
aaagtcattgaaggtcttggctgcgcaatgtcctcaccagcttcctttaaatccacttctggcctgccaggaaaagggttcttcagaacctgacattttaaacgaagagatcaggcaggtcatgagtaaagtgtcattgtccccatggctgtgtcactgctacacct
ccgagacatcacagacacggacactggggcctgcttgtctctcaaactgcccttagatccaaagagggaggaaccaggaaaaaggtctcttattttcctgagaaaagcctggcccctgagcacctggcatggggctctagtcattgtttatgatccccgatgact
tctctgggtcactgaggaaggacagggtcctgagggacccatcgccaacaggaagaccacgtggaggccttgcacccccactccgtgcttctccaccaaatcccaatggcagtgagctcccttggcactgacgtctgtggaaagcagggaaagccccggctcg
caaaaccctgaagtcctgcggagctgacgttccctgagtgacggtgtgagcctaaggaattcaagtgtgggccataggccacctcctggccagggtcaggtggactctgaggggacacatgtagtcacaatcccaccctcccattctccttctcagaggaagga
agcgggcacccatctgcctcatctctgtcccggggggaagatggggagtttcaggggaactttcacatgaatttcactagctcagatcgcctgggaggacggggcccaccatgctcctggtgctgccagaagccctgagcccctccagggtccctgggtttgag
ccagccctatatcatccccaggagctgaatgtcccagcaatgggtagaattagatggaaagagctcccagtttggcctgagagtgtccccagatactcaggaaaaacagggcgtcccacagagtgggtggcaggtgagttccacgctacaggccctgagtttg
agtttgttcccctgtgagaaggccccagcccctcactccattcatacactgggttttaaatggcgcaagataagaggaattttctggtcccaagagcaggaggaagggattttctggggtttcctgaatccagatttgcataagatctcctgcatgtgcattgttctt
tgaggaccattctctgactcaccaggtaagtggttgaattctaatccgtgtaatgagcgttgcagccaataccagttctgaactctacccggtgaccacgggccaggacctttataagatggacggctcggtgtcttccccagactcagctcctggtgaagctccc
agccatcagccatgaaggtcctgtatctcctgttctcgttcctcttcatattcctgatgcctcttccgggtgagatgggccagggaaatgggaggttggaaaaatggaagaatcgggtagaggctgtccatatcctctctgatttcccctaatcctccctctgtttctc
tctccctccctcctctatctctctctcctgcctctctctctcttcttccatctctctctctcctgacccttgctctctcctgcctctctctctcctgcctctctctcttcctctctctctctctctcttctacacacacttttagacggagtagaccgtatgcgttacataatt
gaaccaagcattggttcaatatagaagtttgacaactcgatggacacctcactctctcttctgagccaatatgaaggagcccagtagcttgtaaatctcatctcctcactgctttccatgctataactgctaagactgtggttgaaacctgttaggtgactttttaaa
taaaagtcagaaattttgattttatctaaagaaagtagtccagaatgtcattttctaaatttgtatattgaaagggtagatactgcaacctagaaaattccagataatcttaaggcccagcctatactgtaggagctactgctgcagacactctgcccctaggactt
ttccgatcggaggccctgaaaacagtccctgccacgaggccaccgcaggttcacagtccagagacagcccatggaaagcaacttttaaccgggacatctaaccttcattttctccttgatattatggaaataaaataaaaaccatgaaaggataaaagaggga
gagtggaagggaaggatagagagagggaaaaagaaaatttgagattaaaacctaaaacaattaatcacatagatattatattgtgaaagcatcattttaccaattttatttatgagtcctgggtgttttgagaagaacggggctctgagtggcaccagagacc
tcaaattttccacacctagaacagtatcatgaaggaaggcagggaagtagggaggcagggagggaggcagggaggcagggaggcaggcagggacggagggagggagggaggccgggagggactgaggcagggagggagaaaggcagggaggcagg
cagggagggaagaagggagggaggcagggagggaaaaaggcagggaagcagggagggagggagagaggcagggaggcagggagggagggacagagacagggaaagataaaaaaaagaagaatgaggttgaaaccaggacttagatattagaaa
caagccattacaaaagtttatttcggtggttaattgtggttttcaactgtaagctatgtggtgttaatttcagagatattaaaaacaagccattacaaaagtttatttctgtggttaattgtggttttcaactgtaagctatatggtgttaatttcacattgaacaatttc
tgtgagttgtatctttttatcccatctcagatcaaatacttaacagactaaatgatttgaaaagcaaaaagttactggcttgtgtgttaaaacggaggtagggtggctttgatcttatctacttgtggtggagctgaagtcacaagagaacattaccgagctcctac
cagaccccacccggaggccccagtcactcaggagagatcagggtctttcacaatcagattcgacaaaaactaacatcccccaaaccacagcagtgccagtttccatgtcagaaacttccatccaaatgattgactcgcgtctcatggaaaagccctggcttcag
aaagaagtccactgcagatttattcaaggcaatacagacacagcatttgtgttttgcaacatgagttttcagttctaacacgctgtttgctctttgtgtgtttttttcctttttaggtgtttttggtgatataaggaatcctgttacctgccttaggagtggtgccatatgt
catccaggcttttgccctggaaggtataaacatatcggcgtctgtggtgtccctctaataaaatgctgcaaaaagccatgaggaggccaagaagttgctgtaactgatgtggattcagaaagggctccctcatcagagatgtgtgacatgtaaacaaaattaaa
ctgtggtgttcagatttacggaatcttgatcctagtcattgtggtcattgtgtggtgctggtttgggcaggccaatctctaacatccttggaacacccttttcttctctccaggcaggggtcagggatgccacagcggggcttggagtgctttccagagtacaggcgt
ctgtattatttggatcccttgaccttcccagttattcccgacaatttcatagaacgtgtgctttgctcctcctgcaccctcccccttgtatgcctaccccatgtcttccctaaaaaaagcaagcccaactcaaagaccactttcctcatggaatcatagcggatctgcta
agggaggggatgcccagtcctctgttcttcacaaggactcccttcttctggttaaggtttcctatgcaattacgcctcctacagaggtgcatgaatttttaattctccatttagctatgagagttctactggtgtggactttgtcttattcatttatgtgctggtcattcat
aagctatttcacaaattggatggcaaaatgcagttgtacaagggtttacttacatacaaacacaataggatccaagtaaatactcttaaaaacaagtcttttggagggcataattaaatgaggacagtataccatgggacacaagcagaggggagcaaacct
caagaagaaagacacatcaactctaaagggaagcatcaggatagctccatggccctgctctctctccttgggagggtttgctctttcaatcatgaactttgtgggttgaaaaatcctacatgcagaagaagctctctatttcttcacattgacgctgcactaaatg
tgcatatcatcgacctctgggaaaatgaaccttagttccaaaataatccactgtctttcctcatctcaagacagaaaaaagtaaccagcccttgttagagtaaaagcatttttatctgagtacagatttttccagatgtgatgacagagccaggactagggtgaa
gcgagggaggctctggcctcaggtataaaatataagcacttagatactccatacttaagatagatgatctcataatgcgatagattttttaaaaaatcatagtcatgaaaaaaatcatataatcaacaaaatgtccaaaatgtaaatcaagacagggtctaatc
ctataattgtccaactcagttgcactcacctaccttgataccgggacagtcagaccctgcttttattgagtcatttactattcatcatgtatacattgttattttgaacaaattagtggatttcctggttcccggaagcatatatcagttgattgtataaaagaagtgga
gcttcgcaaaaaaaaattaaaatgtcaccatgacaatggcattcattcctatgttatggggaagcagcaaaactctcttaggaaaaaactctccttttatttgagaagaccctttaatggcaatttaggtcttgaaagacaggacagggaggaagactcccatg
gagaccagcatcctgtggcattgagaagaccgaggtcaaagctcacacaacacaatactccctgtgttagtctgtgtgtgttgctctaaaggaatacctgaggtggataatacggcaggaaaggaagttattttggctcactattctgcaggctgtgtgagaag
catgatacctgcatctgcccctcattaggcctcaatcagccctcaggcctagtggaaagtgaagtgggagcagtaatatcacctggtgagagcagagcgagaaggtgaaaggaggtcccagactcttttaaacaactacatgttgtatgaagtcggagcaaga
actcaatcattattgggacgaaagcacttagtcattcgtcagggatttaccccgcaactcaaatcctccccaccagattccacttccaacattgaggattagatttcagcatgaggtttcgaggggacaaatatctgaaaaatattattcttcccctggcccctcaa
atctcatgtccttctcacattgcaaaatacaatcatcccttcccaagcactcccaaaagtcttaactccttctggcattaactgaatcaaaagtccaaagtccaatgtctcatcctctgagactcaagttccttccacctattagcctgtactatcaaaaataagttat
ttcctttcaagttacaatgatggttcaggcattggctaaacaccaccattccaaaaggcagacactggacaataaaagggattacaggccccacacatgtcagaaactcagcagggaagctgttaaacctcaaagctccaaaacaattcttgattccatgttcc
gaattctggtgtgagggtgggctctgaagaccttgggcagctctgcccctgtggctttccagggtgcagcccacatggctcctgtcacaggtcagaatctgatgcctgtggctcttccatgctgagggtacaagctgtcaatggtgctcctattctcaggtctggag
ggcagtgttccccgtctctcagctccactaggcagtgccccactgtggacactgtgtggggaatccaaccccatatttccccttagcactgcgctagtagaatttttctgttggtattctccctctgcagcagttttctgcctggagaaccaggcttggccacacatcc
tctgaaatgtaggtggaagctgccaagcctctttcctttttgcactctgcatacttgcaggcttaacaccacatgaaagctgccaagacgttatggctttcacccactgaagcagtgaccaatctgtacattaggccctttgagctgaggctggagcctgggcagc
caggatgtgggaacagtgtcttggggctgagcagggcagctgtgccctggacgtggccactgaaccattctttcctcctaggcctgtgggctgtagtaggagggaatgcctcaaacttttcaaaaatgccttgtaagccttttccccattgtcttggccattagcac
atggctccttctcagtcatggaatctctctagcaagtgattctccacagcccacttacattcttcttctcaacatgctttttctttctctactaaattcctaggctgcaaattttccaaacttttatactttgcttcccctttaaatataagttccatatttaaatcatttcttt
gttcctgtgtctcactgtaagcttgtagaagcagccacatatcttgagtactttgctccttggagattacttagaccaaatactctaggtaatcactctaaagttcaactttgcataaatctctaggacatggacacaatgcggccaagctctatactagggtgtaa
cactggtgacccttgtcaattcccaataacttccacttttccatcttacacctgatagcctggacttcactgtccatatctctatgagcactttggtcacatacatttaacaagtttctaaagagttccaactttccctcatcttcctatcttttgctgggctctgcaaact
ctttcagcctctgccagtacccagtgacaaagccacttccatattttcagctgtctttacagcaataccccctgaacagtaccaattttctgtgtcagtttgtttatactgctataaaggaatacctaatgctgggtagttcataaggaaaagaagtttattttggttc
cctattctgtaggctgtatgagtaggatggtgccagcatctgctccttttgaggcctcagaaagcttccatctgggtaggaaagggaagggggaacaggagtatcacaggtcaggagaaggagccagtgtggggaggtgc
```

BAC 65I2 second CNV

```
>CNV2 trimmed
agtgagggctgccccttactttctaaaggtgacatgcaaatataccaaaaaaaaaaaaaggccataaattattaatttaggcagagcacataaaggctttatttcctattccgtttctctgtatgctttcttcaccaggaagaaatagttttagtgtcaggaatgaat
gagtctgcccctcaattccagcctgctcaacacacagggagacaaactcctgacaatccgagtgactccctggtgactgagctccagccctggatgcatatttgtttagcagttctgacaggacttgacccagccctctctttgcatatgccgtcagaaccttcttt
```

ttttttgctttgagactaagtcttgctctgtcataagcgattcccatgcctcaactcccaaatacctggaattataggcataagtcaccacgcctggctaatttttgtatttttcatagagacggggtttttcgtgttggtcaaactggtctcatactcctgacctcacg

tgagccacctgccttggcctcccaaagagctgggatgacaggcgtgagccaccatactaggtccagaaacttccttttataaaaatgtcattaagtatcttggccacacaatatcattaccagcttcctttaaatccacctctggcctgacaggaatcagggttctt

cagaacctgacattttaaaggaagagaccaggcaggccatgagtaaaggctcatcgtccccatggctctgtcacagctgcactgccgagacatcacagacacggacactggggcctgcttgtttctcaaactgcccttagagccaaagagggaggaactagg

atgaaggccactcattttcccgagaaaagtcctctcctgagggcctggcacagggctctgtccgttgcctgtggccgccgattgctcctctgggttacggagaaaggacggggtcctgagaggcctcagagacttcacacagccctgaaaacatcgggctccttc

aaaagtgcgtcccatcaccaacaaggagactatgtggaggccttgcacccccactccgtgccgctccgccaaatcccagcaacagtgagctcccttcacgctgacgtctgtagaaagcagagaaagcctggctcccagagccctgaagtcctgtggagctgac

cttccctgagtgatggtgtgagcctaaggaattcaactgtgggccaaaggccacctcctggcccgggcctaggtggactctgaggggacacatgtagtcacaatcccaccctccattctccttctcagaggaaggaagcgggcacccatctgcctcatctctgtc

ccaggggggaagatggggagtttcaggggagctttcacatgaatttcactagctcggatcacctgggaggacggggcccaccatgctcctggtgcctccagaagccctgagcccctccagggtccttgagtttgagccagccttgtatcagccccaggagctgaa

tgttccagcaatgggtagaattacatggaaagacctctcaatttggcctgagactgcccacaaacactcaggaaaaacagggcatctcacagagtgggcagcaggtgagtggcaggttacagaccctgagtttgagtttgttcccctgtgagacaggtgcagt

ccttcattctactcatacactggattttaagtggcgcaagataagaggaatttttggtcccaagagcaggaggaagggatttttctggggtttcccgagtccaaatttgcataagatctcctgtgtgtgcattgttctttgaggaccattctctgactcaccaaagt

ggttgaattctaatccgtgtaatgagcgttgcagccaataccagttctgaactctacccggtgaccacgggccaggacctttataaggtggaaggctcggtgtcttccccagactcagctcctggtgaagctcccagccatcagccatgaaggtcctgtatctcct

cttctcgttcctcttcatattcctgatgcctcttccaggtgaggtgggccagggaaatgggaggttggaaaaaatggaagaatggggtacaggctccctgtctcctctcatatccctccacctatctctccctcatacctctctttcctgcctctctctctctctgtcc

cctctatccttctttcctgattctctctctccttccctctctctctttcttctacccacgtttttagactgagtagactgaatgccctacttagttgaaccaagcattgcttccttcaatagaaaaggagcttgagaacccgatggacacctcactctctcttctaagccaa

tatgaaggagcccagtagcttgtaaatctcatctcctcactgcttttcacgttacaactgctaagactgtggttgaaacctgttaggtgactttttaaataaaagtcagaaatttaattttatctaaagaaagtagtacagaatgtcattttctaaatttttatattga

aagggtagatactgcaacctagagagttccaaataattttaaggtgcagcctatactgtaggagctactgctgcagacactctgccccaggacttttccgatcggagaccctgagaacagtccctgccacgaggccaccgcaggttcacaatctagaggcggc

ccatggaaagcaacttttaaccgggacgcctaagcttcattttctccttcatattatggaaataaaataaaaaccatgaaaggataaaagagggagagtggaagggaaggatggagagagggaaaaagaaaatttgagactaaaacccaaaacaattaat

cacatagatattatattgtgaaagcgtcatgttaccaattttatttatgagtcctgggtttttgagaagaacggggctctgagtggctccagagacctcaaattttccaacacccaggacagtatcatgaaggaaggcagggaggcagggaggcagggaggga

ggcagagagggagggaggtagggagggagagaggcagggagggaggtagggaggcagagagggacggaggcagggagggagagaggcaggatggcaggcagggagcaaaggagggagggaggcagggagggagggacagaggcagagaggc

agggagggagggagagaggcagggagggataaaaaaagaagaatggggtgaaaccaggacttagatattagaaacaagcaattacaaaagtttatttctatggttaattgtggttttcaactgtaagttacatggtgttaatttcacattgaacaatttctgt

gagttgtatctttttatcccatctcagatcaaatacttaacagactaaatgatttgaaaagcaaaaagttactggcttgtgtgttaaaacggaggtagggtggctttgatcttatctacttgtggtggagctgaagtcacaagagaacattaccgagctcctaccag

accccactgggagacgccagtcactcaggagagatcagggtctttcacaatcagattcgacaaaaactaacatccccaaaaccacagcagtgccagtttccatgtcagaaacttccatccaaatgattgactcgcgtctcatggaaaagccctggcttcagaa

agaagtccactgcagatctattcaaggcaatacagacacagcgtttgtgtttttgcaacatgagttttgagttctaacacgctgtttgctctttttgtgtttttcctgttaggtgtttttggtgatataaggaatcctgttacctgccttaggagtggtgccatatgtcat

ccaggcttttgccctagaaggtataaacatatcggcatctgtggtgtctctgcaataaaatgctgcaaaaagccatgaggaggccaagaagctgctgtaactgatgtggattcagaaagggctccctcgtcagagatgtgcgacatgtaaacaaaattaaact

gtggtgttcagatttacagaatcttgatcctagtcattgtggtcatggtgtggtgctggtttgggcaggccaatctctaacatccttggaacacccttttcttctctccaggcatgggtcaaggatgccacagtggggcttggagtgctttccagggtcacaggcatc

tgtattatttggatcccttgaccttcccagttattcccgacatttttgtagaacgtgtgctttgctcctcctacaccctcccccttgtatgcctaccccacatcttccctaaaaaaagcaagccctactcgaagaccactttcctcatggaatcatagcggatctgccaag

ggaggggatgcccagtcctctgttcttcacaaggactcccttcttctggttaaggtttcctatgcaattacgcctcctacagaggtgcatgaatttttaattctccatttagctatgagagttctactggtgtggactttgtcttattcatttatgtgctgctcattcataa

gctatttcataacttggatggcaaaatgcagttgtacaagggtgtccttacatacaaactcaataggatccaagtaaatgttcttaaaaacaagtcttttggagggcacaattaagtgaggaccatacagcatgggacacaagcagcggggagtaaacctcaa

gaagaaagatgcatcaactttaaggggaagcatcaggatggctccctggccccgctctctctccttgggagggtttgttccttcaatcatgaactctgtgggtgtatcctacatgcagaagaagctctctatttcttcacgttaactctgcactaaatgtgcatatca

tcaacctcagggaaaatgcacctttcttccaaaataatccattgtttttcctaatctcaagattgaaaaaagtaaccagaccttgttagagtaaaaacatttttatctggatgcagataatatccagatgcgatgacagagccaggactagggtgtattgaggga

ggctctggcctcagatataaaatgcaagcattaatatattccatacttaagatagatgatcttataatgtgatagatttttaaaaaaataacattcatgaaaaatcaccataatcaagaaaacgtccaaaatttaaatcaaggcggggtctaatcctgtaattgtc

caactcagtcacactcacctaccttgatgccaggacagtcagaccctgcctttatatagtcatttaatattcaccatgcatatattattattttgaataaattagtggatttcctggttcctggaagcatatatcagttgattgtataaaagaagtgaaacttcacag

gaaaagaaaaattaaaatgtcaccatgacaatggcattcattcctatgttctaggggaagcagcaaaactctcttaggatgaaactctcctttcatttgagaagacccttttaatggccatttgggtcttgaaagataggacagggaggaagactcccatggaga

ccagcatcctgtggcattgagaagaccgaggtcaaagctcacacaacacaatactccttgtgttagtctgtgtgtgttgctctaaaggaatacttgacgtggaataatacagcaggaaaggaagttattttggctcactattctgcaggctgtgtgagaagcatg

atacctgcatctggccctcattaggcctcaataagccttcagtcctagtggaaagtgaagtgggagcagtaatatcacctggtgagagcagagcgagagggtgaaaggaggtcccagattcttttaaacaactacatgttgtatgaagtcggagcaagaactc

actcattattgggacgaaagcacttagtcattcatcagggattcaccccacaactcaaacactccccaccagattccacttccaacattgaggattagatttcagcatgaggtttcgaggggacaaacatctgaaaaatattattcttcccctggaccctcaaatc

tcatgtccttcttacattgcaaaatacaatcatcccttcccaatcactcccaaaagtcttaattccttccagcattaactgaatcaaaagtccaaagtccaaagtctcatcccctgagactcaagttccttccaccttttagcctgtactatcaaaaataagttatttc

ctttcaagttacaatgatggttcaggcattgactagacaccaccattccaaaaggcagaaactggacaataaaagggattacaggccccacacatgtcagaaactcagcagggaagccgttaaaccccaaagctccaaaacaattcttgattccatgttccgt

attctggtgtgagggtgggctctgaagaccttgggcagctctgcccctgtggctttccagggtgcagcccacatggctcctgtcacaggtcagaatctgatgcctgtggctcttctatgttgagggtacaagctgtcaatggttctcctattctcaagtctgtacagc

agtattccccttccctcagctccactaggtaatgccccactggggacactgtgtggggaatccaaccccacttttcccttagcactgcgctagtagaatttttctgttggtattctccctctgcagcagttttctgcctggagaaccaggcttggccacacatcctctg

aaatgtaggtggaagctgccaagcctctttccttttttgcactctgcatacttgcaggcttaacaccacatgaaagctgccaagacgttgtggttttcacccactgaagcagtgaccaatctgtacattaggccctttgagctgaggctggagcctgggcagccagg

atgtgggaacagtgtcttgaggctgtgcagggcagctgtgccctgggtgtggtcactgaaacattctttcctcttaggcctgtgggctgtagtaggagggaatgcctcaaagttttcaaaaatgccttgtaagccttttccccattgtcttggccattagcacatggc

tccttctcagtcatggaatctctctagcaagtgattctccacagcccacttacattcctcttctcaacatgctttttctttctctactaaattcctaggctgcaaattttccaaacttttatactttgcttcccctttaaatataagttccatatttaaatcatttctttgctcc

tgtatctcactgtaagcttgcagaagcagccacatatcttgactattttgctccttggaaattacttagatcagatcctctaagtaatcactctaaagttcaactttgctgaaatccctaggacatggacacaatgcaggcaagttctttactggaatttaacactgg

tgacccttgttccaattcccaataacttccacttttccacctcacaccttgatagcctggacttcactgtccatatctctatgagcattttggtcacatacatttaacaggtttctaaagagttccaactttccctcatcttcctatcttttgctgggccctgcaaactcttt

cagcctctgccagtacccagtgacaaagccacttccatattttcagctgtctttacagcaatgccccttgctcagtaccaattttctgtgtcagtttgtttttactgctataaaggaatacctaatgctgggtagttcataaggaaaagaagtgtattttggttccctat

tctgtaggctgtatgagtagcatggtgccagcatctgctccttttgaggcttcagaaagcttccatctgggtaggaaagggaagggggaacaggattatcacaagtcaggagaaggagccagagttggggaggtgccaca

APPENDIX 3.2 *DEFB2L* annotated coding sequences

Coding sequence name	Sequence
>DEFB2L_clone_47B11	ATGAAGGTCCTGTATCTCCTGTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCGGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAAGGTATAAACATATCGGCGTCTGTGGTGTCCCTCTAATAAAATGCTGCAAAAAGCCATG
>DEFB2L_clone_135L4	ATGAAGGTCCTGTATCTCCTGTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCGGGTGTTTTTGGTGATACAAGGAATCTTTTTACCTGCATTTGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAACGTATAAACATATTGGCATCTGTGGTGTCCCTCTAATAAAATGCTGCAAAAAGCCATG
>DEFB2L_clone_148L4	ATGAAGGTCCTGTATCTCCTGTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCGGGTGTTTTTGGTGATACAAGGAATCTTTTTACCTGCATTTGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAACGTATAAACATATTGGCATCTGTGGTGTCCCTCTAATAAAATGCTGCAAAAAGCCATG
>DEFB2L_clone_201P10	ATGAAGGTCCTGTATCTCCTGTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCGGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAAGGTATAAACATATCGGCGTCTGTGGTGTCCCTCTAATAAAATGCTGCAAAAAGCCATG
>DEFB2L_clone_217D13	ATGAAGGTCCTGTATCTCCTGTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCGGGTGTTTTTGGTGATACAAGGAATCTTTTTACCTGCATTTGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAACGTATAAACATATTGGCATCTGTGGTGTCCCTCTAATAAAATGCTGCAAAAAGCCATG
>DEFB2L_clone_246K23	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTCATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAAGGTATAAACATATCGGCGTCTGTGGTGTCCCTCTAATAAAATGCTGCAAAAAGCCATG
>DEFB2L_BD2_L1	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCATCTGTGGTGTCTCTGCAATAAAATGCTGCAAAAAGCCATGA
>DEFB2L_BD2_L2	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCATTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAAGGTATAAACATATTGGCGTCTGTGGTGTCCCTCTAATAAAATGCTGCAAAAAGCCATGA
>DEFB2L_BD2_L3	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTCTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCGTCTGTGGTGTCTCTGCAATAAAATGCTGCAAAAAGCCATAA
>DEFB2L_BD2_L4	ATGAAGGTCCTGTATCTCCTGTTCTCGTTCCTCTTGATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCGTCTGTGGTGTCTCTGCAATAAAATGCTGCAAAAAGCCATGA
>DEFB2L_BD2_L5	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCGTCTGTGGTGTCTCTGCAATAAAATGCTGCAAAAACCCATAA
>DEFB2L_BD2_L6	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCGTCTGTGGTGTCTCTGCAATAAAATGCTGCAAAAAGCCATGA
>DEFB2L_BD2_L9	ATGAAGGTCCTGTATCTCCTGTTCTCGTTCCTCTTGATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCATCTGTGGTGTCTCTGCAATAAAATGCTGCAAAAAGCCATGA

APPENDIX 3.3. DEFB2L annotated protein variants

Protein name	Sequence
>macaque reference_NP_001122323	MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPGRYKHIGVCGVPLIKCCKKP
>BACs_variant	MKVLYLLFSFLFIFLMPLPGVFGDIRNLFTCLWSGAICHPGFCPGTYKHIGICGVPLIKCCKKP
>EU090139.1_L1	MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPRRYKHIGICGVSAIKCCKKP
>EU090140.1_L2	MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCIRSGAICHPGFCPGRYKHIGVCGVPLIKCCKKP
>EU090141.1_L3	MKVLYLLFSFLFIFLMPLPGVSGDIRNPVTCLRSGAICHPGFCPRRYKHIGVCGVSAIKCCKKP
>EU090142.1_L4	MKVLYLLFSFLLIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPRRYKHIGVCGVSAIKCCKKP
>EU090143.1_L5	MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPRRYKHIGVCGVSAIKCCKNP
>L6	MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPRRYKHIGVCGVSAIKCCKKP
>L7	MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPGRYKHIGVCGVPLIKCCKKP
>L8	MKVLYLLFSFLLIFLMPLPGVFGDIGNPVTCLRGGAICHPGFCPRRYKHIGVCGVSAIKCCKKP
>L9	MKVLYLLFSFLLIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPRRYKHIGICGVSAIKCCKKP
>65I2_DEFB2L_isoform (L1)	MKVLYLLFSFLFIGVFGDIRNPVTCLRSGAICHPGFCPGRYKHIGVCGVPLIKCCKKP

APPENDIX 3.4. Coding sequences annotated for *DEFB2L* orthologous

Coding sequence name	Sequence
>*Macaca mulatta* b-defensin2-like (DEFB2L), mRNA	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTCTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCGTCTGTGGTGTCTCTGCAATAAAATGCTGCAAAAAGCCATAA
>chimpanzee_DEFB4A ENSPTRG00000019965:ENSPTRT00000037007 cds:KNOWN_protein_coding	ATGAGGGTCTTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGGTATAAGCGATCCTGTTACCTGCCTTAAGAGTGGAGCCATATGTCATCCAGTCTTTTGCCCTAGAAGATATAAACAAATTGGCACCTGTGGTCTCCCTGGAACAAAATGCTGCAAAAAGCCATGA
>Gibbon ENSNLEG00000015127:ENSNLET00000019288 cds:NOVEL_protein_coding	ATGAGGGTCTTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAAGAGTGGTGCCATATGTCATCCAGTCTTTTGCCCTAGAAGGTATAAACAAATCGGCACCTGTGGTCTCCCTGGAACAAAATGCTGCCGAAAGCCATGA
>Gorilla ENSGGOG00000027908:ENSGGOT00000026853 cds:NOVEL_protein_coding	TTGCTCTTTTTGTGTGTTTTTTCCCTGTTAGGTGTTTTTGGTGGTATAAGCGATCCTGTTACCTGCCTTAAGAGTGGAGCCATATGTCATCCAGTCTTTTGCCCTAGAAGGTATAAACAAATTGGCACCTGTGGTCTCCCTGGAACAAAATGCTGCAAAAAGCCA
>Horse_DEFB3 ENSECAG00000007755:ENSECAT00000007775 cds:KNOWN_protein_coding	ATGAAGATCCTTCATTTTCTCCTTGTGTTCCTCGTTGTCTTCCTGTTGCCTGTTCCAGGTTTTACTGCAGGCATAGGAAATTCTGTCACTTGCTCTAAGAATGGAGGCTTCTGCATATCCCCTAAGTGCCTTCCAGGGTCAAAACAGATCGGCACCTGTTCCTTGCCTGGGTCAAAATGCTGCAAAAAGAAGTAA
>Horse_DEFB2 ENSECAG00000008136:ENSECAT00000008191 cds:KNOWN_protein_coding	ATGAAGATCCTTCATTTTCTCCTTGCGTTCCTCGTTGTCTTCCTGTTGCCTGTTCCAGGTTTTACTGCAGGCATAGGAAATCCTATTAGTTGCGCCAGGAATCGAGGTGTATGCATTCCCATTGGGTGCCTTCCAGGGATGAAACAGATTGGCACCTGTGGCTTGCCTGGGACAAAATGCTGCAGAAAGAAGTAA
>Horse_DEFB1 ENSECAG00000008347:ENSECAT00000008443 cds:KNOWN_protein_coding	ATGAGGATCCTTCATTTTCTCCTTGCCTTCCTCATTGTCTTCCTGTTGCCTGTTCCAGGTTTTACTGCAGGCATAGAAACTTCGTTCAGTTGCTCTCAGAATGGAGGCTTCTGCATATCCCCTAAGTGCCTTCCAGGGTCAAAACAGATCGGCACTTGTATCTTGCCTGGGTCAAAATGCTGCAGAAAGAAGTAA
>Human_DEFB4A ENSG00000171711:ENST00000302247 cds:KNOWN_protein_coding	ATGAGGGTCTTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGGTATAGGCGATCCTGTTACCTGCCTTAAGAGTGGAGCCATATGTCATCCAGTCTTTTGCCCTAGAAGGTATAAACAAATTGGCACCTGTGGTCTCCCTGGAACAAAATGCTGCAAAAAGCCATGA
>Human_DEFB4B ENSG00000177257:ENST00000318157 cds:KNOWN_protein_coding	ATGAGGGTCTTGTATCTCCTCTTCTCGTTCCTCTTCATATTTCTGATGCCTCTTCCAGGTGTTTTTGGTGGTATAGGCGATCCTGTTACCTGCCTTAAGAGTGGAGCCATATGTCATCCAGTCTTTTGCCCTAGAAGGTATAAACAAATTGGCACCTGTGGTCTCCCTGGAACAAAATGCTGCAAAAAGCCATGA
>Human_DEFB103 lcl|NM_018661.3_cdsid_NP_061131.1 [gene=DEFB103B] [protein=beta-defensin 103 precursor] [protein_id=NP_061131.1]	ATGAGGATCCATTATCTTCTGTTTGCTTTGCTCTTCCTGTTTTTGGTGCCTGTTCCAGGTCATGGAGGAATCATAAACACATTACAGAAATATTATTGCAGAGTCAGAGGCGGCCGGTGTGCTGTGCTCAGCTGCCTTCCAAAGGAGGAACAGATCGGCAAGTGCTCGACGCGTGGCCGAAAATGCTGCCGAAGAAAGAAATAA
>Marmoset ENSCJAG00000003893:ENSCJAT00000007470 cds:KNOWN_protein_coding	ATGAGGGTCCTGTACCTTCTCCTCTCGTTCCTCTTCGTATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAATAATCCCGTTAGCTGCATAAGGAGTGGTGCCATATGTCATCCGGTGTTTTGCCCTAAAAAATTTAAACAAGTCGGCACCTGTGGTCTCCCTGGAACAAAGTGCTGTAAGAAGAAATGA
>Mouse_defb8 ENSMUSG00000031471:ENSMUST00000033854 cds:NOVEL_protein_coding	ATGAGGATCCATTACCTTCTCTTCACATTTCTCCTGGTGCTGCTGTCTCCACTTGCAGCTTTTAGCCAAAAAATCAATGATCCAGTAACTTACATTCGAAACGGAGGCATATGCCAGTATCGGTGCATTGGCCTTAGGCATAAGATTGGAACTTGTGGATCTCCTTTCAAATGCTGCAAGTGA
>Mouse_defb7 ENSMUSG00000037790:ENSMUST00000047851 cds:KNOWN_protein_coding	ATGAGGATCCATTATGTTCTGTTCGCATTTCTCCTGGTGTTGCTGTCTCCATTTGCAGCTTTTAGCCAAGACATCAACAGTAAACGAGCTTGCTATCGGGAAGGAGGCGAATGCCTGCAACGGTGCATTGGCCTTTTTCATAAGATTGGAACTTGTAATTTTCGTTTCAAATGCTGCAAGTTTCAAATCCCAGAGAAGAAGACAAAGATCCTGTGA
>Mouse_defb3 ENSMUSG00000039775:ENSMUST00000033852 cds:KNOWN_protein_coding	ATGAGGATCCATTACCTTCTGTTTGCATTTCTCCTGGTGCTGCTGTCTCCACCTGCAGCTTTTAGCAAAAAAATCAACAATCCAGTAAGTTGTTTGAGGAAAGGAGGCAGATGCTGGAATCGGTGCATTGGCAACACTCGTCAGATTGGCAGTTGTGGAGTTCCTTTCCTCAAATGCTGCAAGAGAAAATAG
>Mouse_defb5 ENSMUSG00000039785:ENSMUST00000110763 cds:NOVEL_protein_coding	ATGAAGATCCATTACCTTCTCTTTGCATTTCTCCTGGTGCTGCTGTCTCCACTTGCAGGTGTCTTTAGCAAAACAATCAACAATCCAGTAAGTTGCTGTATGATTGGAGGCATATGCAGGTATCTGTGCAAGGGCAACATTCTTCAGAATGGCAGTTGTGGAGTTACTAGTCTCAACTGCTGCAAGAGAAAATAG
>Mouse_defb6 ENSMUSG00000050756:ENSMUST00000063112 cds:KNOWN_protein_coding	ATGAAGATCCATTACCTGCTCTTTGCCTTTATCCTGGTGATGCTGTCTCCACTTGCAGCCTTTTCCCAATTAATCAACAGTCCAGTAACATGCATGAGCTATGGAGGCTCATGCCAGCGTTCATGCAATGGAGGTTTTCGACTGGGTGGCCATTGTGGCCATCCTAAAATCAGATGCTGCCGCAGAAAATAG
>Mouse_defb4 ENSMUSG00000059230:ENSMUST00000081017 cds:KNOWN_protein_coding	ATGAGGATCCATTACCTTCTCTTCACATTTCTCCTGGTGCTGCTGTCTCCACTTGCAGCCTTTACCCAAATTATCAACAATCCAATAACATGCATGACCAATGGAGCCATATGCTGGGGTCCGTGCCCTACCGCTTTTCGACAGATTGGCAATTGTGGCCATTTTAAAGTCAGATGCTGTAAGATAAGATAG
>Orangutan ENSPPYG00000018339:ENSPPYT00000021378 cds:KNOWN_protein_coding	ATGAGGGTCCTGTATCTCCTCTTCTCATTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATTAGCAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCACCTGTGGTCTCTCTGTAATAAAATGCTGCAAAAAGCCATGA
>Rat_defb4	ATGAGGATCCATTACCTCCTCTTCTCATTTCTCCTGGTGCTGCTGTCGCCTCTTTCAGCCTTTACTCAAAGTATCAACAATCCAATCA

ENSRNOG00000013939:ENSRNOT00000058029 cds:KNOWN_protein_coding	CATGCCTGACCAAAGGAGGCGTATGCTGGGGTCCATGCACTGGCGGTTTTCGACAGATTGGCACTTGTGGACTGCCTAGAGTGAGATGCTGCAAGAAAAAGTAG
>Rat_defb3 ENSRNOG00000038126:ENSRNOT00000058030 cds:KNOWN_protein_coding	ATGAGGATCCATTACCTTCTCTTCTCATTTCTCCTGGTGCTGCTGTCGCCCCTTTCTGCCTTTAGCAAAAAGGTCTACAATGCAGTATCGTGTATGACCAATGGAGGAATATGCTGGCTTAAGTGCTCTGGCACTTTTCGAGAGATTGGCAGTTGTGGCACTCGTCAGCTCAAATGCTGCAAGAAAAAGTAG
>EU090141.1_protein3	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTCTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCGTCTGTGGTGTCTCTGCAATAAAATGCTGCAAAAAGCCATAA
>EU090142.1_protein4	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTGATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCGTCTGTGGTGTCTCTGCAATAAAATGCTGCAAAAAGCCATAA
>EU090143.1_protein5	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCGTCTGTGGTGTCTCTGCAATAAAATGCTGCAAAAACCCATAA
>EU090139.1_protein1	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCCTTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTAGAAGGTATAAACATATCGGCATCTGTGGTGTCTCTGCAATAAAATGCTGCAAAAAGCCATGA
>EU090140.1_protein2	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTGATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCCTGTTACCTGCATTAGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAAGGTATAAACATATTGGCGTCTGTGGTGTCCCTCTAATAAAATGCTGCAAAAAGCCATGA
>BACs_variant	ATGAAGGTCCTGTATCTCCTCTTCTCGTTCCTCTTCATATTCCTCATGCCTCTTCCAGGTGTTTTTGGTGATATAAGGAATCTTTTTACCTGCCTTTGGAGTGGTGCCATATGTCATCCAGGCTTTTGCCCTGGAACGTATAAACATATTGGCATCTGTGGTGTCCCTCTAATAAAATGCTGCAAAAAGCCATGAG
>Pig_defb1 lcl|NM_213838.1_cdsid_NP_999003.1 [gene=DEFB1] [protein=beta-defensin 1 precursor] [protein_id=NP_999003.1]	ATGAGACTCCACCGCCTCCTCCTTGTATTCCTCCTCATGGTCCTGTTACCTGTGCCAGGTCTACTAAAAAACATAGGAAATTCTGTTAGCTGCTTAAGGAATAAAGGCGTGTGTATGCCGGGCAAGTGTGCTCCAAAGATGAAACAGATCGGCACCTGTGGCATGCCCCAAGTCAAATGCTGCAAAAGGAAGTAA

APPENDIX 3.5 Protein sequences annotated for *DEFB2L* orthologous

Protein name	**Sequence**
>Rhesus_macaque reference_NP_001122323	MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPGRYKHIGVCGVPLIKCCKKP
>Macaque_BACs_variant	MKVLYLLFSFLFIFLMPLPGVFGDIRNLFTCLWSGAICHPGFCPGTYKHIGICGVPLIKCCKKP
>Macaque_EU090139.1_protein1	MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPRRYKHIGICGVSAIKCCKKP
>Macaque_EU090140.1_protein2	MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCIRSGAICHPGFCPGRYKHIGVCGVPLIKCCKKP
>Macaque_EU090141.1_protein3	MKVLYLLFSFLFIFLMPLPGVSGDIRNPVTCLRSGAICHPGFCPRRYKHIGVCGVSAIKCCKKP
>Macaque_EU090142.1_protein4	MKVLYLLFSFLLIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPRRYKHIGVCGVSAIKCCKKP
>Macaque_EU090143.1_protein5	MKVLYLLFSFLFIFLMPLPGVFGDIRNPVTCLRSGAICHPGFCPRRYKHIGVCGVSAIKCCKNP
>Orangutan gi|297682249|ref|XP_002818838.1| PREDICTED: beta-defensin 2-like [Pongo abelii]	MRVLYLLFSFLFIFLMPLPGVFGDISNPVTCLRSGAICHPGFCPRRYKHIGTCGLSVIKCCKKP
>Gibbon gi|332244525|ref|XP_003271423.1| PREDICTED: beta-defensin 4A-like [*Nomascus leucogenys*]	MRVLYLLFSFLFIFLMPLPGVFGDIRNPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCRKP
>Chimpanzee gi|57114023|ref|NP_001009076.1| beta-defensin 4A precursor [*Pan troglodytes*]	MRVLYLLFSFLFIFLMPLPGVFGGISDPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP
>Marmoset gi|296221579|ref|XP_002756827.1| PREDICTED: beta-defensin 2-like [*Callithrix jacchus*]	MRVLYLLLSFLFVFLMPLPGVFGDINNPVSCIRSGAICHPVFCPKKFKQVGTCGLPGTKCCKKK
>*Homo_sapiens*_DEFB4A gi|4826692|ref|NP_004933.1| beta-defensin 4A precursor [Homo sapiens]	MRVLYLLFSFLFIFLMPLPGVFGGIGDPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP
>*Homo_sapiens*_DEFB4B gi|327532771|ref|NP_001192195.1| beta-defensin 4B precursor [Homo sapiens]	MRVLYLLFSFLFIFLMPLPGVFGGIGDPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP
>*Rattus*_defb4 gi|25742583|ref|NP_071989.1| beta-defensin 4 precursor [*Rattus norvegicus*]	MRIHYLLFSFLLVLLSPLSAFTQSINNPITCLTKGGVCWGPCTGGFRQIGTCGLPRVRCCKKK
>*Rattus*_defb3 gi|82830407|ref|NP_001032637.1| beta-defensin 3 precursor [*Rattus norvegicus*]	MRIHYLLFSFLLVLLSPLSAFSKKVYNAVSCMTNGGICWLKCSGTFREIGSCGTRQLKCCKKK
>horse_defb3	MKILHFLLVFLVVFLLPVPGFTAGIGNSVTCSKNGGFCISPKCLPGSKQIGTCSLPGSKCCKKK

ENSECAG00000007755:ENSECAT00000007775 peptide: ENSECAP00000005738 pep:KNOWN_protein_coding[horse defb3]	
>horse_defb2 ENSECAG00000008136:ENSECAT00000008191 peptide: ENSECAP00000006107 pep:KNOWN_protein_coding [horse defb2]	MKILHFLLAFLVVFLLPVPGFTAGIGNPISCARNRGVCIPIGCLPGMKQIGTCGLPGTKCCRKK
>horse_defb1 ENSECAG00000008347:ENSECAT00000008443 peptide: ENSECAP00000006324 pep:KNOWN_protein_coding [horse defb1]	MRILHFLLAFLIVFLLPVPGFTAGIETSFSCSQNGGFCISPKCLPGSKQIGTCILPGSKCCRKK
>mouse_defb3 ENSMUSG00000039775:ENSMUST00000033852 peptide: ENSMUSP00000033852 pep:KNOWN_protein_coding defb3	MRIHYLLFAFLLVLLSPPAAFSKKINNPVSCLRKGGRCWNRCIGNTRQIGSCGVPFLKCCKRK
>mouse_defb4 ENSMUSG00000059230:ENSMUST00000081017 peptide: ENSMUSP00000079808 pep:KNOWN_protein_coding defb4	MRIHYLLFTFLLVLLSPLAAFTQIINNPITCMTNGAICWGPCPTAFRQIGNCGHFKVRCCKIR
>mouse_defb5 ENSMUSG00000039785:ENSMUST00000110763 peptide: ENSMUSP00000106391 pep:NOVEL_protein_coding defb5	MKIHYLLFAFLLVLLSPLAGVFSKTINNPVSCCMIGGICRYLCKGNILQNGSCGVTSLNCCKRK
>mouse_defb6 ENSMUSG00000050756:ENSMUST00000063112 peptide: ENSMUSP00000060836 pep:KNOWN_protein_coding defb6	MKIHYLLFAFILVMLSPLAAFSQLINSPVTCMSYGGSCQRSCNGGFRLGGHCGHPKIRCCRRK
>mouse_defb8 ENSMUSG00000037790:ENSMUST00000047851 peptide: ENSMUSP00000045523 pep:KNOWN_protein_coding defb7	MRIHYVLFAFLLVLLSPFAAFSQDINSKRACYREGGECLQRCIGLFHKIGTCNFRFKCCKFQIPEKKTKIL
>Gorilla ENSGGOG00000027908:ENSGGOT00000026853 peptide: ENSGGOP00000025415 pep:NOVEL_protein_coding	LLFLCVFSLLGVFGGISDPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP
>Pig_defb1	MRLHRLLLVFLLMVLLPVPGLLKNIGNSVSCLRNKGVCMPGKCAPKMKQIGTCGMPQVKCCKRK
Outlier: human DEFB103 (paralogous)>*Homo_sapiens*_DEFB103 gi|8923890|ref|NP_061131.1| beta-defensin 103 precursor [*Homo sapiens*]	MRIHYLLFALLFLFLVPVPGHGGIINTLQKYYCRVRGGRCAVLSCLPKEEQIGKCSTRGRKCCRRKK