



Investigating the functional properties of the multi-allelic copy number variable *DEFA1A3* locus and their effects on gene expression and disease

By Reem Alhamidi, BSc. MSc

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

August 2022

Abstract

Background

Human α -defensins have important roles in innate immunity and exhibit antimicrobial properties against pathogens. The copy variable locus of human α -defensin genes - denoted as *DEFA1A3* - is a multi-allelic copy number variable (CNV) region located on chr8p23.1. The *DEFA1A3* CNV genes cluster in a single tandem repeat array, consisting of a 19 kb full repeat (either *DEFA1* or *DEFA3*) in tandem and a 10 kb partial repeat (either *DEFA1* or *DEFA3*) at the centromeric end of the array. Individuals exhibit 3 to 16 *DEFA1A3* diploid gene copies in a European (EUR) population. Associations of nearby SNPs with IgA Nephropathy (IgAN) and Periodontitis (PD) have been demonstrated by GWAS (genome wide association studies). It was reported that low *DEFA1A3* copy number increases the risk of IgAN but no such conclusion was made for periodontitis. It is expected that gene copy number should correlate with levels of transcripts. However, previous studies of *DEFA1A3* copy number association to RNA levels have been inconclusive, with studies conducted so far having small sample sizes. Long-read sequencing has made it possible to sequence regions uninterruptedly but inferring true haplotype structures is challenging especially with repeated structures.

Aims

In this work, three study approaches were established to study the functional properties of the multi-allelic variants and their functional effects on diseases in the *DEFA1A3* CNV locus. These study approaches include 1) examining the association of total *DEFA1A3* copy number with Aggressive Periodontitis (AgP) in a European-based population, 2) replicating the association study of total *DEFA1A3* copy number with RNA levels and *DEFA1:DEFA3* genomic ratio to transcript ratio with a larger sampler size, and 3) evaluating Oxford Nanopore

Technologies (ONT) sequencing platform to reconstruct the diploid *DEFA1A3* CNV locus using targeted sequencing approaches from samples such as NA12878.

Results

In the case-control association analysis from 700 European controls and 454 (AgP) cases, we found evidence of an association of AgP and local (telomeric) SNPs but not with *DEFA1A3* copy number. In a study of genomic and RNAseq data from a UK-based population of 197 samples, no evidence of association was detected between total *DEFA1A3* copy number and transcript levels in three human immune cells. By contrast, there was evidence of clear correlation between relative ratios of *DEFA1:DEFA3* RNA levels and equivalent gene numbers. In the nanopore sequencing approach, CRISPR-cas9 enrichment and long amplicon sequencing (tiling path) were used. We were able to successfully reconstruct the *DEFA1A3* haplotypic structures for some samples such as NA12878, NA12760, and NA12248 but not others. The CRISPR-cas9 method appears to be the most suitable approach in sequencing repeated complex structures as it requires reduced repeat sorting and less repeat reconstruction.

Acknowledgments

I would like to dedicate this page to the people who stood by me and supported me in this journey. First and foremost – I would like to thank my supervisor John Armour for his incredible guidance, support, and patience.

I would like to thank everyone I had the pleasure to work with from the JALA lab group and for their help throughout my PhD journey: Nzar Shwan, Nouf Sultan, Ignacio Rodriguez Pastrana and Assel Arginbekova. From David Brook's lab group: Anjani Kumari, Tasabeeh Sala, and Xiaoming Xing.

I would like to also thank my good friends back home and all over the world for their moral support and always being there for me when times get difficult; Esraa Aboalwafa, Riem Alhaj, and Manaal Abukar.

Lastly, I would also like to dedicate a huge appreciation to my family for their endless support. My dad Sami Alhamidi and my mom Nida Original. I would not have been able to do this without them. My siblings, Huda, Ahmed, Mohammed, Heba, Hadi, Andrei, and Andrew.

Contents

Abstract.....	i
Acknowledgments	iii
Contents	iv
List of Tables	ix
List of Figures.....	xvii
CHAPTER 1	1
1. Introduction.....	1
1.1. Human Genetic Variation.....	1
1.2. Classes and Mechanisms of Genetic Variations	2
1.3. Copy number Variations	2
1.3.1. CNVs.....	2
1.3.2. Types and examples of CNVs with association to disorders and phenotypes.....	4
1.3.3. Approaches to typing and sequencing highly repeated regions.....	5
1.4. Defensins.....	8
1.4.1. Θ -defensins	9
1.4.2. β -defensins	9
1.4.3. α -defensins	12
1.5. Long-Read Sequencing	18

1.6.	Hypothesis and Aims of Research	23
CHAPTER 2.....		25
2.	Materials and Methods.....	25
2.1.	Aggressive Periodontitis Association Study	25
2.1.1.	Study Samples.....	25
2.1.2.	1000-Genomes Project CNV measurement	26
2.1.3.	Quality control and filtering.....	28
2.1.4.	SNP selection criteria.....	29
2.1.5.	Association Analysis (single-SNP locus vs multi-SNP locus)	31
2.2.	Blueprint Study	31
2.2.1.	Study Samples.....	31
2.2.2.	Data Access.....	31
2.2.3.	Data Bioinformatic Analysis.....	33
2.3.	Nanopore Sequencing analysis.....	35
2.3.1.	CRISPR-Cas9 enrichment sequencing	36
2.3.2.	Amplicon tiling-path sequencing	37
CHAPTER 3.....		43
3.	Aggressive Periodontitis Association Study.....	43
3.1.	1000 Genomes Project <i>DEFA1A3</i> copy number measurement	43
3.2.	<i>DEFA1A3</i> copy number distribution.....	47
3.3.	Top selected SNPs.....	48
3.4.	SNP and AgP association study	50

3.5.	Interaction of Top SNPs with <i>DEFA1A3</i> copy number	55
3.6.	SNP and <i>DEFA1A3</i> CNV association study	57
3.7.	<i>DEFA3</i> Quantitative Association Study.....	58
3.8.	Haplotype-based association analysis	59
	Centromeric Haplotype Analysis	59
	Telomeric Haplotype association	62
CHAPTER 4		65
4.	Association of <i>DEFA1A3</i> copy number with gene expression in human immune cells	65
4.1.	Total <i>DEFA1A3</i> copy number measurement	65
	Read depth measurement	65
	Accuracy of <i>DEFA1A3</i> copy number measurement	69
4.2.	RNA-Seq data analysis.....	69
4.3.	<i>DEFA1A3</i> CNV association with gene expression	71
4.4.	Association of Neutrophils and T-cells' levels of transcripts	72
4.5.	Association of relative ratios of DEFA1:DEFA3 RNA levels and equivalent gene numbers.	72
CHAPTER 5		77
5.	Amplicon Nanopore Sequencing	77
5.1.	Flongle Amplicon Runs	77
	First Amplicon Experiment.....	80
	Second Amplicon Experiment	83

5.2. MinION Amplicon Run	91
NA06993	97
NA07056	101
NA07347	103
NA12248	105
NA12249	107
CHAPTER 6.....	114
6. CRISPR-cas9 Sequencing	114
6.1. Identifying the haplotypes of NA12878.....	114
6.2. Reconstructing the allelic structure	117
6.3. Shared Variants in the 3 copy and the 2 copy alleles.....	121
6.4. Comparing the shared variants called by freebayes to the consensus sequence generated by Canu.....	121
6.5. Accuracy of base calling using known Exchange 1 and Class 2 flanking SNPs	123
6.6. Comparing the current SNPs to a previously sequenced NA12878 using EHF-PCR.....	124
6.7. Comparing the CRISPR-cas9 nanopore sequences to the phased SNPs identified in Tyson et al.'s study	124
6.8. Comparing the same phased SNPs from Tyson et al.'s study and the current study to that of the amplicon-sequencing method.....	129
CHAPTER 7.....	131
7. Discussion.....	131

7.1.	Aggressive Periodontitis Association Study	131
7.2.	Blueprint Study	136
7.3.	Nanopore Sequencing	139
7.3.1	Amplicon Sequencing	139
7.3.2	CRISPR-Cas9 Sequencing	144
7.4	Summary and Implications of PhD Thesis.....	146
REFERENCES.....		148

List of Tables

Table 1.1. Beta and alpha defensin genes with their corresponding peptides and their amino acid sequences. A.A: amino acid, BM: Bone Marrow, GIT: gastrointestinal tract. Highlighted in red font are the differences between HNP-1, HNP-2 and HNP-3's first amino acid in the sequence from the alpha defensin category.....	10
Table 2.1. Five European subpopulations included from the 1000 Genomes Project in the case-control study.	26
Table 2.2. Four superpopulations from the 1000 Genomes project. EUR: European, ASN: Asian, AMR: American, and AFR: African population	26
Table 2.3. Bed files corresponding to the relevant chromosomal locations (GRCh37/hg19) of the DEFA1A3 CNV repeats and the flanking regions.	28
Table 2.4. Overview of the study population, QC and filtering. *No copy number information available for German controls in this study. 1kGP: 1000 Genomes Project, Pooled-analysis: includes 197 UK-based samples from the Blueprint Epigenome Study [143]. Dataset obtained from Munz et al. [102].	30
Table 2.5. Access to eight datasets from the Blueprint Epigenome Project was granted which includes one WGS dataset, three RNA-seq datasets, three ChIP-Seq datasets, and one DNA methylation profile dataset.....	32
Table 2.6. The on-target scores below indicate the predicted genome editing performance of gRNA at the target site, the higher the score the better the gRNA performance, while the off-target scores indicate that the higher the score the lower the risk of off-target genome editing.	36

Table 2.7. Forward and reverse primers designed using Primer3. PCR1_old and PCR1_new only differ in their reverse primers. Reverse PCR1_new amplifies a region that extends to include the four centromeric tag SNPs in the centromeric repeat of DEFA1A3.	38
Table 2.8. Samples used for nanopore sequencing. Samples number 1 and 2 were only used in the flongle experiment. Sample 2 was used for the CRISPR-Cas9 experiment.....	41
Table 2.9. Phased SNPs from Tyson et al. were used to assist in the sorting of the repeats and characterizing their location in a haplotype, in addition to identify the DEFA1(G)/DEFA3(T) genes. Variant number 1 is the DEFA1/DEFA3 SNP.....	42
Table 3.1. Top SNPs selected at 7 suggestive risk loci on chromosome 8. SNP: single nucleotide polymorphism, r^2 : coefficient of determination BP: base-pair position (hg19:chr8: 6700000-7100000), minor: minor allele, major: major allele, MAF: minor allele frequency, OR: odds ratio, CI: confidence interval, P: P-value, AgP-GERx1kGP-EUR: case control study of German cases and 1000 Genomes Project European controls, Pooled: Blueprint samples (see Chapter 2 section 2.2) were also included as controls in this study. P-values presented here are not adjusted for multiple testing.....	49
Table 3.2. Case-control association analysis. SNP: single nucleotide polymorphism, BP: base-pair position (chr8), minor: minor allele, major: major allele, MAF: minor allele frequency, ADJ: adjusted, OR: odds ratio, CI: confidence interval, P: p-value, CN: DEFA1A3 copy number, ALL: DEFA1A3 copy number and sex	52
Table 3.3. Case-control association analysis. *: variants from Munz et al., associated as risk loci for periodontitis disease.	54

Table 3.4. Interaction analysis between top SNPs and DEFA1A3 copy number. *: SNPs are not as highly significant as lead SNPs, but were identified as risk loci for periodontitis by Munz et al. A1: effect allele.	56
Table 3.5 Association analysis between genotype and DEFA1A3 copy number. Agp-GERx1kGP-EUR: German cases with AgP vs 1kGP European controls. Pooled: German cases with AgP vs 1kGP European controls plus UK based samples from the blueprint study (see chapter 2, section 2.2).	58
Table 3.6. The four DEFA1A3 neighboring tag SNPs that are in linkage disequilibrium.	60
Table 3.7. Haplotype association test. SNPs used for the test: rs4300027, rs7825750, rs7826487. FA: frequency of affected (cases), FU: frequency of unaffected (controls), CHISQ: chi-square test, DF: degree of freedom. TCG: Class 1, TCA: Exchange 2, CCA: Class 2, TTA: Reference sequence or Exchange 1..	61
Table 3.8. Haplotype association test. SNPs used for the test: rs4300027, rs7825750, rs7826487. FA: frequency of affected (cases), FU: frequency of unaffected (controls), CHISQ: chi-square test, DF: degree of freedom. TCG: Class 1, TCA: Exchange 2, CCA: Class 2, TTA: Reference sequence or Exchange 1..	61
Table 3.9. Haplotype association test. SNPs used for the test: rs74556447, rs7016650, rs6998500, rs59878581, rs57688776. FA: frequency of affected (cases), FU: frequency of unaffected (controls), CHISQ: chi-square test, DF: degree of freedom.	62
Table 3.10. Haplotype association test. FA: frequency of affected (cases), FU: frequency of unaffected (controls), CHISQ: chi-square test, DF: degree of freedom.	63
Table 4.1. Statistical correlation analysis for two methods used to test genomic DEFA1:DEFA3 association to DEFA1:DEFA3 transcripts. AD: allele depth.....	73

Table 5.1. Summary of the flongle runs. PCR1 primers*: we used different primers for each run (see Chapter 2, section 2.3 from Materials and Methods). New PCR1 primers include essential centromeric tag SNPS found on DEFA1A3 haplotypes that are useful to classify five haplotypic structures in a European population [2].	78
Table 5.2. Summary of sequencing statistics for samples NA12878 and NA12760 in the primary and secondary runs.	78
Table 5.3. Number of reads reported per PCR bam file per sample. Fold change between the primary and secondary run is reported for comparison.	79
Table 5.4. DEFA1:DEFA3 ratio was determined by using the information of DEFA1(G) and DEFA3(T) alleles in the nanopore reads.....	82
Table 5.5. Observed ratios for a) DEFA1(G):DEFA3(T) alleles found in PCR1 and PCR5 reads, b) centromeric and non-centromeric alleles from PCR1 reads, and c) telomeric and non-telomeric alleles from PCR4 reads.	85
Table 5.6. Eight phased SNPs from Tyson et al. study that were determined from emulsion haplotype fusion PCR. (*): variants originally called by bcftools as a mixed position, and the more frequent allele was reported here. (^): the variants reported here differ from the variants reported in the Tyson et al. study. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). tel: telomeric repeat, mid: middle repeat, cent: centromeric repeat.	87
Table 5.7. Summary of number of mapped reads per repeat per PCR bam file. Repeats number 4 and 5 represent the centromeric repeats. These repeats do not include PCRs 2 and 3 which span the DEFT1P pseudogene that is only found within the telomeric and middle regions of any given DEFA1A3 haplotype.....	91
Table 5.8. Summary of sequencing statistics for one pooled sample made up of 5 barcoded samples. Samples sequenced in this study are from Northern and Western	

European ancestry (CEPH; Utah residents). Min: minimum, Max: maximum, No.: number, hrs: hours, M: million, Gb: gigabases, bp: base pair.	92
Table 5.9. Summary statistics per barcoded sample. BC: barcoded sample.....	97
Table 5.10. Number of mapped reads per PCR bam file per sample.....	97
Table 5.11. Eight phased SNPs from Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). Each column represents a DEFA1 or DEFA3 repeat, and the first row displays labels given for each repeat sorted by the last variant position chosen for repeat separation. This table is from PCR1 reads only. PCR5 repeat separation was incomplete at this stage due to lack of candidate SNPs, although there are 8 variant positions that may be used for sorting, but there were no mixed positions identified in the reads. NC: non-centromeric repeat, C: centromeric repeat. These positions are found in the overlaps between PCR1 and PCR5 amplicons.....	101
Table 5.12. Eight phased SNPs from Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). Each column represents a DEFA1 or DEFA3 repeat, and the first row are labels given for each repeat sorted by the last variant position chosen for repeat separation. NC: non-centromeric repeat, C: centromeric repeat. MP: mixed position. These positions are found in the overlaps between PCR1 and PCR5 amplicons.	103
Table 5.13. Eight phased SNPs from Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). Each column represents a DEFA1 or DEFA3 repeat, and the first row are labels given for each repeat sorted by the last variant position chosen for repeat separation. NC: non-centromeric repeat, C: centromeric repeat, (*): not called by bcftools. These positions are found in the overlaps between PCR1 and PCR5 amplicons.....	104

Table 5.14. Eight phased SNPs from Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). Each column represents a DEFA1 or DEFA3 repeat, and the first row are labels given for each repeat sorted by the last variant position chosen for repeat separation. NC: non-centromeric repeat, C: centromeric repeat. These positions are found in the overlaps between PCR1 and PCR5 amplicons.....104

Table 5.15. Eight phased SNPs from the Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). Each column represents a DEFA1 or DEFA3 repeat, and the first row are labels given for each repeat sorted by the last variant position chosen for repeat separation. NC: non-centromeric repeat, C: centromeric repeat. These positions are found in the overlaps between PCR1 and PCR5 amplicons.105

Table 5.16. Summary of number of mapped reads per repeat per PCR bam file. Repeats number 5 and 6 represent the centromeric repeats. These repeats do not include PCRs 2 and 3 which span the DEFT1P pseudogene that is only found within the telomeric and middle regions of any given DEFA1A3 haplotype.....107

Table 5.17. Summary of number of mapped reads per PCR bam file. The chromosomal locations for each PCR are demonstrated from the GRCh37/hg19 genome assembly. PCR1 and PCR5 includes the DEFA1(G)/DEFA3(T) gene.107

Table 5.18. Eight phased SNPs from Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). Each column represents a DEFA1 or DEFA3 repeat, and the first row displays labels given for each repeat sorted by the last variant position chosen for repeat separation. This table is from PCR5 reads only. PCR1 repeat separation was incomplete at this stage due to lack of candidate SNPs. POS: variant position. Variants highlighted in red correspond to unique difference compared to variants from the other repeats in a single reported position.....110

Table 5.19. DEFA3 sorted repeats extracted using `samtools+awk` command across the five samples from the MinION experiment. Each column represents a single DEFA3 repeat. SNP position 35383 distinguishes the DEFA1 (35383G) from DEFA3 (35383T). C: centromeric repeat, NC: non-centromeric repeat, *: unknown location within the DEFA1A3 haplotype.....112

Table 5.20. Centromeric DEFA1 sorted repeats that were defined from three samples only. Each column represents a single centromeric DEFA1 repeat. MP: mixed position.....113

Table 6.1. SNPs and indels in each read reported from 2 copy and the 3 copy alleles bam files. (*): variation in called SNPs and indels. The 3-copy allele consists of 13 reads only of ~60kb long. When the sequence is not perfectly accurate, all reads at each of these positions would not agree on a single allelic state. The artificial reference sequence used for mapping, the third repeat of the 3 copy allele and the second repeat of the 2 copy allele both have the 4bp deletion in exon 2, therefore relative to the reference sequence, the indel InsTAGA in the reads at those positions and GATA in other positions are all in the same allelic state.120

Table 6.2. Variation of SNPs in both methods, freebayes and Canu in calling variants. 3 copy contigs and 2 copy contigs are produced by Canu. (SNPs: single nucleotide polymorphisms, BP: base-pair, 3C: 3 copy reads, 2C: 2 copy reads). 3C and 2C shared: are variants produced by freebayes in a VCF format at those specific locations telomerically. 3C and 2C contigs are produced by Canu.....122

Table 6.3. Flanking SNPs as reported by Black et al. compared to the 3 copy and 2 copy variants. (BP: base-pair, 3C: 3 copy reads, 2C: 2 copy reads).....123

Table 6.4. The SNPs listed are from Tyson et al. study [92] (haplotype emulsion fusion PCR was used in their experiment) and are compared to the current variants called from CRISPR-cas9 nanopore sequencing. The first repeat corresponds to the most telomeric end of the haplotype while the third repeat corresponds to the most centromeric end of the haplotype. The variants reported on the table are from the

dbSNP build 135. The top SNP with an (*) is the non-synonymous variant that distinguishes DEFA1 (G) from DEFA3 (T) genes. (SNP: single nucleotide polymorphism; 3C: 3 copy reads, 2C: 2 copy reads). First row of each variant are results from this current study, and the second row is what Tyson et al. found. 126

Table 6.5. The variants reported on the table are from the dbSNP build 135. The top SNP with an (*) is the variant base that distinguishes the DEFA1 gene (G) from DEFA3 (T). The last column represents the predicted outcome ratio for the variants of interest from Tyson et al. [92] and what is reported in the current CRISPR analysis..... 129

Table 7.1. Comparison of two variants identified as suggestive risk loci for periodontitis from the Munz et al study. AgP-Gerx1kGP: 1000 Genomes Project European controls and German cases. Pooled: 1000 Genomes Project European controls and German cases, plus the Blueprint GBR controls. SNP: single nucleotide polymorphism, A2: minor allele, A1: major allele, ADJ: adjusted, OR: odds ratio, CI: confidence interval, P: P-value. 134

List of Figures

Figure 1.1. Top image: Schematic map of the DEFA1A3 copy number variable locus on chromosome 8, based on the GRCh37/hg19 human genome assembly, representing the predicted structural arrangement of a two-copy allele. Four tag SNPs (single nucleotide polymorphisms) that distinguish the different haplotypes are located at the centromeric region of the repeat: (1) rs7825750 (2) rs62487514, (3) rs4300027 and (4) rs7826487. Bottom image: Five DEFA1A3 haplotypic structures commonly observed with European ancestry. Ins5 and Del5 is a 5 bp indel found external of the DEFA1A3 coding region which helps characterize the arrangement of haplotypes. Source of data from Black et al. [2]. Image created with BioRender.com13

Figure 1.2. Top image: Schematic map of the DEFA1A3 copy number variable locus on chromosome 8, based on the GRCh37/hg19 human genome assembly, representing the predicted structural arrangement of a two-copy allele. Four tag SNPs (single nucleotide polymorphisms) that distinguish the different haplotypes are located at the centromeric region of the repeat: (1) rs7825750 (2) rs62487514, (3) rs4300027 and (4) rs7826487. Bottom image: Five DEFA1A3 haplotypic structures commonly observed with European ancestry. Ins5 and Del5 is a 5 bp indel found external of the DEFA1A3 coding region which helps characterize the arrangement of haplotypes. Source of data from Black et al. [2]. Image created in Bio Render.13

Figure 2.1. A flowchart of the steps performed to obtain the DEFA1A3 copy number measurements from WGS, and normalized gene expression counts from RNA-Seq from neutrophils, T-cells, and monocytes. (N: neutrophils, M: monocytes, EGA: European Genome-Phenome Archive).....33

Figure 2.2. Allele sorting using a <code>samtools+awk</code> syntax command by filtering reads based on the status of a single nucleotide.....	34
Figure 2.3. Steps performed on nanopore-sequencing data (CRISPR-Cas9 enrichment method in a MinION, and amplicon sequencing method in a flongle and a MinION).....	36
Figure 2.4. The position of gRNAs relative to the DEFA1A3 repeat region and the neighboring flanking sequences. The gene symbols indicate the order and content only, and positions are not shown to scale.....	37
Figure 2.5. Visual representation of the 3 repeats as viewed on the UCSC Genome Browser from the 2009 human genome assembly(Ch37/hg19) and the corresponding positions of PCR products (top). Steps of amplicon sequencing for NA12878 (BC01) and NA12760 (BC02) prior to loading into a flongle. The chromosomal locations are not accurately to scale but only for illustration purposes.	39
Figure 3.1. DEFA1A3 copy number estimates of DEFA1core plotted against the copy number estimates of DEFA2core+2 from 2,503 samples (1 sample excluded as an outlier), showing the correlation between the two estimates (r^2 : 0.9761) and evidence of clustering around integer values from 3 to 8.....	45
Figure 3.2. DEFA1A3 copy number distribution in four super populations. EUR: European, AMR: American, ASN: Asian, AFR: African.....	46
Figure 3.3. DEFA1A3 copy number distribution in 503 European controls from the 1000 Genomes Project (1kGP-EUR) and 566 German cases (AgP-GER). This figure only reports the DEFA1A3 deduced copy number from the 1kGP-EUR and, out of 744 AgP-GER cases, only 566 cases were included as the rest of the samples (178 AgP-GER cases) are not included in the data available from the case-control study. Deduced copy number in controls were measured via read depth while cases	

were measured experimentally using the PRT (paralogue ratio test) by a previous MRes student.	47
Figure 3.4. Manhattan plot produced on LocusZoom [152]. The figure demonstrates SNPs within the region of chromosome 8 (6.65MB to 7MB) from the GRCh37/hg19 assembly. The top SNPs with the lowest P-values are located at the telomeric (left) flanking region of DEFA1A3 CNV (highlighted in the red box, corresponding to the gap without any SNPs).....	50
Figure 3.5. DEFA1A3 haplotype frequency in cases and controls. RS: Reference seq, C1: Class 1, C2: Class 2, X1: Exchange1, X2: Exchange 2.....	59
Figure 4.1. DEFA1A3 copy number estimates. (A) up to four copy number measurements obtained per individual, (B) average copy number measurement per individual.	67
Figure 4.2. Unrounded DEFA1A3 copy number distribution around each integer value.....	68
Figure 4.3. DEFA1A3 copy number distribution in the UK (United Kingdom) population (197 samples).....	68
Figure 4.4. Distribution of normalized values (based on the mean of the two independent DEFA1A3 measurements) in 197 individuals.....	69
Figure 4.5. Association of total DEFA1A3 copy number with gene expression in Neutrophils and T-cells. Note the very different scales on the y-axis for the two analyses.	71
Figure 4.6. DEFA3:DEFA1 ratio. A) Using bcftools mpileup + AD syntax to obtain the allele depth for DEFA1 and DEFA3 alleles; B) samtools+awk syntax command was used. C) Hybrid approach – bcftools mpileup+AD on x-axis for genomic and samtools+awk on y-axis for transcripts.	75

Figure 4.7. DEFA1, DEFA3, and total DEFA1A3 copy number deduced from bcftools mpileup + AD.	76
Figure 5.1. Histogram of read distribution for NA12878 (top) and NA12760 (bottom).....	81
Figure 5.2. An example of repeat sorting example from the MinION amplicon experiment for sample NA12248 from PCR1 data. Six repeats from PCR1 were separated. The numbers in the figure represent the number of reads per step.....	83
Figure 5.3. Raw read lengths plotted against the number of reads produced from the flongle sequencing run. Figures were created from a NanoPlot [1] python script.	84
Figure 5.4. The repeat structures characterized by nanopore sequencing at the DEFA1A3 locus following the tiling path of the PCRs to their corresponding repeats based on SNP information a) between PCR overlapping regions, b) centromeric/non-centromeric from PCR1, and c) telomeric/non-telomeric from PCR4. NA12878 (top) and NA12760 (bottom). Del5/Ins5: 5bp indel variant (ACAGA).....	89
Figure 5.5. The figures here demonstrate the frequency of the reads outputted per DEFA1A3 repeat post-repeat separation via the samtools+awk command. NA12878 (top) and NA12760 (bottom) have 5 DEFA1A3 repeats each.	90
Figure 5.6. Summary of the 72-hour MinION experiment. Figures were obtained from the Minotour DeepSeq webpage (http://nanoseq.nottingham.ac.uk:8000/login/). A) Mean read length by genome mapped to chromosomes 1-22, X and M (human mitochondrial chromosome). B) Zoomed-in genome coverage focused on chromosome 8 from 5 barcoded samples and the unclassified group. C) Histogram of total number of reads by read lengths and D) total number of bases by read lengths. E) Percentage of total reads sequenced per barcoded sample. The orange and blue bars (found in	

Figures C, D, and E) represent reads that passed and failed the quality threshold respectively. Barcode 1: NA06993, Barcode 2: NA7056, Barcode 3: NA07347, Barcode 4: NA12248, Barcode 5: NA12249.96

Figure 5.7. Possible general haplotype structures for NA06993. A) 6 + 2 copy haplotype, B) 5 + 3 copy, and C) 4 + 4 copy haplotype.98

Figure 5.8. An IGV (version 2.5.2) snapshot of PCR1 (top) and PCR4 repeats (bottom) - in VCF format - from 8 repeats for sample NA06993 (zoomed out). The red rectangle represents a block of SNPs that differ in the centromeric and non-centromeric PCR1 repeats, and in the telomeric and non-telomeric PCR4 repeats. Light blue indicates SNPs different from the reference sequence, while dark blue indicates a mixed position relative to the reference sequence.100

Figure 5.9. The expected Class 2 structure of NA07056 consisting of 5 copies of DEFA1A3 (3 DEFA1 copies + 2 DEFA3 copies). The DEFA3 copies are expected to be positioned in the telomeric region of the haplotype [2]. Del5/Ins5 is a 5bp indel situated at position chr8:40301 in our sequences (ACAGA).....102

Figure 5.10. Possible haplotype structural organization for NA12248 based on other known structures of class 1 and class 2 alleles [2]. A) A 4 by 2 copy structure and a B) 3 by 3 copy structure. Del5/Ins5 is a 5bp indel situated at position chr8:40301 in our sequences (ACAGA). The DEFA3 copy is shown in the telomeric position of the haplotype, as frequently observed in class 2 haplotypes [2].106

Figure 5.11. Full haplotype structure of NA12248 was constructed successfully for the DEFA1A3 repeats. Del5/Ins5 is a 5bp indel situated at position chr8:40301 in our sequences (ACAGA).106

Figure 5.12. Possible general haplotype structures for NA12249. A) 7 + 2 copy and B) 6 + 3 copy and C) 5 + 4 copy haplotype structure.108

Figure 6.1. (-) and (+) strands of sample NA12878. Chromosomal location of reads on chromosome 8 mapped between chr8:6750000-6910000. The red rectangular box represents the location of DEFA1A3 CNV and regions flanking the gene. Black lines represent the full length of the 3 copy allele reads. Total number of reads is approximately 260 reads.	117
Figure 6.2. The reconstructed haplotypic structure of NA12878 from the 3 copy and 2 copy haplotypes using the CRISPR Cas9 sequencing.	118
Figure 6.3. Shared variants between the 3C (red circle) and 2C (blue circle) haplotypes for sample NA12878.	121

CHAPTER 1

1. Introduction

1.1. Human Genetic Variation

Variation in the human genome can have an impact on how genes are expressed, which then may influence physiological, biochemical, morphological, and pathological traits in humans. The genetic variation between different individuals provides the opportunity to explain genetic contributions to phenotypic differences and disorders. In 2001, the Human Genome Project provided a benchmark for the identification of genetic variations with the publication of a near-complete genome sequence [3]. This has led to several large-scale research projects that focused on investigating human genetic variations such as the International HapMap Project and the 1000 Genomes Project. The International HapMap Project genotyped over one million variants in four super-populations [4]. Subsequently, the 1000 Genomes Project derived sequences from 26 subpopulations of 2,504 individuals using high-throughput sequencing technologies to enable whole-genome and whole-exome sequencing [5]. Such projects were important to document genetic diversity within and between populations with frequencies as low as 1%. Additionally, genetic association analyses - such as genome-wide association studies (GWAS) – have led to the discovery of important genetic markers that exhibit strong associations with certain phenotypic traits and disorders [6-10].

1.2. Classes and Mechanisms of Genetic Variations

By studying the variations observed between two individual sequences in any given population, mutational pathways and their consequences are commonly observed. The major types of mutations found in humans are: 1) single nucleotide polymorphisms (SNPs) 2) insertions or deletions (indels), and 3) structural variations (SV). The different forms are largely defined based on the genomic size and mechanism of formation. SVs account for variation in a greater number of nucleotides when comparing any two genomes, even though they occur less frequently than SNVs (single nucleotide variants) and indels [11]. SNVs are single base pair substitutions such as transitions or transversions. However, if the frequency of the substitution in a human population is $>1\%$, it is described as a SNP (single nucleotide polymorphism), and SNPs are the most common form of variation. Indels are the gain or loss of 1-50 bases. Lastly, SVs are categorized based on the length of the sequence, and involve rearrangements such as inversions, translocations, tandem repeats (microsatellites and microsatellites), and copy number variations.

1.3. Copy number Variations

1.3.1. CNVs

Copy number variation (CNV) is one form of human genetic variation and is an important subset of SVs that can lead to phenotypic, genetic, and morphological diversity in addition to various predispositions to disorders [11-13]. However, many genes can be completely deleted with no apparent deleterious consequences caused [14]. CNVs consist of deletions, duplications, or insertions of more than 1000 bp in length observed in variable copies relative to the human reference genome. It has been estimated that 4.8% to 9.5% of the human genome

is implicated in CNVs [15]. There are over 7 million CNVs in the human genome reported by the Database of Genomic Variants (DGV) that range from 50 bp to over >1 Mb, but the most commonly documented CNVs span between 1 Kb and 10 Kb. CNVs cause gains or losses of sequence copy number that can directly impact the levels of gene expression [16, 17].

Most CNV studies have been based on two different approaches: genome wide assessments, and locus-specific studies. In the former approach, CNV discovery is strongly biased towards copy number changes that result in a significant change in gene dosage. The latter approach depends on small sample sets with diploid copy number calling that is error prone. Taken as a whole, reduced purifying selection on copy number has been suggested for *Drosophila melanogaster* and humans in genome wide assessments [18-20]. However, neutral evolution (such as that affecting human chemosensory receptor genes) and positive selection (such as at the amylase locus and *UGT2B17* gene) have also been suggested [21-23]. CNVs can be classified into two groups: Bi-allelic and multi-allelic CNVs. Most of the CNVs that are descended from a single ancestral duplication or deletion event follow a simple bi-allelic model, while multi-allelic CNVs occur when several independent or recurrent deletions or duplications happen in the same locus, which results in a variety of potential copy number consequences [11]. The recurrent rearrangements lead to a dynamic range of copy number that is more than 3 copies per diploid genome [13, 24].

Multi-allelic CNVs are less studied and understood, due to the complexity and difficulty in measuring the number of copies accurately and determining the structural differences between alleles. When new multi-allelic CNV alleles are formed, recurrent mutation means that they may not show strong linkage disequilibrium (LD) with nearby SNPs. Multi-allelic CNVs have higher mutation rates than the simple bi-allelic form [17]. Previously, it has been shown that multi-allelic CNVs only comprise about 15% of overall CNVs studied, but give rise to more than 85% of variations of gene dosage in humans. A statistical evaluation

suggests evidence that multi-allelic CNVs contribute to gene dosage variation seven times more than bi-allelic CNVs, although bi-allelic CNVs are much numerous in number compared to multi-allelic CNVs [17].

There are several molecular mechanisms behind the formation of CNVs: non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ), Fork Stalling and Template Switching (FoSTeS), L1-mediated retrotransposition, and (most commonly) non-allelic homologous recombination (NAHR) [25]. NAHR is a process of recombining very similar sequences that occurs between long segments of DNA (≥ 300 bp long). NAHR can result in recurring changes in copy number and can structure a substantial copy number variable region. This mechanism can take place either between homologous chromosomes, sister chromatids or in an intra-chromatid approach.

1.3.2. Types and examples of CNVs with association to disorders and phenotypes

Human genomes exhibit segmental CNVs at numerous loci. In many cases, varying gene copy numbers do not cause any type of phenotypic changes or disorder [26, 27]. However, several CNVs have been associated with conditions that range from rare to common complex disorders [28]. A dosage-sensitive gene, *PMP22* (peripheral myelin protein 22), has a duplication with an outcome of increased protein production and has been associated with Charcot-Marie-Tooth Type 1A [29]. A further example is the *RAI1* (retinoic acid induced 1) gene, which causes either Smith-Magenis syndrome or Potocki-Lupski syndrome depending on whether the gene is deleted or duplicated [30]. Rare and *de novo* CNVs greater than 100 kb in length with an allele frequency of $<1\%$ are known as major risk factors for many human disorders such as schizophrenia and autism [7, 17, 31]. Rare CNVs, that usually produce dramatic effects on phenotype, are most likely to be under substantial negative selection pressure because they cause deleterious effects on the host [17]. Multi-allelic CNVs have also been reported to have

associations with diseases, such as *C4* (complement component 4), *HP* (haptoglobin), *FCGR3B* (Fc fragment of IgG with low-affinity IIIb), *CCL3L1* (C-C chemokine ligand-3-like-1), β -defensin, and *AMY1* (salivary amylase) genes. However, some of these findings have not been replicated by subsequent work which will be further discussed in this section in detail.

Common deletions and duplications with varying copy numbers in the human genome have a frequency $>5\%$ [32]. For example, the *FCGR3B* CNV region uncovered an association with glomerulonephritis in humans and rats [33]. In addition, an association of systemic lupus erythematosus (SLE) with low *FCGR3B* copies was found [34, 35]. Nonetheless, this gene-dosage explanation was disputed after the observation of natural killer (NK) cells yielding atypical transcription from the *FCGR2B* gene and not *FCGR3B* [36]. High copies of β -defensins have been associated with psoriasis (an immune-mediated skin disease) [37, 38]. In addition, two independent studies revealed that high and low copies of β -defensins predispose to Crohn's disease risk [39, 40]. However, as well as contradicting each other, neither of the findings was replicated in later studies [26, 41]. As a further example, low *CCL3L1* copy number has been correlated as a susceptibility locus for HIV, and high *CCL3L1* copies has been associated with rheumatoid arthritis (RA) [42, 43]. However, because of inaccurate copy number measurements in preliminary studies, both findings have been subsequently disregarded [44-46]. Other copy number variations are regarded to have adaptive advantage despite the change in phenotype [47].

1.3.3. Approaches to typing and sequencing highly repeated regions

As the copy number increases, it becomes challenging to differentiate integer states as the relative difference among CNV classes becomes smaller, and in general it becomes challenging to type multi-allelic CNVs accurately [48]. Replicating association studies can be problematic due to the difficulty with

genotyping sequences of high identity in CNV regions. For example, The Wellcome Trust Case Control Consortium GWA study of CNVs was conducted on eight common diseases that included 2000 cases of rheumatoid arthritis (RA) and Crohn's disease, but there was no evidence supporting previous reports from candidate-gene studies of association with *CCL3L1*, *FCGR3B*, and β -defensin [26]. A further challenge is the ability to obtain copy number values in integer numbers (discrete) instead of a continuous variable and the limitations of accurately measuring copy numbers of more than 4 copies. For example, in 2014, a study generated a detailed copy number map from 14 different human populations of the 1000 Genomes project (Phase 1) of ~850 genomes. The study mainly focused on the ability to obtain discrete copy number clusters on CNV loci. It was discovered that the range of discrete diploid copy numbers was from 0 to 15 [17].

Several molecular approaches have been developed to analyse multi-allelic CNVs, but no one approach is applicable for examining all CNV classes due to their various structural properties. Determining gene copy numbers by molecular approaches can be categorized as follows: i) locus-specific hybridization based methods such as Fluorescent In Situ Hybridization (FISH), Southern blotting and pulsed field gel electrophoresis, ii) array based methods, iii) PCR based methods such as quantitative real-time PCR, multiplex amplifiable probe hybridization (MAPH), multiplex ligation-dependent probe amplification (MLPA), droplet digital PCR (ddPCR), and paralogue ratio test (PRT), and lastly iv) sequencing based methods [49, 50].

Locus specific hybridization-based methods

FISH is a commonly used approach that detects DNA sequences using specific fluorescent probe often covering a substantial extent of genomic sequences. The probes bind to complementary DNA sequences. The limitation of this technique is the inability to detect shorter DNA sequences of the order of 1 kb long due to the limiting fluorescent probe signal. FIBRE-FISH is a modified

version that is a powerful tool in determining complex structural rearrangements, estimating integer copy number correctly and deducing the copy number per haplotype [22, 51]. A limiting factor of FIBRE-FISH is that high quality DNA is required for genotyping multi-allelic CNVs, in addition to the relatively low yield of individual experimental observations made from a laborious and lengthy procedure. Probe cross-hybridization with paralogues in highly repeated regions is also a key limitation [49]. Similarly, Southern blotting of pulsed-field gels requires high quality DNA and is laborious [52].

Array-based methods

Array CGH technique is a high throughput method that has been popular in large-scale studies for copy number estimation [11, 26]. The method compares a test sample and a reference DNA on a microarray in which each sample is marked with a fluorescent label. Array hybridisation methods can also be successful in detecting copy number changes when applied in a non-comparative (i.e, internally controlled) format. As with other DNA-based methods, there is no absolute way to calibrate quantification per cell, but instead relative quantification is usually employed, based on a normalising average value taken to be representative of the autosomal genome as a whole.

PCR-based methods

qPCR is a useful and convenient method to quantify DNA copy number and can detect small deletions and duplications. The main disadvantage of qPCR is that it has limited accuracy, and can distort associations at CNV regions, most likely because of differential effects of sample condition on the test result [41, 49]. Multiple studies have used qPCR to investigate association at the multi-allelic locus such as the salivary amylase gene and pancreatic amylase gene [22, 53]. Nevertheless, a study provided evidence that qPCR is error-prone for measuring copy number accurately, and reproducibility is low [54].

The key difference between MAPH and MLPA is in the type of specific probes; MAPH uses double stranded, amplifiable probes, while MLPA uses single stranded half probes that need template mediated ligation to become amplifiable. Both techniques are PCR based methods and can detect deletions and duplications in association to certain genetic traits or disorders [52, 55]. MLPA appears to call most integer copy numbers with confidence, but it is difficult to scale conveniently to thousands of samples, and copy number only appears correctly measured for some loci [56]. On the other hand, PRT experiments have been effective in measuring copy numbers accurately, especially for some of the more complex loci [46, 57, 58]. In many ways, PRT is a simpler and cost-effective approach to accurately measure CNVs with accuracy sufficient to distinguish integer states reliably, and the amount of DNA input required is 10-20x less compared to MLPA. Lastly, sequencing-based techniques will be further discussed in detail in chapter 1.5.

1.4. Defensins

Antimicrobial peptides (AMPs), also known as host-defence peptides, are part of the innate immune response. They form the first line of defence against potential pathogens (i.e., gram-positive, gram-negative bacteria, enveloped viruses, and fungi) [59]. Cathelicidins and defensins are the most widespread AMPs in mammals, and are expressed in neutrophils and macrophages [60]. Defensins are β -sheet rich, small cationic peptides that range between 25-45 amino acids. Their molecular weight is between 2-6 kDa, and they contain six cysteine residues that form three intramolecular disulphide bridges [61]. Due to the differences in the size and structure of the disulphide bonding, mammalian defensins are categorized into three different sub-families: alpha (α), beta (β), and theta (θ) defensins. The primary function of defensins is their antimicrobial action against microorganisms by damaging the cytoplasmic membranes [62]. Defensins strike pathogens by binding to the microbial membranes, creating voltage dependent ion channels that enable other enzymes to invade the pathogens and

cause cytotoxicity [63-65]. Defensins are efficient against negatively charged microorganisms because of their cationic and hydrophobic properties [66-71]. Additionally, defensins can suppress pro-inflammatory cytokine induction by bacterial antigens and can trigger an antigen-specific immune response [72]. In addition to LL-37 (a cathelicidin AMP), defensins can stimulate and de-granulate mast cells, can enhance phagocytosis in blood by macrophages, work as chemo-attractants for various cells (i.e., monocytes, macrophages, T-cells, and undeveloped dendritic cells) and can regulate the complement system [73-76]. Overall, defensins combine effector (microbicidal) functions with signalling roles that link innate immunity to the adaptive immune system. The main types of defensins expressed in humans are α -defensins and β -defensins[77]. The genes of the copy variable defensins cluster on chromosome 8 (8p22-p23) at the telomeric region [78].

1.4.1. Θ -defensins

Θ -defensins are reported to have evolved from α -defensins and active peptides have only been identified in leukocytes of old-world monkeys such as rhesus macaques and baboons (*Papio anubis*) [79, 80]. A Θ -defensin pseudogene called *DEFTIP* is the only gene of this family found in humans; the pseudogene does not produce a functional peptide because of a stop codon in the sequence [81].

1.4.2. β -defensins

β -defensins are synthesized in epithelial cells at mucosal surfaces and exhibit broad antimicrobial properties that are stimulated upon pathogen entry [82]. The β -defensin peptides and their respective genes located in the 8p23.1 chromosome cluster are presented in Table 1.1. The β -defensin copy number commonly ranges from 2 to 7 copies [55], but higher than 7 is observed occasionally.

Table 1.1. Beta and alpha defensin genes with their corresponding peptides and their amino acid sequences. A.A: amino acid, BM: Bone Marrow, GIT: gastrointestinal tract. Highlighted in red font are the differences between HNP-1, HNP-2 and HNP-3's first amino acid in the sequence from the alpha defensin category.

Gene	Peptide	A.A sequence	Tissue Distribution
β -defensins			
<i>DEFB1</i>	HBD-1	MRTSYLLLFTLCLLLSEMASGGNFLTGLGHRSDHYNVCVSSGGQCLYSACPIFTKIQTGTYRGKAKCCK	Epithelial cells
<i>DEFB103A</i>	HBD-3	MRIHYLLFALLFLFLVPVPGHGGIINTLQKYYCRMRIHYLLFALLFLFLVPVPGHGGIINTLQKYYC RVRGGRCAVLSCLPKKEEQIGKCSTRGRKCCRRKK	Epithelial cells
<i>DEFB4</i>	HBD-2	MRVLYLLFSFLFIFLMPLPGVFGGIGDPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP	Epithelial cells
<i>DEFB104</i>	HBD-4	MQRLVLLLAISLLLYQDLPRSEFELDRICGYGTARCRKKCRSQEYRIGRCPNTYACCLKWDESL LNRTKP	Testis
<i>DEFB105</i>	BD-5	MALIRKTFYFLFAMFFILVQLPSGCQAGLDFSQPFPSGEFAVCESCKLGRGKCRKECLENEKPDG NCRLNFLCCRQRI	Testis
<i>DEFB106</i>	BD-6	QECIFDEKCEDELKGACKKHCEKNEELTSFCQKSLKCCRTI	Epididymis and lung.

<i>DEFB107</i>	BD-7	MPGAMKIFVFILAALILLAQIFQARTAIHRALIS KRMEGHCEAECLTFEVKIGGCRAELAPFCCKN RKKH	Testis
<i>SPAG11</i>	HE2	MRQRLLPSTSLLLVALLFPSSQARHVNHSA TEALGELRERAPGQGTNGFQLLRHAVKRDLLP PRTPPYQVHISHQEARGPSFKICVGFLGPRWAR GCSTGN	Epididymal epithelium and on the sperm surface
α-defensins			
<i>DEFA1</i>	Human Neutrophil Peptide 1 (HNP1)	A CYCRIPACIAGERRYGTCTIYQGRLWAFCC	BM & lymphoid tissues
	Human Neutrophil Peptide 2 (HNP2)	- CYCRIPACIAGERRYGTCTIYQGRLWAFCC	
<i>DEFA3</i>	Human Neutrophil Peptide 3 (HNP3)	D CYCRIPACIAGERRYGTCTIYQGRLWAFCC	BM & lymphoid tissues
<i>DEFA4</i>	Human Neutrophil Peptide 4 (HNP4)	VCSCRLVFCRRTELRVGNCLIGGVSTYCCTRV	BM & lymphoid tissues
<i>DEFA5</i>	Human Defensin 5 (HD5)	ATCYCRHGRCATRESLSGVCEISGRLYRLCCR	GIT
<i>DEFA6</i>	Human Defensin 6 (HD6)	AFTCHCRRSCYSTEYSYGTCTVMGINHRFCCL	GIT

1.4.3. α -defensins

Peptide Composition

Human α -defensins are small (29-34 amino acid) cationic peptides with fundamental roles in host defence of innate immunity, as the first line of defence when an infection arises [83]. α -defensins were discovered earlier and are closely related to β -defensins. α -defensins descended from β -defensins and evolved much later given that they were detected in mammals and marsupials only [78, 84, 85]. They possess antimicrobial and antiviral properties against gram-positive and gram-negative bacteria and enveloped viruses, respectively [59]. Human α -defensins are also known as Human Neutrophil Peptides (HNP) because they were first isolated from the azurophilic granules of neutrophils [71]. There are six types of α -defensins described in the literature (Table 1.1). HNP-1, HNP-2, HNP-3 and HNP-4 are primarily expressed in the granules of neutrophils, while HD-5 (Human defensin 5) and HD-6 are expressed in the small intestine (Paneth cells) [52, 85-87]. Daher *et al.* found that HNP-1, HNP-2 and HNP-3 make up approximately 30% of the total granular proteins of neutrophils [88]. HNP-1 to HNP-3 are produced by promyelocytes, the bone marrow precursors to neutrophils [89]. HNP-1 and HNP-3 are encoded by the copy variable genes *DEFA1* and *DEFA3* respectively. Both peptides have similar sequences, and they contain 30 amino acids each except they differ in their first amino acid (N-terminal) of the sequence. The first amino acid of HNP-1 is alanine, whereas HNP-3 contains aspartic acid instead. There is no evidence to suggest that HNP-2 is encoded by any specific gene and *DEFA2* does not exist [90]. The HNP-2 peptide lacks the first amino acid in its sequence, and therefore it is believed that HNP-2 is a proteolytic product expressed from either *DEFA1* or *DEFA3* gene. These peptides are made by neutrophils, and they interact with microbial surface membranes to disrupt them which will eventually lead to apoptosis.

Genomic Organization

DEFA1A3 is a multi-allelic copy number variable region in which each haplotype consists of a 19 kb full repeat in tandem (consisting of either *DEFA1* or *DEFA3* and *DEFTIP* pseudogene) and a single copy of a 10 kb partial repeat (including either *DEFA1* or *DEFA3*) at the centromeric end of the array (Figure 1.1). In various human populations, the lack of *DEFA3* gene varies in frequency between 10% and 37% [52, 87, 91]; chimpanzee, orangutan, bonobo and gorilla lack *DEFA3* completely which suggests that *DEFA3* is human-specific and has emerged from *DEFA1* [52]. The difference between the α -defensin genes is that *DEFA4*, *DEFA5* and *DEFA6* are commonly found at 2 copies per diploid genome, while *DEFA1* and *DEFA3* show extensive variation in copy number states within

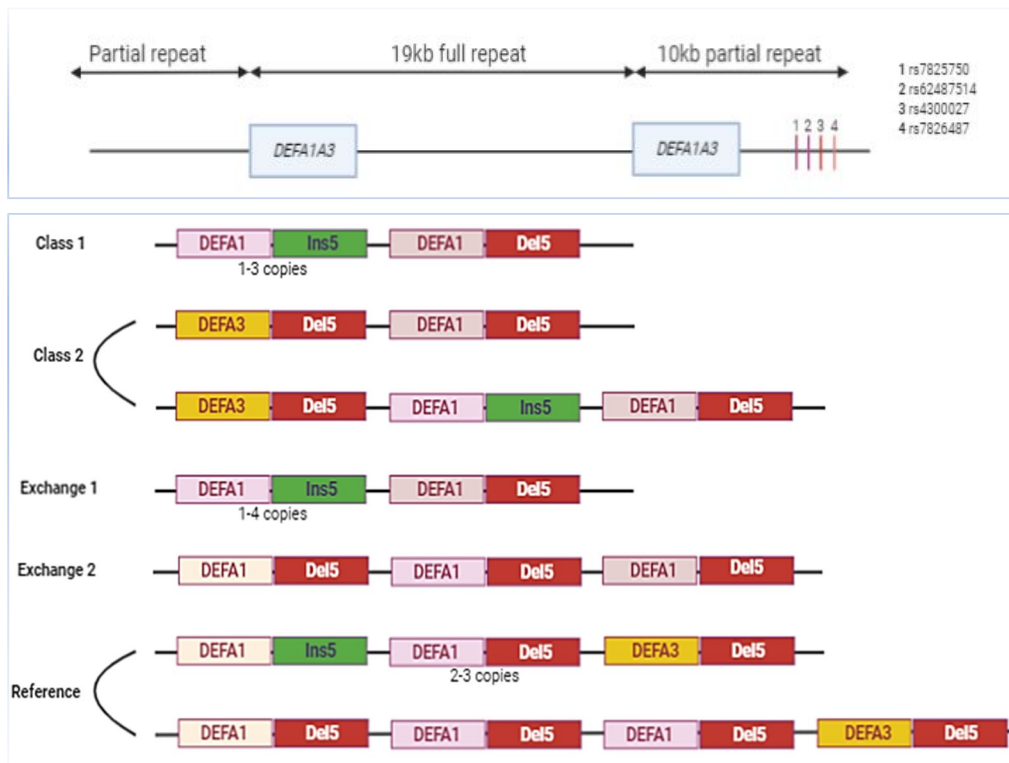


Figure 1.1. Top image: Schematic map of the *DEFA1A3* copy number variable locus on chromosome 8, based on the GRCh37/hg19 human genome assembly, representing the predicted structural arrangement of a two-copy allele. Four tag SNPs (single nucleotide polymorphisms) that distinguish the different haplotypes are located at the centromeric region of the repeat: (1) rs7825750 (2) rs62487514, (3) rs4300027 and (4) rs7826487. Bottom image: Five *DEFA1A3* haplotypic structures commonly observed with European ancestry. Ins5 and Del5 is a 5 bp indel found external of the *DEFA1A3* coding region which helps characterize the arrangement of haplotypes. Source of data from Black et al. [2]. Image created with BioRender.com

a single repeated region. Additionally, due to their unique features and high similarity in sequences with each other, the locus containing *DEFA1* and *DEFA3* was given the compound symbol *DEFA1A3* [52].

The copy variable genes of *DEFA1A3* cluster in a single tandem repeat array on chromosome 8p23.1. In the (GRCh37/hg19) human genome assembly, the *DEFA1A3* structure includes three repeats, of which two are 19 kb full repeats and one 10 kb partial repeat. The telomeric repeat (left) of *DEFA1* is 99.8% and 99.1% identical to the internal repeat and the centromeric repeat (right), respectively. The *DEFA1A3* gene cluster is a multifaceted entity as it consists of structural rearrangements between and within repeats that distinguish alleles of the gene from each other. Firstly, it consists of multi-allelic copy number polymorphism varying in the number of gene coding regions. Secondly, the ratio of *DEFA1* and *DEFA3* differs in number from one individual to another [52, 92], and some individuals lack *DEFA3* completely. Moreover, the haplotypic structural rearrangement of *DEFA1* and *DEFA3* is distinctive between individuals, including from the same population [2, 52] (Figure 1.1).

Flanking SNPs surrounding the *DEFA1A3* locus exhibit high linkage disequilibrium (LD) [2, 4], suggesting crossover events are rare within the *DEFA1A3* locus. Khan et al. identified the SNP rs4300027 as a proxy that significantly correlates with *DEFA1A3* copy number in a European population [86]. Furthermore, additional SNPs were identified within the flanking region of *DEFA1A3* CNV that may help the classification and prediction of haplotype diversity within a European population [2]. The four tag SNPs identified are as follows: rs4300027, rs7826487, rs7825750 and rs62487514. These SNPs have characteristic patterns of gene copy number and content. This has led to the identification of five *DEFA1A3* haplotypic structures commonly detected in European samples [2]. The haplotype information was obtained by using the emulsion haplotype fusion PCR (EHF-PCR) and sequencing. The structures are named as follows: Class 1, Class 2, Exchange 1, Exchange 2, and Reference

Sequence. For each haplotype, unique features are presented, with each class showing different characteristic gene copy levels as shown on Figure 1.1. The *DEFA1A3* total genomic copy number ranges from 3 to 16 copies in a European population [52, 86], with haplotypes known to range from 1 copy [92] to at least 8 copies, and there are different distributions in other population groups but for these, there has only been small-scale study [90].

α -defensins CNV expression

It is expected with gene expression, that the number of genes are likely to express equal levels of mRNA. Therefore in the case of CNVs, fewer copies should yield lower RNA output. For instance, Linzmeier *et al.* discovered that the levels of HNP-1 to HNP-3 peptide expression in neutrophils had a linear relationship to the number of gene copies of *DEFA1* and *DEFA3*, which suggests that the number of *DEFA1A3* copy number correlates to mRNA expression. The peptides are easily distinguishable from one another as they differ only in their N-terminal amino acid. In their study, the peptides were captured qualitatively by two methods (Coomassie-stained acid urea gels and Western immunoblot). However, the analysis was of a small-scale study that consisted of only 8 samples. In parallel, Aldred *et al.* showed different outcomes [52]. No association was found between total *DEFA1A3* copy number and total mRNA levels from whole blood. Nevertheless, there was a clear association between relative *DEFA1:DEFA3* mRNA levels and the ratios of equivalent gene numbers. No further studies were done to replicate these findings.

Disorders associated with α -defensins CNV

The variable *DEFA1A3* locus has been reported to be associated with several disorders. A previous study observed that high *DEFA1A3* copy number is strongly correlated with increased risk of severe sepsis in the Southeast-Asian population [93]. Similarly, high *DEFA1A3* copy number was reported to be associated with Crohn's disease in Northern-European individuals [91]. However

the quality of the CNV typing suggests that both these candidate gene studies may not be reliable, and they have not been independently replicated to date. Subsequently, various reliable and replicated GWAS studies found that variation at the *DEFA1A3* locus was associated with IgA nephropathy (IgAN) and Periodontitis [10, 94, 95]. IgAN is the most common primary glomerulonephritis, and its prevalence is higher in Asia than other continents [94]. It was recently proposed that low copy number of *DEFA1A3* increased the risk of IgAN [9], but no such conclusion was made for periodontitis. Periodontitis is the most common inflammatory disease in humans, affecting ~20-50% of the worldwide population [96]. Recently, variants near the *DEFA1A3* locus were identified as risk loci for the progression of periodontitis [95], but it is still unclear if *DEFA1A3* copy number is the key factor predisposing to periodontitis.

IgA Nephropathy

IgA nephropathy (IgAN), also known as Berger's disease, was first defined by Berger and Hinglais in 1968 [97]. IgA is an antibody named immunoglobulin A that has an essential role in mucosal immunity and there are two classes (IgA1: found in serum and derived from bone marrow, and IgA2: a secretory form of IgA). IgAN is a kidney disorder that develops when immune complexes containing immunoglobulin A antibodies (IgA) aggregate in the glomerulus, causing an inflammation and will gradually damage the normal functions of the kidneys [98]. It is the most common primary glomerulonephritis worldwide with an incidence of 2.5 per 100,000 population per year [98, 99]. IgAN symptoms include haematuria, proteinuria, edema, GIT, or UTI, but symptoms are often unnoticed during the early stages. Up to 50% of affected individuals with IgAN have persistent haematuria or proteinuria [100]. There is solid evidence that suggests IgAN is caused by genetic factors, and IgAN is more commonly found in Asia than other

continents [101]. Recently, multiple genome wide association studies (GWAS) have detected several susceptibility loci for IgAN. Yu *et al.* conducted a case-control GWAS study in a Han Chinese population (4,270 controls and 1,434 patients with IgAN) and identified susceptibility loci (*TNFSF14* and *DEFA* genes) [94]. Subsequently, Kiryluk *et al.* introduced the largest GWAS cohort which included 20,600 IgAN patients from European and East Asian ancestries. The study led to the discovery of two loci: *HLA-DBQ1* and *DEFA* genes [10]. In both studies, SNP rs2738048G at *DEFA* was detected as a protective allele against IgAN and was later confirmed to be an Asian population specific allele. In 2017, Ai *et al.* conducted an association study in two different southern Han Chinese cohorts, to evaluate the role of *DEFA1A3* CNV with IgAN. Interestingly, low levels of *DEFA1A3* copy number increased the risk of developing IgAN and renal dysfunction. This indicates that *DEFA1A3* CNV region plays a role in the pathogenesis of IgAN and is believed to be a risk factor [9].

Periodontitis

Periodontitis is the most common inflammatory disorder in humans and its prevalence is approximately 20-50% of the worldwide population, with an estimated heritability of 50% [96, 102, 103]. It is a serious infection that affects the condition of the gums. Periodontitis causes bleeding in the gums that begins with gingivitis and plaque, and then continues to cause severe destruction to the alveolar bone and the gums start to recede. The inflammation course is generally chronic and often persists long-term because periodontitis is painless. When left untreated, periodontitis can become extremely severe and will lead to tooth decay and tooth loss. Adults above 40 years of age experience tooth loss due to the disease [8, 102]. Periodontitis is mostly associated with the presence of certain bacteria that are predominantly gram-negative anaerobic microorganisms (such *Campylobacter*

rectus, *Aggregatibacter actinomycetemcomitans*, *Porphyromonas gingivalis*, *Prevotella intermedia*, *Bacteroides forsythus*, *Eubacterium nodatum*, *Peptostreptococcus micros*, *Staphylococcus intermedius*, and *Fusobacterium nucleatum* [104, 105]. Although some of these bacteria are the normal flora in the oral cavity of healthy individuals, when the right conditions are present these microorganisms present a threat to the immune system. There are several factors identified that increase the risk of periodontitis such as long-term smoking and diabetes. Other factors identified contribute to the development of periodontitis such as age and the lack of proper oral hygiene and diet. Two types of periodontitis are distinguished: chronic periodontitis (CP) and aggressive periodontitis (AgP). CP occurs in patients above 40 years of age and is the most common form. It is regarded as a slowly progressive disease unlike AgP. CP is also the major cause of tooth loss in patients above the age of 40. AgP is a rare but a highly severe form that occurs in patients below 35 years old.

Recently, SNPs near the *DEFA1A3* locus were identified as risk loci for the progression of periodontitis [102]. However, it is still unclear if *DEFA1A3* copy number is the driving factor predisposing to periodontitis. In genome-wide studies, three genes were identified as the strongest genetic risk factors for AgP: glycosyltransferase 6 domain containing 1 (*GLT6D1*), sialic acid-binding Ig-like lectin 5 (*SIGLEC5*) and Defensin Alpha 1 and Alpha 3 (*DEFA1A3*) [106]. Although genetic risk factors for AgP were identified, no genes were discovered for CP, thus suggesting that AgP is more strongly associated with genetic factors while CP is more associated with environmental factors [102, 107].

1.5. Long-Read Sequencing

With the rapid growth and advances of sequencing technologies (first, second, and third generations), it has become progressively easier, quicker, and more accurate to study the *DEFA1A3* CNV sequence diversity within and between populations. The four gold standards in genome sequencing are to obtain high

accuracy rate, high throughput, the ability to produce long reads and cost-effectiveness. DNA sequencing is categorized in three groups: first generation sequencing (such as the Sanger sequencing), second generation sequencing (also known as next generation sequencing, e.g Illumina sequencing), and third generation sequencing (also known as long-read sequencing). In the last decade, long-read sequencing technologies – such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) - have made it feasible to sequence regions uninterruptedly, overcoming the limitations of short-read sequencing in constructing repeated structures [108]. Short-read sequencing such as Illumina generates reads of 100 to 200 bp long only. Long-read sequencing technologies allow the production of accurate and significantly more contiguous genomic assemblies [109]. Although PacBio and ONT sequencing produce longer reads compared to Illumina or Ion sequencing, the per-base error rate of both PacBio and ONT base-calling remain higher. Pacbio sequences a molecule several times to deliver a high-quality consensus sequence, while ONT can sequence a molecule twice only. Lower basecalling error rates are expected from consensus PacBio compared to ONT and overall, a marginally better performance for purposes such as transcriptome complexity detection and sensitive detections of isoforms [110-113].

Nanopore sequencing is a third-generation sequencing platform, and this evolving technology can sequence DNA or RNA directly without prior amplification or reverse transcription [114]. It can sequence thousands of molecules in a shorter period, can produce long reads, single-molecule detection, portability, and low instrument cost [115]. The concept of the nanopore technology was first proposed by Kasianowicz and colleagues in the late 1980s, when it was suggested that it is possible to sequence and detect DNA using nanopore signals. These signals are disturbances in the electric current that arise as DNA goes through the pore. In over the past thirty years, the technology has been developed and refined [116-118]. Nanopore sequencing approaches are broadly divided into two categories; biological and solid state [119]. The biological state relies on

naturally arising biological nanopores, whereas the solid state relies on the use of artificial membranes including pores that are nanometre-sized. Nanopore sequencing depends on the detection of ionic current variations during the passage of DNA molecules through a pore [120]. The way the system works is that a motor protein enables a single strand of DNA to pass through a protein pore, that is embedded in a membrane which separates two different compartments from one another, by forcibly slowing down the movement of the DNA sequence to correctly read the bases. The nucleotides passing through the pore changes the ionic flow, thus producing different current signals. This distinction in signals identifies the different nucleotides, from which the sequence is analysed in real-time.

Oxford Nanopore Technologies (ONT) was the first group to release the nanopore technology to the scientific community in 2014. ONT manufactured the first ever portable sequencing device, the MinION. The MinION is a pocket-sized and a lightweight device (less than 100g). This gives users access to sequence their samples outside the research laboratory environment, or where the resources are very limited [121-124]. The MinION has been used when it was first released by early access users for research purposes only, to test its accuracy and reliability [125]. Using the MinION has many advantages. Firstly, the ability to produce longer reads in real time [125], hence any length of DNA molecules can be sequenced. The longest read that has so far been produced (2,272,580 bases in length) was in May 2018 [126]. Recently, an entire human reference genome (GM12878 Utah/Ceph cell line) has been sequenced and assembled for the first time, producing ultra-long reads [125]. As a result, 91.2 Gb of raw sequence data was produced, representing 30x coverage.

Additionally, sequencing per human genome costs about \$1000, which makes the MinION cheaper than other short-read technologies such as Illumina. Furthermore, it can detect base modifications (epigenome sequencing) [127]; since the amplification step is eliminated, the modified sequence is preserved. Unlike Illumina, the MinION can also provide very direct evidence about large-scale

structural variation, by sequencing across the entire variant region [128-130]. Additionally, it has the advantage in *de novo* genome assembly and sequencing [130]. Finally, targeted and selective sequencing approaches can be used to sequence the region of interest of highly repetitive regions [128, 130] instead of the whole-genome such as the *BRCA1* gene[128], the method that was used in their study being called Cas9-assisted targeting of chromosome segments (CATCH).

One major caveat of MinION sequencing is the high error basecalling rate compared to the short-read sequencing platforms, when the MinION was first introduced, the accuracy rate of basecalling was less than 60% [131, 132]. However, over recent years, the accuracy rate has increased to up to 85% [125, 133-135]. Therefore, the results produced from the MinION are being compared with second-generation sequencing technologies such as Illumina to validate the accuracy of the results. For this reason, Illumina is still the best option to use for sequencing in terms of accuracy. Another disadvantage of the MinION is the inability to call homopolymers accurately. A homopolymer has the same nucleotides repetitively. Additionally, the read distribution and the length of reads are dependent on the quality of the DNA. Degraded DNA results in shorter reads, whereas an intact DNA sequence results in high quality reads.

The GridION and PromethION are also manufactured by ONT, but these sequencing technologies are not hand-held although are suitable for large scale projects. The Flongle was recently developed by ONT, which is an adaptor introduced to either MinION or GridION devices that facilitates direct, rapid, and real time sequencing on a smaller and single use flow cell. It is suitable for smaller assays, or for targeted approaches and is cost-efficient as minimal as \$90 per flow cell (inexpensive compared to other sequencing devices). It is similar to MinION, but it has a smaller flowcell (126 channels instead of 512). It has been introduced in several reports to study gene fusions, pathogens (bacterial and viral), epigenetic changes and structural variations [136-140].

Nanopore sequencing provides ample opportunities in discovering the genetic and epigenetic effects on various organisms of interest. Many approaches are possible with MinION to investigate different research questions such as population genetics, cancer genetics or epidemiological approaches [129].

It can produce high yields for large genomes, it is a rapid approach, library preparations are straightforward, and it produces real-time analysis. The rate of accuracy is gradually increasing with the improvements made on the MinION, thus this will yield in higher contiguity. Thus far, Oxford Nanopore has been used in numerous studies in conjunction with other complementary technologies, such as Illumina and PacBio, due to its high error rate. Although Illumina is a short-read sequencing technology, the accuracy of sequencing and assembly of any genome is higher. Therefore, it is a still preferred approach to use in large scale studies such as the 1000 Genomes Project. However, the appeal in the future is to sequence DNA and RNA with long-read approaches, such as ONT is offering. One of the key challenges in next generation sequencing is inferring the correct haplotypes due to their short read lengths. The initial attempt to assemble a diploid human genome depended on highly accurate and relatively long reads (approximately 500 to 1000 bp long) from Sanger sequencing [141]. However, these assemblies were fragmented and are unable to resolve many repetitive regions. Structural variations (SV) and high sequence similarity are usually collapsed into one region after mapping to genome assemblies [142]. This issue becomes a challenge in characterizing true haplotype structures such as CNVs in tandem repeats. Therefore, the use of long-reads may potentially be the best approach in delivering novel capabilities - such as the capture of SV resolution - with targeted sequencing. *DEFA1A3* CNVs will be used as an example to demonstrate the need of long read sequencing and applying a targeted (selective) approach to capture only the CNV region which is more cost and time-efficient.

1.6. Hypothesis and Aims of Research

The aims of this PhD research are to investigate key features of variation in the *DEFA1A3* CNV region relevant to functional impact. Firstly, two SNPs - downstream of the *DEFA1A3* CNV - were found to be significantly associated with periodontitis [102]. However, no evidence has been gathered to study the association between the *DEFA1A3* CNV and periodontitis – specifically aggressive periodontitis. Thus, the aim of this study is to investigate this association by using the 1000 Genomes Project as controls (European population only) and German cases (from the Munz *et al.* study [102]) with their relevant *DEFA1A3* copy number information.

Secondly, the amount of HNP1-3 in neutrophils has been reported to be proportional to total *DEFA1A3* gene copy number [90], but the analysis was of a small scale study. In parallel, Aldred *et al.* [52] showed different outcomes and found no relationship between the total *DEFA1A3* copy number and total mRNA levels from whole blood, but there was a clear association between relative *DEFA1:DEFA3* mRNA levels and the ratios of equivalent gene numbers. Therefore, this study aims to investigate such association: a) between total *DEFA1A3* copy number and RNA levels in neutrophils, T-cells and monocytes and b) between *DEFA1:DEFA3* genomic ratio and *DEFA1:DEFA3* transcript ratio. This work will use a larger sample size from the Blueprint Epigenome Study [143] as the results of previous studies did not agree on the consequences of gene copy number for gene expression [9, 52, 71].

Thirdly, using a third-generation sequencing platform (Oxford Nanopore Technology), we will assess the feasibility of constructing full *DEFA1A3* CNV haplotypic sequences by utilizing two different sequencing approaches: a) CRISPR-cas9 enrichment system in a MinION sequencer and b) Long amplicon sequencing in a flongle adaptor and a MinION sequencer. The objective of the long amplicon sequencing is to include five long amplicons of 5.5 kb to 7.5 kb long per

DEFA1A3 repeat with each amplicon encompassing overlapping regions to one another (tiling path). The flongle adaptor will be used to assess its accuracy and performance in sequencing the *DEFA1A3* CNV. The two sequencing approaches will then be used to compare the results with previous studies and experimental methods in inferring the *DEFA1A3* haplotype structures [2].

CHAPTER 2

2. Materials and Methods

2.1. Aggressive Periodontitis Association Study

2.1.1. Study Samples

Data from 503 European controls (1kGP-EUR) in the 1000-Genomes Project (1kGP) from 5 European subpopulations (Table 2.1) and 717 German cases (AgP-GER) with AgP (Aggressive Periodontitis) were used to perform case-control association analysis; periodontitis case data were obtained under the terms of a Materials Transfer Agreement with CHARITÉ– Universitätsmedizin Berlin (Charitéplatz 1, D-10117 Berlin) from Arne Schaefer and Matthias Munz [6]. As preliminary data, the combined dataset provided 5581 variants (hg19;chr8:6700000-7100000) and 1220 individuals. Additionally, for a pooled analysis, 197 further UK controls (BP-GBR) from the Blueprint study below (section 2.2) were included in the analysis [143]. *DEFA1A3* copy number measurements were obtained by using the read-depth [51] method on the 1kGP-EUR and BP-GBR controls, while a previous MRes student (Eric Moise) typed the *DEFA1A3* copy number in 744 German cases using the PRT assay (paralogue ratio test).

Table 2.1. Five European subpopulations included from the 1000 Genomes Project in the case-control study.

Symbol	Description
GBR	British in England and Scotland
TSI	Toscani in Italy
FIN	Finnish in Finland
IBS	Iberian populations in Spain
CEU	Utah residents with Northern and Western European ancestry

2.1.2. 1000-Genomes Project CNV measurement

Sequence data for 2,504 samples from phase 3 of the 1000 Genomes Project or 1kGP were included in the analysis. These samples consist of four super populations which are presented on Table 2.2.

Table 2.2. Four superpopulations from the 1000 Genomes project. EUR: European, ASN: Asian, AMR: American, and AFR: African population

Super population	Number of samples
EUR	503
ASN	1003
AMR	347
AFR	651

Read depth was determined for the *DEFA1A3* region to obtain estimated *DEFA1A3* copy number by using `samtools` version 1.9 [51] with the read-counting command `samtools view -c`, defining the intervals with a corresponding bed files (Table 2.3), and directed by a Perl script produced previously by Prof John Armour. Two independent genomic intervals were used for *DEFA1A3* copy number estimation (DEFA1core and DEFA2core+2). DEFA1core measures the parts of *DEFA1A3* CNV that are represented three times (3 *DEFA1A3* copies) on the GRCh37/hg19 genome assembly while DEFA2core+2 measures the parts of the CNV represented two times on the assembly. In diploid genomic DNA, DEFA2core therefore has 2 copies fewer (*DEFA1A3* gene copy

minus 2). The following bed files correspond to the relevant locations of *DEFA1A3* CNV and the flanking regions (table 2.3): DEFA1core covers 9kb which includes the *DEFA1A3* CNV repeats; DEFA2core covers 5 kb which includes the *DEFTIP* pseudogene; DEFAflank which includes the (copy number variant) flanking regions that cover 140 kb. Read counts from DEFAflank were scaled by a factor of 1.06 to correct for small but reproducible differences in read mapping density in this region relative to the *DEFA1A3* repeats. The value of 1.06 was determined empirically based on observation of a small but consistent difference in mapping densities between the repeat and flanking regions in the same samples, using integer clusters as reference points for the repeat regions. After we obtained the read counts from the `samtools` command, the estimated copy numbers were then derived from read density per kb using these two calculations:

$$DEFA1_CN: DEFA1/9 \times 140/(DEFAflank * 1.06) \times 2$$

$$DEFA2_CN: DEFA2/5 \times 140/(DEFAflank * 1.06) \times 2$$

To justify the accuracy of the read-depth method, the mean copy number estimate was calculated from measurements from both DEFA1core and DEFA2core+2 on the same samples. Next, the obtained mean values were normalized by their nearest integer value. For example, if a mean value of a sample was 8.97, then the rounded integer value will be 9, and therefore the normalized value of that sample is 0.9966. The closer the normalized ratios are to 1, the better calibrated the test is.

Table 2.3. Bed files corresponding to the relevant chromosomal locations (GRCh37/hg19) of the DEFA1A3 CNV repeats and the flanking regions.

Bed files	Chromosomal location
DEFA1core	chr8:6829200-6836206
	chr8:6837208-6839209
	chr8:6848360-6855366
	chr8:6856368-6858369
	chr8:6867463-6874469
	chr8:6875471-6877472
DEFA2core+2	chr8:6841527-6846531
	chr8:6860632-6865636
DEFAflank	chr8:6710000-6790000
	chr8:6900000-6910000
	chr8:6920000-6940000
	chr8:6960000-6990000

2.1.3. Quality control and filtering

Quality control (QC) and filtering were applied to SNP genotyping data from Berlin collaborators prior to case-control association (Table 2.4). PLINK v1.90 software was used [144]. This reduced the number of SNPs and individuals. Firstly, SNP information in EUR controls was from the GRCh38/hg38 assembly, and therefore coordinates from the AgP-GER and BP-GBR SNPs were lifted over to match coordinates to the EUR controls. SNPs were excluded with genotype call rate <98%, MAF (minor allele frequency) <1%, or HWE (Hardy-Weinberg Equilibrium) deviation with a value of $P < 10^{-4}$. This resulted in the removal of 5043 SNPs out of 5581. Additionally, the genotyping call rate per individual <98% excluded 145 samples from further analysis. Lastly, samples with missing copy number information were removed, which excluded 118 samples.

2.1.4. SNP selection criteria

Top SNPs (see chapter 3, section 3.3, Table 3.1) were selected (from the 1000 Genomes Project European controls and German cases) based on **1)** the lowest P -values from the logistic-regression analysis without adjusting for covariates (discovery stage), **2)** Variants with $MAF > 0.05$ **3)** and SNPs with strong LD ($r^2 > 0.8$).

Table 2.4. Overview of the study population, QC and filtering. *No copy number information available for German controls in this study. 1kGP: 1000 Genomes Project, Pooled-analysis: includes 197 UK-based samples from the Blueprint Epigenome Study [143]. Dataset obtained from Munz et al. [102].

	German cases + 1kGP controls		German cases + German controls		Pooled-Analysis*	
Type	Cases	Controls	Cases	Controls	Cases	Controls
Pre-QC						
No. of individuals	717	503	717	4199	717	700
Males	302	240	302	2012	302	330
Females	415	263	415	2187	415	367
Unknown sex	-	-	-	-	-	3
SNPs	1022	5534		1022		
Data merged			Pooled			
No. of samples		1220		4916		1417
SNPs		5581		1022		6524
QC						
Genotype Call Rate <0.98		4880		361		5043
MAF < 0.01		0		0		0
HWE < 10^-4		0		0		2 SNPs removed
Genotype Call Rate per individual <0.02	117 individuals removed		544 individuals removed		145 individuals removed	
After QC						
SNPs after QC		701		661		585
Individuals after QC		1101		4372		1272
No of samples with CN information	566	503		NA	572	700
*No of samples after QC and exclusion criteria	479	503	605	3767	454	700

2.1.5. Association Analysis (single-SNP locus vs multi-SNP locus)

SNP association to AgP in AgP-GERx1kGP-EUR and the pooled analysis (including the BP-GBR controls) was tested using logistic regression with discrete (*DEFA1A3* copy number) and binary variables (sex) used as covariates, respectively. SNP association to *DEFA1A3* copy number in both stages of analysis was tested using the quantitative association analysis.

2.2. Blueprint Study

2.2.1. Study Samples

Chen *et al.* collected blood samples from 200 healthy individuals from a UK based population on which multiple analysis methods were applied to highly pure sorted cell subsets: a) whole-genome sequencing (WGS) was performed on genomic DNA, and b) RNA-seq, DNA-methylation, and ChIP-sequencing for two histone modification marks (H3K4me1 and H3K27ac) were performed on samples from neutrophils, monocytes, and naïve T-cells [143].

2.2.2. Data Access

An application for data access was made to the Blueprint Epigenome Access Committee (<http://dcc.blueprint-epigenome.eu/>), and access to eight datasets was granted (Table 2.5). WGS dataset (EGAD00001002663) was used to measure read-depth [2, 51] and to obtain *DEFA1A3* total genomic copy number for each individual (see section 2.1 from the 1000 Genomes Project CNV measurement).

Table 2.5. Access to eight datasets from the Blueprint Epigenome Project was granted which includes one WGS dataset, three RNA-seq datasets, three ChIP-Seq datasets, and one DNA methylation profile dataset.

Dataset ID	Description	Samples
EGAD00001002663	Whole-genome sequencing	197
EGAD00001002670	ChIP-Seq data for mature neutrophil samples	186
EGAD00001002671	RNA-Seq data for CD4+ T-cell samples	212
EGAD00001002672	ChIP-Seq data for CD14+CD16- classical monocyte samples	174
EGAD00001002673	ChIP-Seq data for CD4+ T-cell samples	158
EGAD00001002674	RNA-Seq data for CD14+CD16- classical monocyte samples	197
EGAD00001002675	RNA-Seq data for mature neutrophil samples	205
EGAD00010000850	DNA methylation profiles of neutrophils, T-cells and monocytes	525

781 cram files from WGS sequencing from the European Genome-Phenome Archive (<https://ega-archive.org/>) were downloaded. The cram files for samples from 197 donors were then decrypted, and reads aligned to the *DEFA1A3* region were extracted and converted to bam files; finally, the original cram files were deleted to efficiently save storage space on the local DeepSeq server. RNA-Seq datasets for neutrophils (EGAD00001002675) and monocytes (EGAD00001002674) consists of single-end reads while naïve T-cells (EGAD00001002671) had data consisting of paired-end reads. Therefore, the approach taken in normalising the RNA-seq data depended on the type of reads produced. Single-end read normalisation used the RPKM (Reads Per Kilobase of transcript, per Million mapped reads) method while paired-end reads used FPKM (Fragments Per Kilobase of transcript per Million mapped reads) [145] (Figure 2.1).

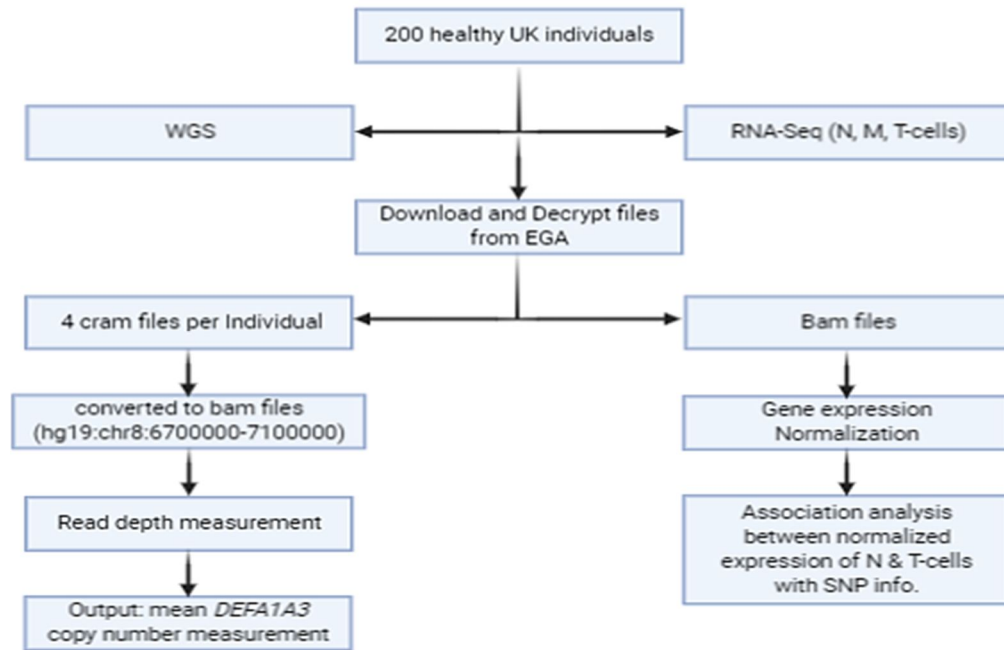


Figure 2.1. A flowchart of the steps performed to obtain the *DEFA1A3* copy number measurements from WGS, and normalized gene expression counts from RNA-Seq from neutrophils, T-cells, and monocytes. (N: neutrophils, M: monocytes, EGA: European Genome-Phenome Archive).

2.2.3. Data Bioinformatic Analysis

A `samtools+awk` (v1.9 and v4.0.2) command originating from Masella *et al.* [146] was used on genomic and RNA-seq bam files to filter reads based on the status of a single nucleotide, to help in identifying the *DEFA1:DEFA3* ratio (Figure 2.2).



Figure 2.2. Allele sorting using a *samtools+awk* syntax command by filtering reads based on the status of a single nucleotide.

Both genomic and RNA bam files were re-aligned to a single telomeric *DEFA1A3* repeat reference sequence (hg19:chr8:6790000-6850000) to determine the count of *DEFA1*(G) and *DEFA3*(T) alleles. The *samtools* syntax is as follows:

```
$ samtools view -h $1 input.bam | gawk '{{ position =
35383; mapped_position = 0; actLength = 0}{ split($6,
len , /([A-Z]+)/); split($6, op , /([0-9]+)/); }{ for
(i = 0; i < length(op); i++){if (op[i+1] ~ /M/ || op[i+1]
~ /D/ || op[i+1] ~ /N/ || op[i+1] ~ /X/) actLength =
actLength + len[i]; }} { if (and($2,
16)){required_offset = $4 + actLength - position-1;}
else {required_offset = position - $4; }}{ for(i = 1; i
< length(op); i++){ for (incr = 0; incr < len[i] &&
required_offset >= 0; incr++){if (op[i+1] ~ /M/ ||
op[i+1] ~ /D/ || op[i+1] ~ /N/ || op[i+1] ~ /X/)
required_offset--;if(op[i+1] ~ /M/ || op[i+1] ~ /I/ ||
op[i+1] ~ /S/ || op[i+1] ~ /X/) mapped_position++;}}}}
(/^@/) || (position >= $4 && $4+actLength >= position
```

```
&& substr($10,mapped_position) ~ /^G/ && $6 !~ /*/) {
print $0 }'|samtools view -b -S -o $2 output.bam -
```

The most important specific characters in the syntax are highlighted in red. The first one ‘35383’ represents the bp-position used as the basis of sorting, and the second one ‘G’ represents the specific allele selected for sorting into the output. In parallel, `bcftools mpileup` [147] and allele depth switch were also used to determine the depth of both *DEFA1*(G) and *DEFA3*(T) alleles. The syntax of the relevant commands is as follows:

- a. `bcftools mpileup -f /PATH/TO/xxx.fasta --annotate INFO/AD input.bam -Ou | bcftools call -mv -Ou | bcftools filter -s LowQual -e '%QUAL<20 || DP>100' > output.vcf`
- b. `bcftools query -f '%CHROM\t%POS\t%REF\t%ALT\t%INFO/AD{0}x\t%INFO/AD{1}\n' input.vcf > output.AD`

2.3. Nanopore Sequencing analysis

Figure 2.3 shows the stages used to analyse results post-sequencing. Raw fastq files were used to align reads to a reference sequence with Minimap version 2.16-r922 [148]. Mapping statistics of each read were used to identify read length and bedtools version 2.25.0 [149] was used to identify strand direction. Canu version 1.8 [150] was used to generate consensus sequences and *freebayes* and `bcftools` version 1.5 [147] to call variants.

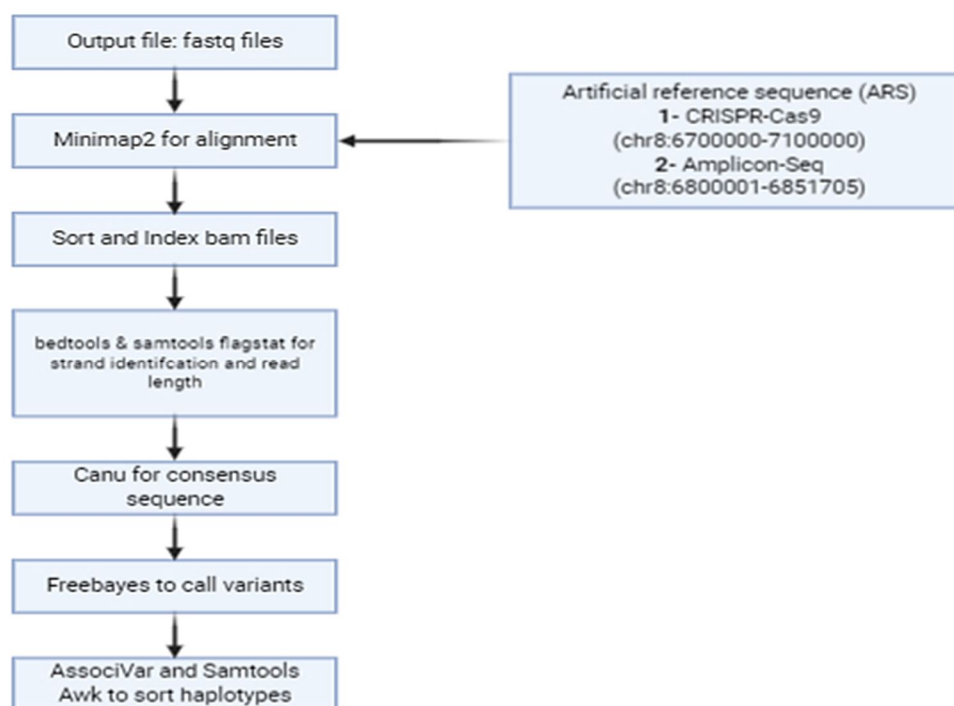


Figure 2.3. Steps performed on nanopore-sequence data (CRISPR-Cas9 enrichment method in a MinION, and amplicon sequencing method in a flongle and a MinION).

2.3.1. CRISPR-Cas9 enrichment sequencing

Guide RNAs (gRNAs) for the top and bottom strands were designed using the CHOPCHOP (<https://chopchop.cbu.uib.no/>) and IDT (<https://eu.idtdna.com/>) websites. gRNAs used were: gRNA1: TCCATACACACGGCGTGAGT and gRNA2: TGTGAGGGTCTCCAAACAAT (Table 2.6, Figure 2.4). Nanopore Cas9 Targeted-Sequencing (nCATS) was performed by DeepSeq following the protocol published recently [139]. An artificial reduced reference sequence containing only the *DEFA* region was created in fasta format with the corresponding coordinates (hg19:chr8:6700000-7100000).

Table 2.6. The on-target scores below indicate the predicted genome editing performance of gRNA at the target site, the higher the score the better the gRNA performance, while the off-target scores indicate that the higher the score the lower the risk of off-target genome editing.

gRNA strands	On-target score	Off-target score
DEFA1A3gRNA1	62	91
DEFA1A3gRNA2	82	58

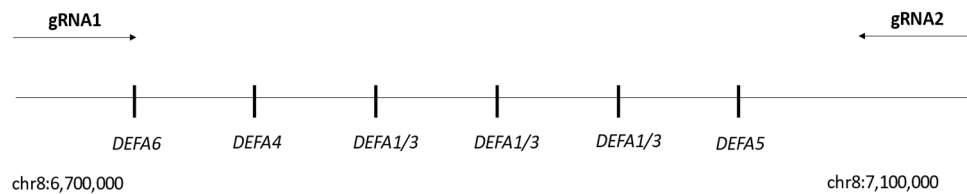


Figure 2.4. The position of gRNAs relative to the *DEFA1A3* repeat region and the neighboring flanking sequences. The gene symbols indicate the order and content only, and positions are not shown to scale.

2.3.2. Amplicon tiling-path sequencing

To limit PCR-mediated recombination (to produce chimeric amplicons) arising from long PCR, 5 pairs of primers were designed from Primer3web version 4.1.0 (<https://primer3.ut.ee/>) to tile across the *DEFA1A3* repeat in amplicons of 5-7kb and were ordered from Integrated DNA Technologies or IDT (<https://eu.idtdna.com/>) (Table 2.7). The predicted product length is estimated as ~5-7kb. Each amplicon has overlapping regions with its neighbours (tiling PCR) (Figure 2.5). Primers were designed to avoid common positions of substitutional variation in the repeat unit, as determined from custom analysis of 1000 Genomes data by former Ph.D. student Xiao Xu, and tested on UCSC In-Silico PCR browser GRCh37/hg19 (<https://genome.ucsc.edu/cgi-bin/hgPcr>) to evaluate their predicted corresponding locations on the genome assembly. PCR1, PCR4, and PCR5 have 3 products present each, while PCR2 and PCR3 show 2 products each. The corresponding positions for each PCR on the GRCh37/hg19 assembly are as follows: old PCR1 (chr8:6835034-6841077, chr8:6854151-6860182, and chr8:6873254-6879278); new PCR1(chr8:6835034-6842185, chr8:6854151-

6861290, and chr8:6873254-6880498); PCR2 (chr8:6837642-6844108 and chr8:6856752-6863211); PCR3 (chr8:6842128-6848588 and chr8:6861233-6867691); PCR4 (chr8:6828084-6834693, chr8:6847240-6853810, and chr8:6866343-6872913); and PCR5 (chr8:6830738-6836557, chr8:6849897-6855674, and chr8:6869000-6874773).

Table 2.7. Forward and reverse primers designed using Primer3. PCR1_old and PCR1_new only differ in their reverse primers. Reverse PCR1_new amplifies a region that extends to include the four centromeric tag SNPs in the centromeric repeat of DEFA1A3.

Primer Name	Forward 5'-3'	Reverse 5'3'
NanoPCR1_old	GCCCAAACCTCCTGCTTGAAA	TATGTSTGCTGCTCCTGGAA
NanoPCR1_new	GCCCAAACCTCCTGCTTGAAA	TCCTAGAAAGCTGGTTGCTC
NanoPCR2	AGGTCTTTTGGGATGAGGTGT	AAGACGCACTCTCCTCTGTT
NanoPCR3	GAACTGAACCCACCTGCAAG	TGCCACACATCAGGAATTCAC
NanoPCR4	CTCCTGTAGCAAGAGCAGCC	CACAGCATCTCACATTGGCAC
NanoPCR5	TTGGACTTCTCAGCCTCGAG	AACAGCAGGACCCATCTTCA

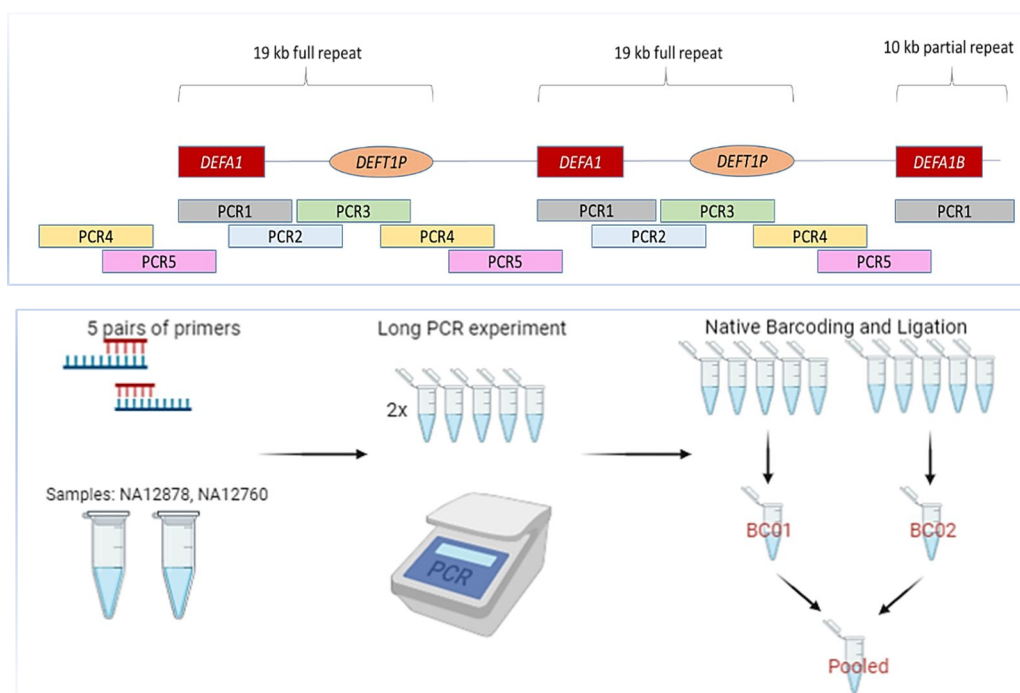


Figure 2.5. Visual representation of the 3 repeats as viewed on the UCSC Genome Browser from the 2009 human genome assembly (Ch37/hg19) and the corresponding positions of PCR products (top). Steps of amplicon sequencing for NA12878 (BC01) and NA12760 (BC02) prior to loading into a flongle. The chromosomal locations are not accurately to scale but only for illustration purposes.

Stock primers were reconstituted with PCR-grade water (H₂O) to make 100 μ M stock solution and working primers to make 10 μ M of primer solution were prepared. LongAmp[®] Taq DNA Polymerase reagents were purchased from New England Biolabs Inc (<https://international.neb.com/>). The general master mix for Long PCR experiments was as follows (per 20 μ L reaction): 12.4 μ L of H₂O; 4 μ L of 5x LongAmp buffer; 0.6 μ L each of 10 μ M forward and reverse primers; 0.6 μ L of 10 mM dNTPs; 2 units of LongAmp DNA polymerase; and 1 μ L of (20 ng/ μ L) input DNA. The cycle conditions were as follows: the initial denaturation step 95 °C for 30 sec followed by 32 cycles of 95 °C for 30 sec, X °C for 15 sec, and 65 °C for 7 min; and the final extension step was 65 °C for 20 min. The optimal annealing temperature was determined after a series of long PCR reactions using different annealing temperatures. 0.8% agarose gels in 0.5x TBE were prepared for gel electrophoresis.

Sequencing of 5 amplicons per sample was performed in each sequencing experiment, between them, the amplicons covered all of the region included by CRISPR-Cas9 sequencing (Chapter 6):

1. A flongle (a flow-cell adaptor for MinION or GridION) was loaded to a GridION X5 machine for sequencing. Native Barcoding Expansion (EXP-NBD104) and Ligation Sequencing Kit (SQK-LSK109) were purchased from Oxford Nanopore Technologies (<https://nanoporetech.com/>). 1D Native barcoding genomic DNA protocol was performed according to the manufacturer's instructions [151], except that 5 PCR products from each sample were mixed in an Eppendorf tube each with equimolar measurements instead of using genomic DNA as input. DeepSeq staff loaded the samples into the flongle. Fastq files were produced for each barcoded sample using Porechop. An artificial reference sequence that contains only the first (telomeric) *DEFA1A3* repeat was created in a fasta format with the corresponding coordinates (hg19::chr8:6800001-6851705).
2. A MinION Spot-ON flowcell was added to a GridION X5 Machine for sequencing. A pooled sample consisting of 5 barcoded samples (5 amplicons per sample) was loaded. Originally, 16 samples were included for the long PCR experiment and subsequently were going to be sequenced (Table 2.8, samples number 3 to 18). However due to time constraints and the inability to produce 5 complete amplicons per sample at a time, we only culminated with 5 samples (5 amplicons per sample) of high quality.

Table 2.8. Samples used for nanopore sequencing. Samples number 1 and 2 were only used in the flongle experiment. Sample 2 was used for the CRISPR-Cas9 experiment.

Sample No.	Sample Name	Sample Concentration (ng/uL)
1	NA12760	20
2	NA12878	20
3	NA06985	10
4	NA12248	10
5	NA07347	10
6	NA11993	10
7	NA11930	10
8	NA11931	10
9	NA07056	10
10	NA06993	10
11	NA07055	10
12	NA11832	10
13	NA12249	10
14	NA07034	10
15	NA07345	10
16	NA11829	20
17	NA07000	20
18	NA06994	20

Repeat Sorting process

For the repeat sorting experiment, see Section 2.2, Figure 2.2. The table below (Table 2.9) shows phased SNPs from the Tyson *et al.* study that were used as a reference point in our current analysis, to look at the variations in the overlaps between repeats after sorting (these phased SNPs are found in the overlaps between PCR1 and PCR5 amplicons).

Table 2.9. Phased SNPs from Tyson et al. were used to assist in the sorting of the repeats and characterizing their location in a haplotype, in addition to identify the DEFA1(G)/DEFA3(T) genes. Variant number 1 is the DEFA1/DEFA3 SNP.

Variant No.	Variant bases	SNP	Position
1	G/T	rs145076681	chr8:6835383,6854499,6873603
2	A/C	rs148298380	chr8:6835489,6854605,6873708
3	T/C	rs139164630	chr8:6835506,6854622,6873725
4	G/T	rs147777993	chr8:6835515,6854631,6873734
5	A/T	rs77140339	chr8:6835640,6854756,6873859
6	T/G	rs77649030	chr8:6835646,6854762,6873865
		chr8:6835780	
7	C/T	chr8:6845897	chr8:6835780,6854897,6873996
		chr8:6873996	
8	G/T	rs77855524	chr8:6835821,6854938,6874037

CHAPTER 3

3. Aggressive Periodontitis Association Study

3.1. 1000 Genomes Project *DEFA1A3* copy number measurement

This part of the work made use of 1000 Genomes Project sequence data, derived from 2,504 samples from four super populations (American, Asian, African, and European). The aim was to measure the *DEFA1A3* copy number for each individual sample, and to specifically use the European samples as controls for the case-control association analysis with Aggressive periodontitis (AgP). The Munz *et al.* [102] dataset that we have access to includes cases from the German and Dutch samples. However, the Dutch samples are excluded from the study as they do not have any *DEFA1A3* copy number information.

The distribution of the copy number estimates, normalized relative to the closest integer value in each case, follows an approximately Gaussian distribution (data not shown). The copy number ranged from 3 to 17 copies as shown on Figure 3.1. The DEFA1core and DEFA2core+2 measures show a strong correlation ($r^2=0.9761$) (see chapter 2, section 2.1, section *1000 Genomes Project CNV measurement*). The definite clustering around integer values is prominent from 3 to 8 copies. As expected, after 8 copies the clustering appears to gradually become less defined as the copy number increases. Sample HG02554, from an African-

Caribbean ancestry (ACB), was excluded from Figure 3.1 as an outlier far out of the normal copy number range (28 copies).

Figure 3.2 demonstrates the deduced integer copy number distribution in the four different super populations (European, Asian, African, American) defined in the 1000 Genomes Project. The frequency of *DEFB1A3* copy number appears to decrease after 8 copies in EUR and 6 copies in ASN, AMR and AFR samples. The modal copy number observed in ASN, AMR and AFR is 6 copies and 8 for EUR samples. Individuals with copy numbers above 10 are more frequent in the ASN and AFR populations, and the highest copy numbers observed (15-17) are not seen at all among the EUR and AMR samples.

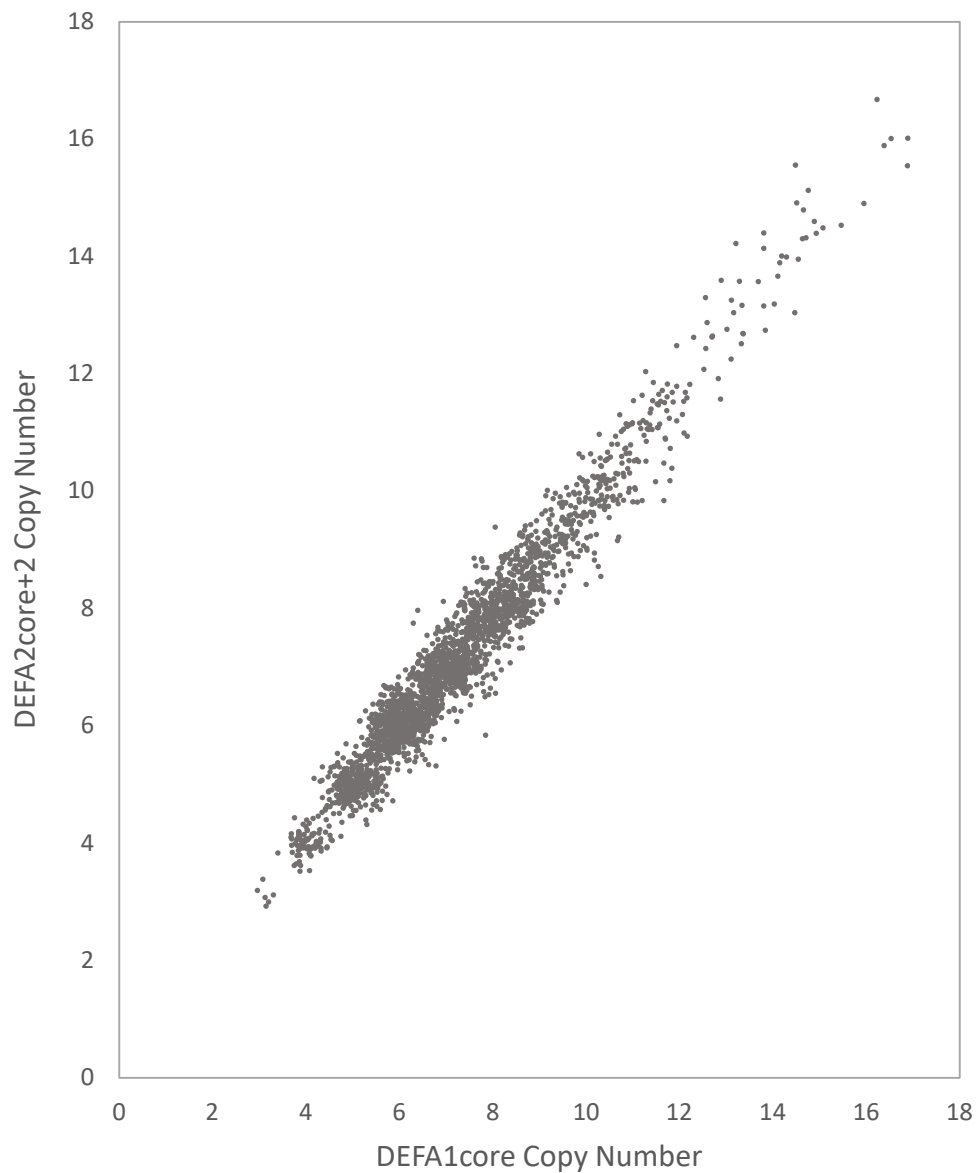


Figure 3.1. DEFA1A3 copy number estimates of DEFA1core plotted against the copy number estimates of DEFA2core+2 from 2,503 samples (1 sample excluded as an outlier), showing the correlation between the two estimates (r^2 : 0.9761) and evidence of clustering around integer values from 3 to 8.

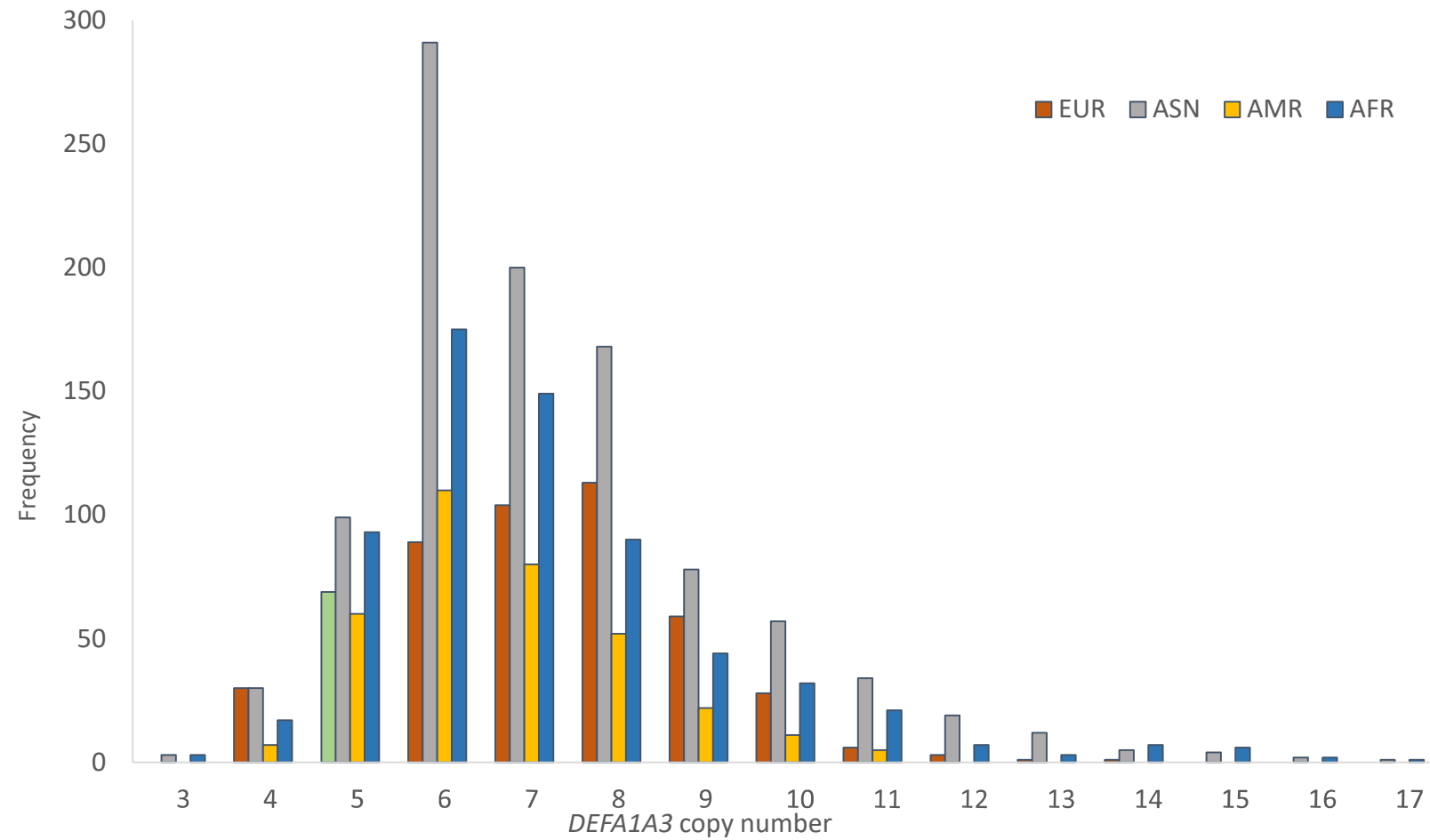


Figure 3.2. *DEFA1A3* copy number distribution in four super populations. EUR: European, AMR: American, ASN: Asian, AFR: African.

3.2. *DEFA1A3* copy number distribution

We aimed to use the *DEFA1A3* copy number estimates as a covariate in the SNP association analysis to understand the effect it might have on predisposition to AgP. *DEFA1A3* copy number in (European) cases and controls ranged from 4 to 14 copies per diploid genome (Figure 3.3: mean: 6.98, median: 7.00; standard deviation: 1.61), with 6 and 7 copies being the most common in cases and 7 and 8 copies in controls. In Figure 3.3, deduced *DEFA1A3* copy number in controls were measured using the read depth method, while *DEFA1A3* copy number for cases was measured experimentally using the paralogue ratio test (PRT) by a previous MRes student from our lab group (Eric Moise, unpublished data).

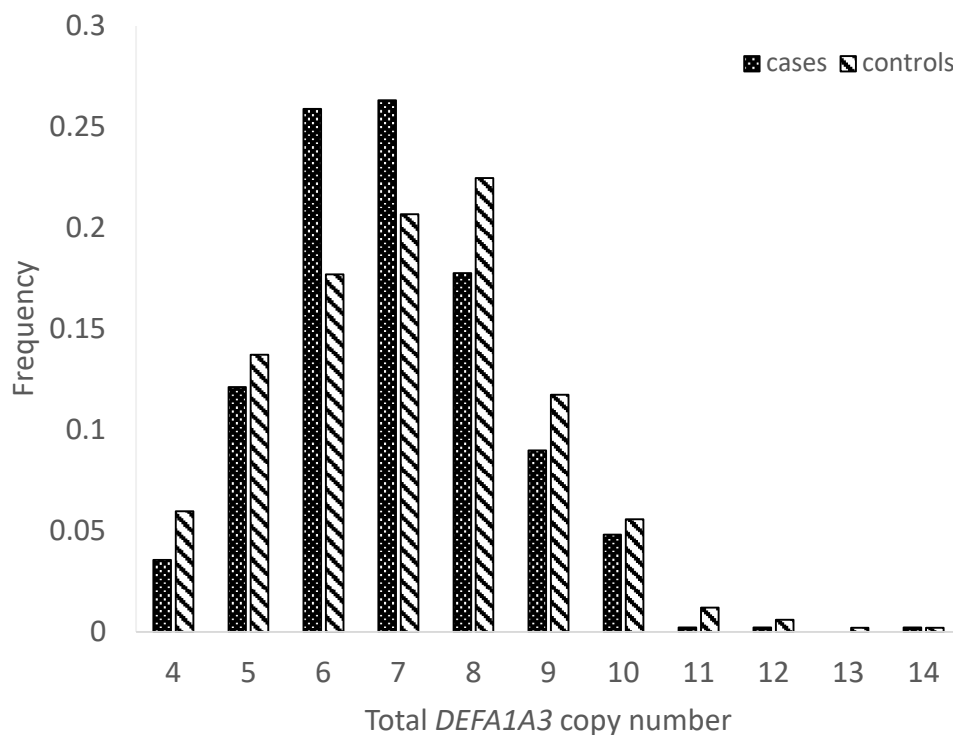


Figure 3.3. *DEFA1A3* copy number distribution in 503 European controls from the 1000 Genomes Project (1kGP-EUR) and 566 German cases (AgP-GER). This figure only reports the *DEFA1A3* deduced copy number from the 1kGP-EUR and, out of 744 AgP-GER cases, only 566 cases were included as the rest of the samples (178 AgP-GER cases) are not included in the data available from the case-control study. Deduced copy number in controls were measured via read depth while cases were measured experimentally using the PRT (paralogue ratio test) by a previous MRes student.

3.3. Top selected SNPs

Logistic regression was used without adjustment and reported 144 SNP variants out of 702 that were associated with AgP with at least nominal statistical significance (P -values < 0.05) (Figure 3.4, Table 3.1). Among those SNPs, 20 variants with the lowest P -values were selected, and after further investigation of LD (linkage disequilibrium), were reduced to 7 variants. Some variants were in strong LD with other neighbouring SNPs (such as rs4841776 in LD with 7 other SNPs) and therefore reduced the number of lead variants. The variants that were considered for further analysis showed consistent MAF with other neighbouring SNPs in LD except for rs59878581 (MAF: 0.16; rs76333965 MAF: 0.08; rs2615771 MAF: 0.32).

Lead SNPs rs57688776 and rs75098331 are in strong LD (r^2 : 0.93) with approximately equal MAF (rs57688776: 0.16 and rs75098331: 0.15), and therefore rs57688776 was selected for analysis. SNP rs7016650 was selected, as it is in strong but incomplete LD with rs74556447 (r^2 : 0.89) and rs6998500 (r^2 : 0.73). SNP rs59878581 was selected, because it is in weak LD with rs76333965 (r^2 : 0.44) and rs2615771 (r^2 : 0.42). SNP rs11774240 was chosen (near-complete LD with rs35397308, r^2 : 0.94; and rs17078234, r^2 : 0.95). SNP rs4841776 was selected, and it is in almost complete LD with rs2951863 (r^2 : 0.97) and rs2978887 (r^2 : 0.98). SNP rs2075835 was selected for analysis and is in strong LD with rs6998500 (r^2 : 0.90). Finally, SNP rs2702934 was selected and is an independent variant.

Table 3.1. Top SNPs selected at 7 suggestive risk loci on chromosome 8. SNP: single nucleotide polymorphism, r^2 : coefficient of determination BP: base-pair position (hg19:chr8: 6700000-7100000), minor: minor allele, major: major allele, MAF: minor allele frequency, OR: odds ratio, CI: confidence interval, P: P-value, AgP-GERx1kGP-EUR: case control study of German cases and 1000 Genomes Project European controls, Pooled: Blueprint samples (see Chapter 2 section 2.2) were also included as controls in this study. P-values presented here are not adjusted for multiple testing.

	AgP-GERx1kGP-EUR										Pooled			
	SNP	r ²	BP	minor	major	MAF	OR	95% CI		P	OR	95% CI		P
1	rs57688776		6962046	G	A	0.16	0.61	0.48	0.79	1.14E-04	0.67	0.53	0.85	9.37E-04
1.1	rs75098331	0.93	6962733	C	T	0.15	0.61	0.48	0.79	1.49E-04	NA	NA	NA	NA
2	rs7016650		6949736	G	A	0.16	0.63	0.49	0.81	2.86E-04	0.70	0.55	0.88	2.61E-03
2.1	rs74556447	0.89	6948870	T	C	0.15	0.63	0.49	0.81	3.06E-04	0.70	0.55	0.89	3.33E-03
2.3	rs6998500	0.73	6950859	A	T	0.12	0.59	0.45	0.79	3.54E-04	0.66	0.50	0.87	3.13E-03
3	rs59878581		6955158	T	G	0.16	0.63	0.5	0.81	3.02E-04	0.70	0.55	0.89	2.88E-03
3.1	rs76333965	0.44	6957350	G	T	0.08	0.61	0.43	0.87	6.56E-03	0.70	0.50	0.98	4.02E-02
3.2	rs2615771	0.42	6955237	G	A	0.32	0.81	0.67	0.98	2.70E-02	0.84	0.70	1.00	5.38E-02
4	rs11774240		6857380	T	C	0.44	0.73	0.62	0.88	5.71E-04	0.75	0.63	0.88	5.96E-04
4.1	rs35397308	0.94	6857395	G	A	0.44	0.75	0.62	0.89	1.26E-03	0.76	0.64	0.89	1.04E-03
4.2	rs17078234	0.95	6857409	T	A	0.45	0.75	0.63	0.90	1.52E-03	0.75	0.63	0.89	6.89E-04
5	rs4841776		6854697	G	T	0.48	0.74	0.62	0.89	9.06E-04	0.75	0.63	0.88	6.63E-04
5.1	rs2951863	0.97	6852980	C	A	0.48	0.74	0.62	0.89	9.14E-04	0.76	0.64	0.90	1.17E-03
5.2	rs2978887	0.98	6853357	T	C	0.48	0.75	0.63	0.89	1.07E-03	0.76	0.65	0.90	1.23E-03
6	rs2075835		6951366	T	C	0.11	0.62	0.46	0.82	1.12E-03	0.67	0.51	0.89	5.80E-03
6.1	rs6998500	0.90	6950859	A	T	0.12	0.59	0.45	0.79	3.54E-04	0.66	0.50	0.87	3.13E-03
7	rs2702934		6850363	T	C	0.15	1.47	1.15	1.88	2.16E-03	1.38	1.11	1.73	4.40E-03

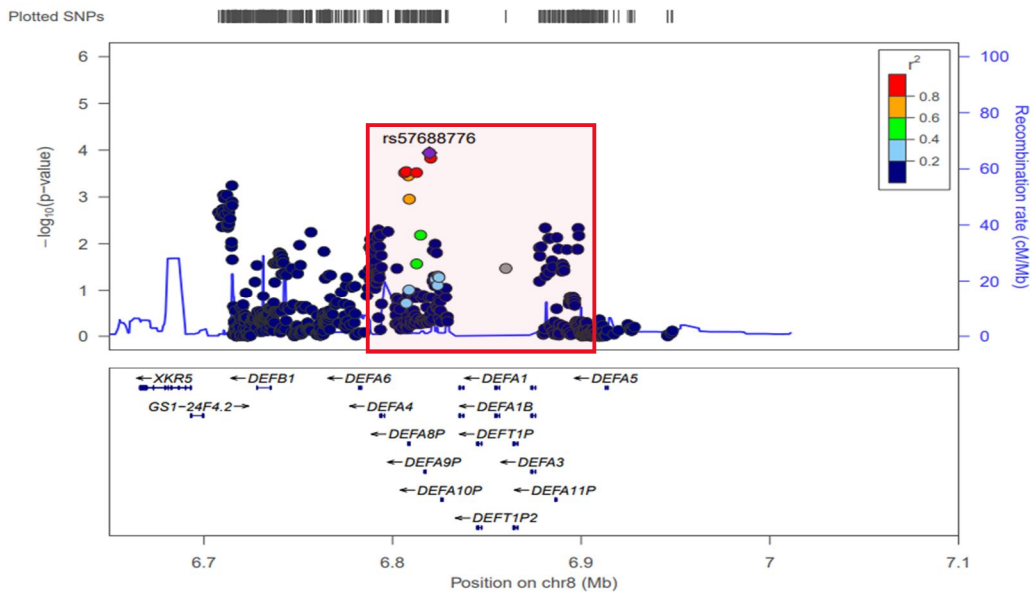


Figure 3.4. Manhattan plot produced on LocusZoom [152]. The figure demonstrates SNPs within the region of chromosome 8 (6.65MB to 7MB) from the GRCh37/hg19 assembly. The top SNPs with the lowest P -values are located at the telomeric (left) flanking region of *DEFA1A3* CNV (highlighted in the red box, corresponding to the gap without any SNPs).

3.4. SNP and AgP association study

The most significantly associated SNP is rs57688776 (P -value: 1.14×10^{-4} ; OR: 0.61 [95% CI: 0.48-0.79]). In Table 3.2, 7 SNPs were included with and without controlling for covariates (*DEFA1A3* copy number, sex, and *DEFA1A3* copy number + sex combined). The P -values for each test showed only small variation, suggesting that the covariates have minimal effect on the level of association to AgP. The odds ratio of all variants is below 1 (except for rs2702934) which suggests that the minor alleles are associated with decreased frequency of the outcome (AgP) and are therefore considered as protective alleles against the disease.

Additionally, we tested the association of rs2978951 and rs2738058 to AgP (Table 3.3), and although the statistical significance of both variants is not as high as the lead SNPs, both variants were identified as risk loci for periodontitis previously and therefore are essential in this study [6]. We also tested the

association of the four tag SNPs identified previously that characterize the different haplotypic classes of *DEFA1A3* [2]. Among the four tag SNPs, rs7826487 was the only variant found to be significantly associated with AgP in the study (P -value: 4.59×10^{-3} ; OR: 0.65 [95% CI: 0.49-0.88]). Interestingly SNP rs2978951 showed no evidence of significant association in the pooled analysis when controlling for *DEFA1A3* copy number and *DEFA1A3* copy number + SEX.

We also performed a separate case-control association analysis based on *DEFA1A3* copy number without SNP genotype information, and we found no significant association of copy number with AgP in the German cases and European 1000 Genomes Project controls (P -value: 0.09; OR: 0.94 [95% CI: 0.87-1.01]) and the pooled analysis (P -value: 0.1174, OR: 0.94 [95% CI: 0.88-1.02]). This suggests that the *DEFA1A3* copy number itself has no direct effect on the progression or development of AgP.

Table 3.2. Case-control association analysis. SNP: single nucleotide polymorphism, BP: base-pair position (chr8), minor: minor allele, major: major allele, MAF: minor allele frequency, ADJ: adjusted, OR: odds ratio, CI: confidence interval, P: p-value, CN: DEFA1A3 copy number, ALL: DEFA1A3 copy number and sex

							AgP-GERx1kGP-EUR				Pooled			
	SNP	BP	minor	major	MAF	ADJ	OR	95% CI		P	OR	95% CI		P
1	rs57688776	6962046	G	A	0.16	-	0.61	0.48	0.79	1.14E-04	0.67	0.53	0.85	9.37E-04
						SEX	0.62	0.48	0.79	1.37E-04	0.67	0.53	0.85	9.94E-04
						CN	0.62	0.49	0.80	2.17E-04	0.68	0.54	0.86	1.56E-03
						ALL	NA	NA	NA	NA	0.68	0.54	0.87	1.68E-03
2	rs7016650	6949736	G	A	0.16	-	0.63	0.49	0.81	2.86E-04	0.70	0.55	0.88	2.61E-03
						SEX	0.63	0.50	0.81	3.38E-04	0.70	0.55	0.88	2.81E-03
						CN	0.64	0.50	0.82	5.00E-04	0.71	0.56	0.90	3.96E-03
						ALL	NA	NA	NA	NA	0.71	0.56	0.90	4.34E-03
3	rs59878581	6955158	T	G	0.16	-	0.63	0.50	0.81	3.02E-04	0.70	0.55	0.89	2.88E-03
						SEX	0.64	0.50	0.82	3.61E-04	0.70	0.55	0.89	3.12E-03
						CN	0.65	0.50	0.83	5.28E-04	0.71	0.56	0.90	4.48E-03
						ALL	NA	NA	NA	NA	0.71	0.56	0.90	4.93E-03
4	rs11774240	6857380	T	C	0.44	-	0.73	0.62	0.88	5.71E-04	0.75	0.63	0.88	5.96E-04
						SEX	0.74	0.62	0.88	8.18E-04	0.75	0.64	0.89	9.20E-04
						CN	0.73	0.61	0.87	5.22E-04	0.74	0.63	0.88	4.96E-04

						ALL	0.74	0.62	0.88	8.29E-04	0.75	0.64	0.89	7.78E-04
5	rs4841776	6857380	T	T	0.48	-	0.74	0.62	0.89	9.06E-04	0.75	0.63	0.88	6.63E-04
						SEX	0.75	0.63	0.89	1.20E-03	0.76	0.64	0.89	1.01E-03
						CN	0.74	0.62	0.88	8.18E-04	0.75	0.63	0.88	6.21E-04
						ALL	0.75	0.63	0.89	1.11E-03	0.76	0.64	0.89	9.66E-04
6	rs2075835	6951366	T	C	0.11	-	0.62	0.46	0.82	1.12E-03	0.67	0.51	0.89	5.80E-03
						SEX	0.62	0.46	0.82	1.15E-03	0.67	0.50	0.89	5.30E-03
						CN	0.63	0.47	0.84	1.72E-03	0.68	0.52	0.91	8.03E-03
						ALL	NA	NA	NA	NA	0.68	0.51	0.90	7.44E-03
7	rs2702934	6850363	T	C	0.15	-	1.47	1.15	1.88	2.16E-03	1.38	1.11	1.73	4.40E-03
						SEX	1.44	1.12	1.84	4.12E-03	1.35	1.08	1.69	7.80E-03
						CN	1.46	1.15	1.87	2.32E-03	1.38	1.10	1.72	4.63E-03
						ALL	1.44	1.13	1.85	3.54E-03	1.35	1.08	1.69	8.30E-03

Table 3.3. Case-control association analysis. *: variants from Munz et al., associated as risk loci for periodontitis disease.

	SNP	r ²	BP	minor	major	MAF	ADJ	OR	95% CI		P	OR	95% CI		P
1	rs2978951*		6965773	A	G	0.42	-	1.24	1.04	1.47	1.57E-02	1.19	1.01	1.4	3.52E-02
							SEX	1.23	1.03	1.46	2.06E-02	1.19	1.01	1.40	4.18E-02
							CN	1.201	0.9964	1.449	0.05452	1.16	0.97	1.38	1.04E-01
							ALL	NA	NA	NA	NA	1.15	0.96	1.37	1.27E-01
2	rs2738058*	0.27	6964095	T	C	0.45	-	1.24	1.05	1.48	1.39E-02	1.23	1.05	1.45	1.21E-02
							SEX	1.24	1.04	1.48	1.57E-02	1.23	1.04	1.45	1.32E-02
							CN	1.22	1.00	1.48	0.0519	1.21	1.01	1.45	4.03E-02
							ALL	NA	NA	NA	NA	1.21	1.00	1.45	4.65E-02
3	rs7826487		7023403	G	A	0.11	-	0.65	0.49	0.88	4.59E-03	0.62	0.46	0.83	1.19E-03
							SEX	0.66	0.49	0.88	5.40E-03	0.62	0.46	0.83	1.41E-03
							CN	0.63	0.47	0.85	2.20E-03	0.59	0.44	0.8	5.39E-04
							ALL	NA	NA	NA	NA	0.60	0.44	0.80	6.28E-04

3.5. Interaction of Top SNPs with *DEFA1A3* copy number

We tested the interaction between the genotype and covariate (i.e., top SNPs and *DEFA1A3* copy number) to determine if together they exert a significantly stronger combined effect on AgP. There is no significant interaction observed for each variant (Table 3.4), confirming that *DEFA1A3* copy number does not affect the phenotype risk when combined with top SNPs as shown on the previous table (except for rs4300027).

Table 3.4. Interaction analysis between top SNPs and DEFA1A3 copy number. *: SNPs are not as highly significant as lead SNPs, but were identified as risk loci for periodontitis by Munz et al. A1: effect allele.

Interaction Analysis			AgP-GERx1kGP-EUR				Pooled			
SNP	BP	A1	OR	95% CI		P	OR	95% CI		P
rs57688776	6962046	G	0.95	0.80	1.12	0.5065	0.97	0.83	1.14	0.716
rs7016650	6949736	G	0.96	0.82	1.13	0.6354	0.98	0.84	1.15	0.8198
rs59878581	6955158	T	0.96	0.82	1.13	0.6206	0.97	0.83	1.14	0.7297
rs11774240	6857380	T	0.96	0.86	1.07	0.4073	0.96	0.87	1.07	0.4571
rs4841776	6854697	G	0.94	0.85	1.05	0.2859	0.94	0.85	1.04	0.2144
rs2075835	6951366	T	0.88	0.73	1.06	0.1855	0.91	0.76	1.09	0.2882
rs2702934	6850363	T	1.06	0.90	1.25	0.4969	1.06	0.92	1.22	0.4436
rs7826487	7023403	G	1.01	0.81	1.26	0.9183	1.04	0.83	1.29	0.7613
rs2978951*	6965773	A	1.08	0.97	1.21	0.1675	1.08	0.98	1.20	0.1287
rs2738058*	6964095	T	1.08	0.96	1.21	0.1842	1.07	0.96	1.19	0.2038
rs4300027	7023053	T	0.87	0.77	0.97	1.48E-02	0.86	0.77	0.96	7.60E-03

3.6. SNP and *DEFA1A3* CNV association study

Next, we asked whether the top SNPs for AgP association showed any significant association with the *DEFA1A3* copy number. There was no evidence of association detected except for SNPs rs2978951 and rs2738058 in both analyses (Table 3.5). Other relevant SNPs showed strong association with *DEFA1A3* copy number and the most significant SNP is rs4300027, which confirms the findings of Khan *et al.* in which *DEFA1A3* copy number was strongly associated with rs4300027 in European samples [86]. However, rs4300027 showed no evidence of association with AgP. The four tag SNPs - that between them classify 5 different *DEFA1A3* haplotypic structures in European populations, namely rs4300027, rs7825750, rs7826487, and rs62487514 - are found at the centromeric border of the *DEFA1A3* CNV. SNP rs7826487 is associated with AgP and interestingly is weakly associated with the *DEFA1A3* copy number (P -value: 0.0358). However, rs7826487 loses that association with *DEFA1A3* copy number in the pooled analysis (Table 3.5).

Previously, SNP rs2738048 was identified as a locus associated with IgAN in Han Chinese patients [153]. Here, in European samples rs2738048 did not show any significant association with AgP (P -value: 0.06; OR: 0.83 [95% CI: 0.69-1.01]). However, it showed a positive significant correlation with the *DEFA1A3* copy number (beta: 0.4523, P -value: 4.64×10^{-9}) (data not shown).

Table 3.5 Association analysis between genotype and *DEFA1A3* copy number. Agp-GERx1kGP-EUR: German cases with AgP vs 1kGP European controls. Pooled: German cases with AgP vs 1kGP European controls plus UK based samples from the blueprint study (see chapter 2, section 2.2).

		AgP-GERx1kGP-EUR		Pooled	
SNP	BP	BETA	P	BETA	P
rs57688776	6962046	0.135	0.1703	0.09025	0.3182
rs7016650	6949736	0.1696	0.08634	0.08956	0.3226
rs59878581	6955158	0.1646	0.09477	0.09624	0.2865
rs11774240	6857380	-0.03789	0.5973	-0.04048	0.5392
rs4841776	6854697	-0.06752	0.3462	-0.02659	0.6875
rs2075835	6951366	0.1434	0.2171	0.007603	0.9438
rs2702934	6850363	-0.09269	0.3506	0.1534	0.09008
rs7826487	7023403	-0.2483	0.0358	-0.1416	0.2025
rs2978951	6965773	-0.3665	1.89E-07	-0.4047	4.75E-10
rs2738058	6964095	-0.5122	3.07E-13	-0.4344	1.76E-11
rs4300027	7023053	0.7924	5.20E-28	0.6947	4.43E-25
rs7825750	7022281	-0.6296	1.50E-16	-0.6002	1.71E-17

3.7. *DEFA3* Quantitative Association Study

Firstly, we wanted to explore whether the genomic proportion of *DEFA3* (i.e., the fraction $DEFA3/(DEFA1+DEFA3)$) was associated with AgP. The genomic copy number of *DEFA1* and *DEFA3* had not been measured explicitly in the AgP samples, and so we used the typical composition of the corresponding *DEFA1A3* haplotypes as a proxy. Although data from HNP1-3 peptides and mRNA are not available in the study, the *DEFA3:DEFA1* peptide ratio might be most relevant to the disease. For this reason, we first wanted to extract the haplotype information from everyone in the study according to the four tag SNPs defining common haplotypes in Europeans (Figure 3.5, and see chapter 1, section 1.4, Figure 1.1) [2]. After acquiring the haplotype information, expected *DEFA3* proportions for each haplotype were calculated by Prof John Armour (using unpublished results from jointly phasing SNPs and CNV content with

MOCsphaser [154]) to be used as a phenotype in the quantitative association analysis.

The strongest signal of association with *DEFA3* proportion came from the Class 2 SNP (rs7825750; P -value: 3.47×10^{-46}), the second strongest signal was from rs2738110 (P -value: 2.562×10^{-29}), which is found at the telomeric region of the *DEFA1A3* repeat (~50.6kb away from Class 2 variant). Both variants are in LD (r^2 : 0.604). The third strongest signal is from rs2977789 (P -value: 9.879×10^{-29} , found at the telomeric region). rs2977789 is in LD with rs2738058 (r^2 : 0.569) and rs2978951 (r^2 : 0.625). This leads to the conclusion that the predicted *DEFA1A3* proportions are most strongly associated with the Class 2-defining variant (rs7825750) in comparison to other variants.

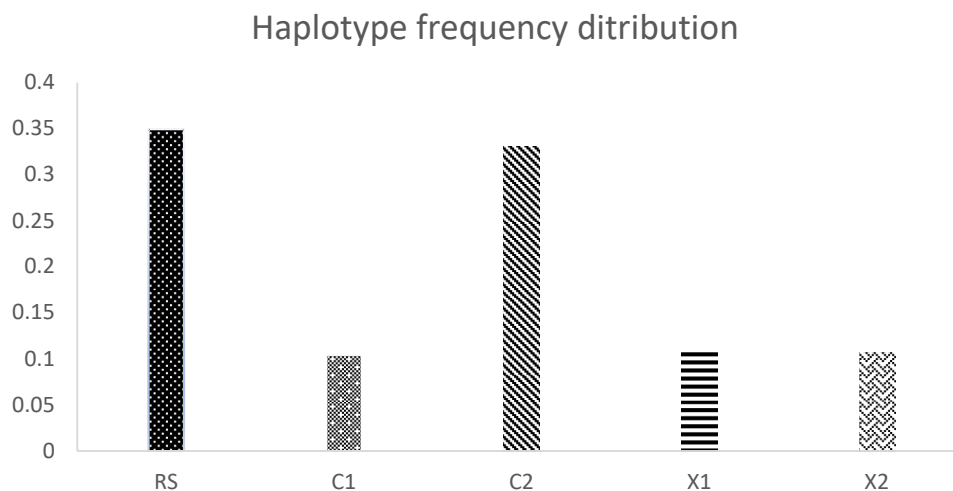


Figure 3.5. *DEFA1A3* haplotype frequency in cases and controls. RS: Reference seq, C1: Class 1, C2: Class 2, X1: Exchange1, X2: Exchange 2.

3.8. Haplotype-based association analysis

Centromeric Haplotype Analysis

Previously, SNP and *DEFA1A3* copy number association analysis was done on German cases and 1kGP controls, to understand the effect of such

components to Aggressive periodontitis (AgP). Next, we asked if a specific multi-SNP haplotype would exert a stronger association with AgP. Firstly, we performed a haplotype-based analysis from the four tag SNPs reported in an earlier study and found centromeric from *DEFA1A3* CNV (Table 3.6).

Table 3.6. The four *DEFA1A3* neighboring tag SNPs that are in linkage disequilibrium.

Class	rs4300027	rs7825750	rs7826487	rs62487514
Ref Seq	T	T	A	C
Class 1	C	T	G	C
Class 2	C	C	A	C
Exchange 1	T	T	A	A
Exchange 2	C	T	A	C

A) German cases & 1000 Genomes Project (European) controls

We started the analysis with German cases and 1000 Genomes Project European controls as that was the original dataset used to perform the SNP and copy number association study. We completed the primary haplotype-based association and the omnibus test by conditional association. An omnibus test is used to examine all haplotypes included in the study collectively for association with the disease or trait. An observed association indicates which haplotype is driving the association

In this step, three centromeric flanking tag SNPs only were included because rs62487514 was filtered out from the dataset in the quality control process. Without SNP rs62487514 in the analysis, it is not possible to distinguish between two haplotypes (Reference from Exchange 1) that are quite different in their expected *DEFA3* proportions. Firstly, the switch used to test the association was `--hap-assoc`:

```
plink      --bfile      input_file_name      --hap-snp
rs4300027,rs7825750,rs7826487      --hap-assoc      --out
output_file_name.
```

Table 3.7. Haplotype association test. SNPs used for the test: rs4300027, rs7825750, rs7826487. FA: frequency of affected (cases), FU: frequency of unaffected (controls), CHISQ: chi-square test, DF: degree of freedom. TCG: Class 1, TCA: Exchange 2, CCA: Class 2, TTA: Reference sequence or Exchange 1.

HAPLOTYPE	FA	FU	CHISQ	DF	P-value
OMNIBUS	NA	NA	17.8	3	0.000483
TCG	0.0793	0.122	10.67	1	0.001088
TTA	0.4635	0.4645	0.001965	1	0.9646
CCA	0.326	0.3223	0.03373	1	0.8543
TCA	0.1312	0.09117	9.217	1	0.002397

B) German cases & controls

Priority to quality control (QC), there were 717 German cases and 4199 German controls (2314 males and 2602 females). After QC, 361 SNPs and 544 individuals were removed from the study. We used the German cases and German controls for haplotype-based analysis to ensure population stratification is eliminated as a factor in the analysis (Table 3.8). Tag SNP rs62487514 was still removed from the dataset after QC and filtering.

Table 3.8. Haplotype association test. SNPs used for the test: rs4300027, rs7825750, rs7826487. FA: frequency of affected (cases), FU: frequency of unaffected (controls), CHISQ: chi-square test, DF: degree of freedom. TCG: Class 1, TCA: Exchange 2, CCA: Class 2, TTA: Reference sequence or Exchange 1.

HAPLOTYPE	FA	FU	CHISQ	DF	P-value
OMNIBUS	NA	NA	19.95	3	0.0001742
TCG	0.08685	0.1033	3.097	1	0.07843
TTA	0.4448	0.4959	10.89	1	0.0009649
CCA	0.3433	0.2932	12.41	1	0.0004279
TCA	0.1251	0.1076	3.27	1	0.07057

As shown on both tables above, there is still a significant association with AgP in the omnibus haplotype association test (German cases and 1kGP controls

P -value: 0.000483; German cases and German controls P -value: 0.0001742). In addition, haplotypes TTA (rs4300027) and CCA (rs7825750/C2) were not significant in Table 3.7 (German cases and 1kGP controls), while rs7826487/C1 showed a significant association with AgP (which confirms what was found in the single SNP association test). However, in Table 3.8 (German cases and German controls), those variants showed opposite outcomes compared to the first table.

Telomeric Haplotype association

A) German cases and 1000 Genomes Project European controls

We performed the omnibus haplotype-based association test on top SNPs that are found in the telomeric region of *DEFA1A3* (i.e., to the left of the CNV as shown in the reference assembly). Five top SNPs were chosen without considering the LD threshold; the SNPs were solely chosen based on the lowest p -values reported. The omnibus association shows that there is a nominally significant association of AgP with haplotypes of rs74556447, rs7016650, rs6998500, rs59878581, and rs57688776 (Table 3.9).

Table 3.9. Haplotype association test. SNPs used for the test: rs74556447, rs7016650, rs6998500, rs59878581, rs57688776. FA: frequency of affected (cases), FU: frequency of unaffected (controls), CHISQ: chi-square test, DF: degree of freedom.

HAPLOTYPE	FA	FU	CHISQ	DF	P
OMNIBUS	NA	NA	9.146	3	0.02741
TGATG	0.08703	0.1237	7.539	1	0.006038
CGATG	0.009978	0.01026	0.004303	1	0.9477
TGTTG	0.03048	0.03894	1.136	1	0.2865
CATGA	0.8725	0.8271	8.588	1	0.003384

B) German cases and German controls

A single-locus association test was performed to find the SNPs with lowest *P*-value, which was rs2951869 (*P*-value: $1.08\text{E} \times 10^{-6}$; OR:1.37 [95% CI: 1.21-1.56]. rs2951869 is in strong LD with rs2977789, which is the second top SNP in the analysis. rs2951869 is in LD with six other variants, which also includes SNP rs2978951 (risk locus for periodontitis from Munz *et al.* study). We then performed a telomeric haplotype-based test on the SNP with the lowest *p*-value. We kept adding a new SNP in LD with rs2951869 to test the strength of association as shown on the tables below consecutively (Table 3.10). The omnibus association shows that there is a statistically significant association with AgP. In the third table, the omnibus association is weaker compared to the second table after the addition of rs2977788.

Table 3.10. Haplotype association test. FA: frequency of affected (cases), FU: frequency of unaffected (controls), CHISQ: chi-square test, DF: degree of freedom.

Haplotype	FA	FU	Chisq	DF	P	SNPs
GT	0.3565	0.2889	22.46	1	2.15-06	rs2951869
AC	0.6435	0.7111	22.46	1	2.15E-06	rs2977789

Haplotype	FA	FU	Chisq	DF	P	SNPs
OMNIBUS	NA	NA	25.23	2	3.32E-06	rs2951869
GTA	0.3564	0.2886	22.72	1	1.88E-06	rs2977789
ACA	0.116	0.1117	0.1912	1	0.662	rs2978951
ACG	0.5275	0.5997	22.21	1	2.45E-06	

Haplotype	FA	FU	Chisq	DF	P	SNPs
OMNIBUS	NA	NA	25.32	3	1.32E-05	rs2951869
GTAA	0.3564	0.2887	22.59	1	2.01E-06	rs2977789
ACAA	0.1161	0.1118	0.1912	1	0.6619	rs2977788
ACAG	0.01083	0.01424	0.8846	1	0.347	rs2978951

ACCG	0.5167	0.5852	19.86	1	8.34E-06
------	--------	--------	-------	---	----------

In conclusion, we were unable to identify an association between *DEFA1A3* copy number and AgP using the same subjects as analysed by Munz *et al.* to discover SNP associations at *DEFA1A3*. However, we were still able to confirm their primary observation that telomeric neighbouring SNPs found in the flanking region of *DEFA1A3* are associated with the disease.

CHAPTER 4

4. Association of *DEFA1A3* copy number with gene expression in human immune cells

This study aims to investigate the association between *DEFA1A3* copy number and RNA levels in white blood cell subsets. In addition, this work explored the association between *DEFA1:DEFA3* genomic ratio to its relevant transcript ratio. The analysis used sequence data from 197 samples from a UK based population from the Blueprint Epigenome Study (see chapter 2, section 2.2). A multistep purification procedure was used to obtain pure selections of neutrophils, naive T-cells and classical monocytes in the Blueprint Epigenome Study. The subsets were purified using an antibody/magnetic bead approach. The purity was 98% for neutrophils, 93% for naive T-cells and 95% for monocytes.

4.1. Total *DEFA1A3* copy number measurement

Read depth measurement

Read depth measurement on 197 genomic DNA samples was achieved using the methods outlined in (Chapter 2, section 2.2). Each sample has 3 or 4 genomic bam files, and we first calculated the estimated copy number for each bam file separately (Figure 4.1A). We then averaged the copy number estimate per

individual (Figure 4.1B) by aggregating results from that individual's bam files. The copy number estimates obtained from DEFA1core and DEFA2core+2 (two independent approaches to *DEFA1A3* copy number estimation by read-depth, see Chapter 2, section 2.1) correlated well with each other (Figure 4.1A $r^2 = 0.8382$; Figure 4.1B $r^2 = 0.9504$, P-value: $3.9E-129$). The definite clustering around integer values was prominent from 4 to 8 copies. After 8 copies the clustering seemed to gradually become less distinct as the copy number increased (Figure 4.1B, Figure 4.2). The total *DEFA1A3* copy number ranged from 4 to 11 copies per diploid genome (Figure 4.3). The highest frequency in the population is for 8 copies, and the median is 7 copies.

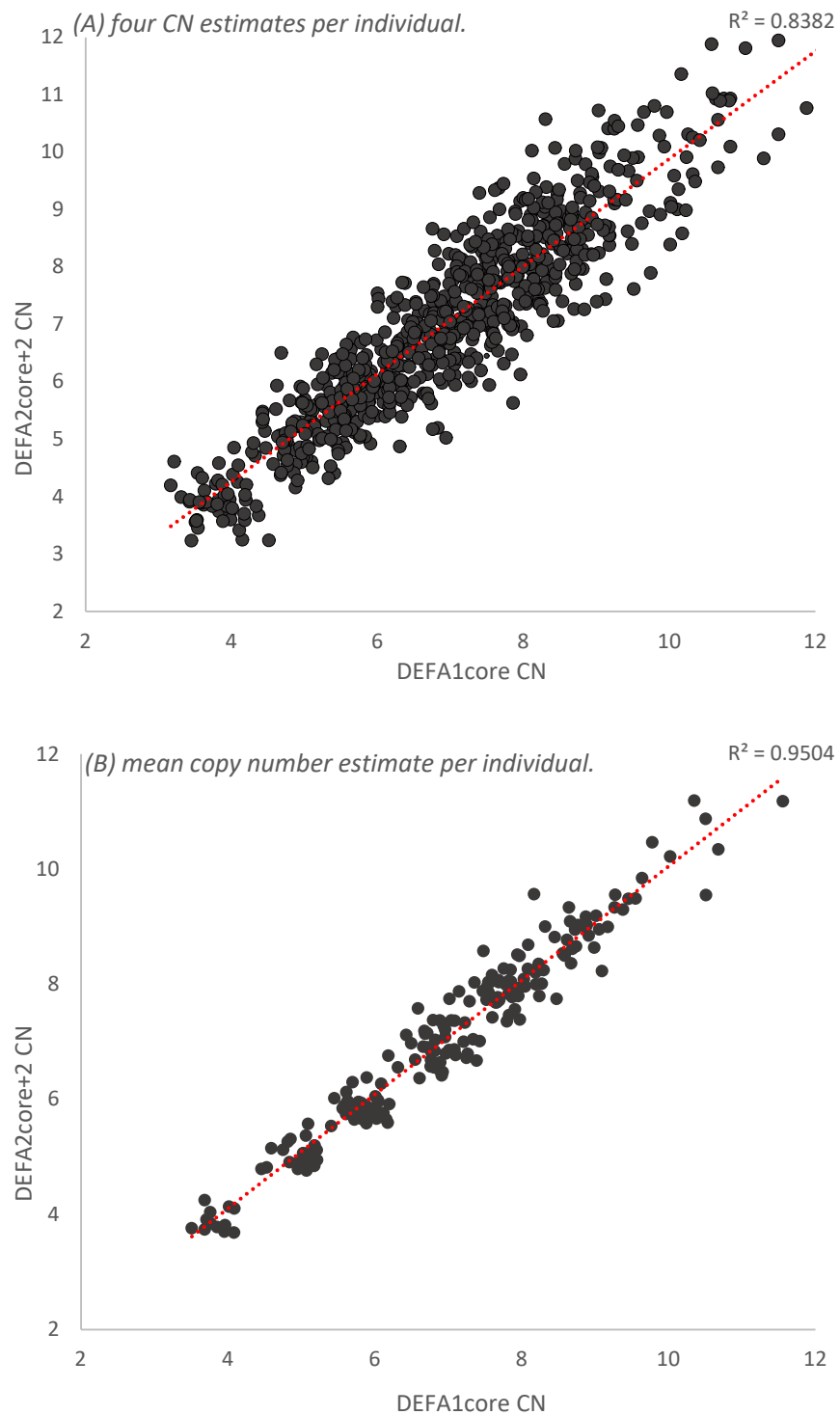


Figure 4.1. *DEFA1A3* copy number estimates. (A) up to four copy number measurements obtained per individual, (B) average copy number measurement per individual.

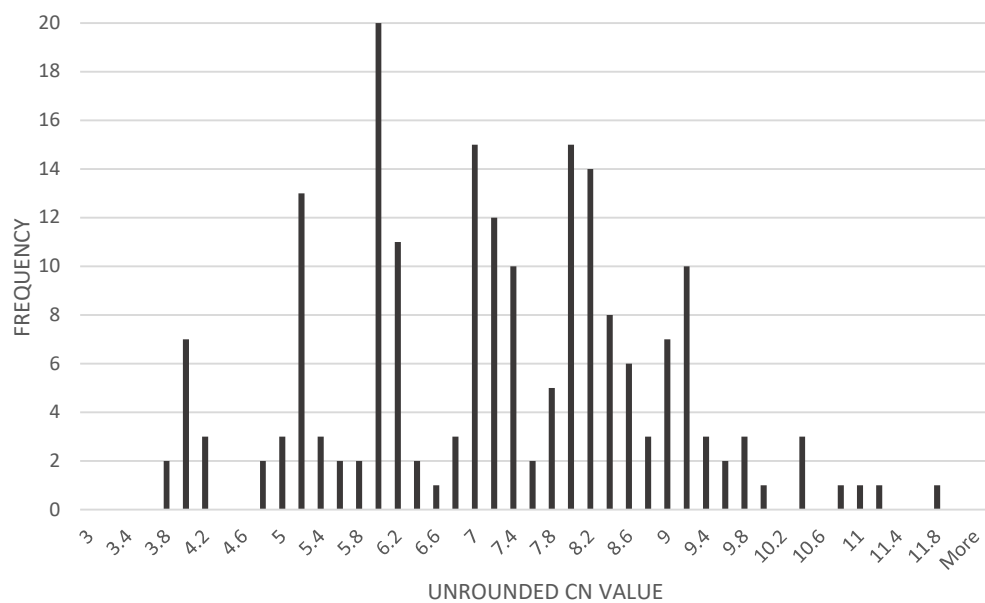


Figure 4.2. Unrounded *DEFA1A3* copy number distribution around each integer value

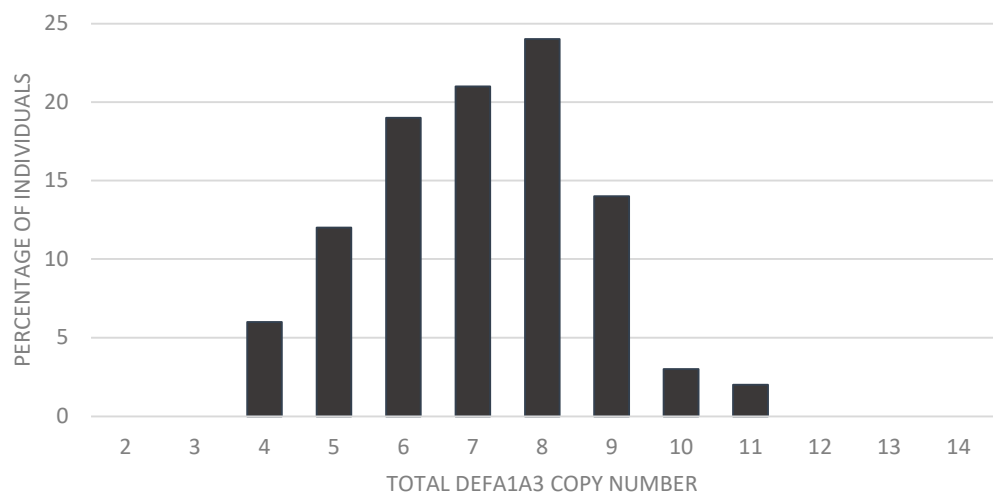


Figure 4.3. *DEFA1A3* copy number distribution in the UK (United Kingdom) population (197 samples).

Accuracy of *DEFA1A3* copy number measurement

The mean copy number estimate was calculated from measurements of DEFA1core and DEFA2core+2 on 197 samples. Next, the obtained mean values were normalized relative to their nearest integer value (Figure 4.4). The normalized ratios shown here have a mean value of 1.00159 (median: 1.00077) and a standard deviation of 0.03013.

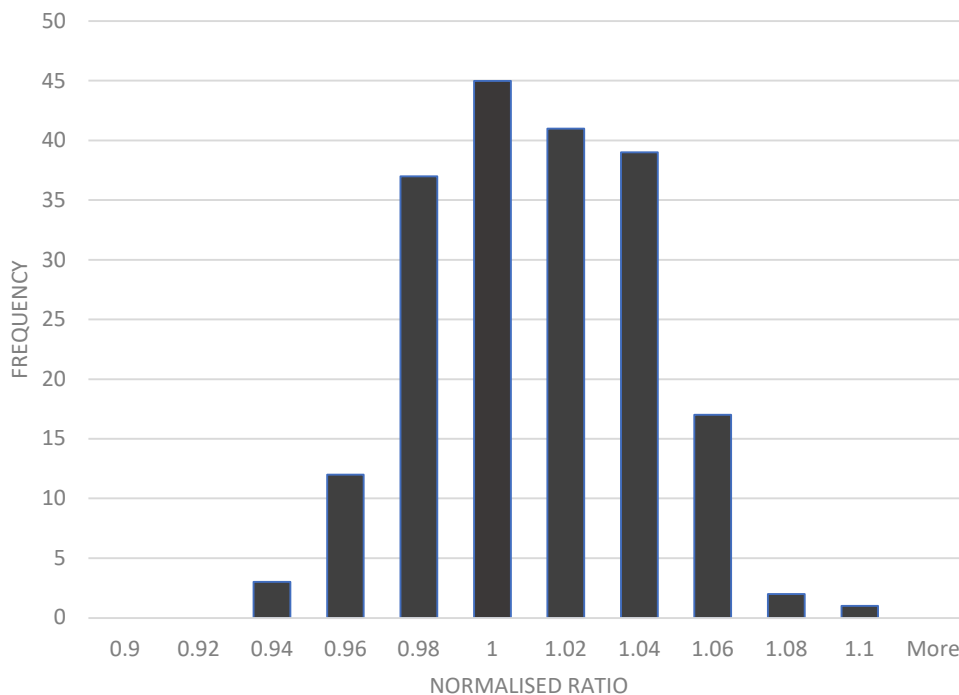


Figure 4.4. Distribution of normalized values (based on the mean of the two independent *DEFA1A3* measurements) in 197 individuals.

4.2. RNA-Seq data analysis

205 bam files of mapped RNAseq reads from 196 mature neutrophil (CD16+) samples were downloaded. To normalise for sequencing depth and gene length, the total mapped reads from each sample were counted, and the scaling factor (sf) per million was calculated by dividing the total mapped reads by 10^6 . The read count from *DEFA1A3* was divided by sf which normalises for sequencing

depth, and the normalised sequencing depth is divided by the length of the transcript in kb. We also measured read depth for a selection of genes known to be preferentially expressed in neutrophils such as *NCF1* (neutrophil cytosolic factor 1), *ELANE* (elastase, neutrophil expressed), and *CEBPE* (CCAAT enhancer binding protein epsilon) for comparison with *DEFA1A3*. *NCF1* exhibited the highest levels of expression relative to other genes, approximately 9x greater than *DEFA1A3*. Additionally, we investigated the correlation of *DEFA1A3* with *NCF1*, *ELANE* and *CEBPE* transcript levels and we found no evidence of a linear association; however, there was a significant but weak correlation found between *DEFA1A3* and *ELANE* (*NCF1*: $r^2=0.0367$, $P\text{-value}=0.6096$; *ELANE*: $r^2=0.2697$, $P\text{-value}=1.32\text{E-}04$, *CEBPE* $r^2=0.0565$, $P\text{-value}=0.4315$).

212 bam files from 172 T-cell (CD4+) RNA samples were downloaded. 40 samples had 2 files, for which the average was calculated. 53% of T-cell samples showed no *DEFA1A3* reads at all. Known T-cell specific genes were also measured to compare the levels of transcript levels with *DEFA1A3*, such as *CD4* and *CCL5* (chemokine CC motif ligand 5). *CD4* expression levels are ~3000 and 15 times higher than *DEFA1A3* and *CCL5*, respectively. We also investigated the correlation of *DEFA1A3* with *CD4* and *CCL5* transcript levels and we found no evidence of linear association, *DEFA1A3* vs *CD4* ($r^2=0.0062$, $P\text{-value}=0.9357$) and *DEFA1A3* vs *CCL5* ($r^2=0.0022$, $P\text{-value}=0.9772$).

197 bam files from classical monocyte (CD14+CD16-) samples were downloaded. The analysis showed no evidence of *DEFA1A3* gene expression in monocytes from any of the reads/files. We then compared the levels of RNA expression from each human immune cell, which reveals that the mean levels of *DEFA1A3* RNA in neutrophils are approximately 150 times those of T cells and no expression in monocytes was detected.

4.3. *DEFA1A3* CNV association with gene expression

Although *DEFA1A3* expression levels are high in neutrophils, no association was detected between *DEFA1A3* total genomic copy number and levels of expression in neutrophils ($r^2 = 0.000379$, P -value= 0.78658). Similarly, no association was detected in RNA from T cells ($r^2 = 0.000899$, P -value= 0.699584) (Figure 4.5).

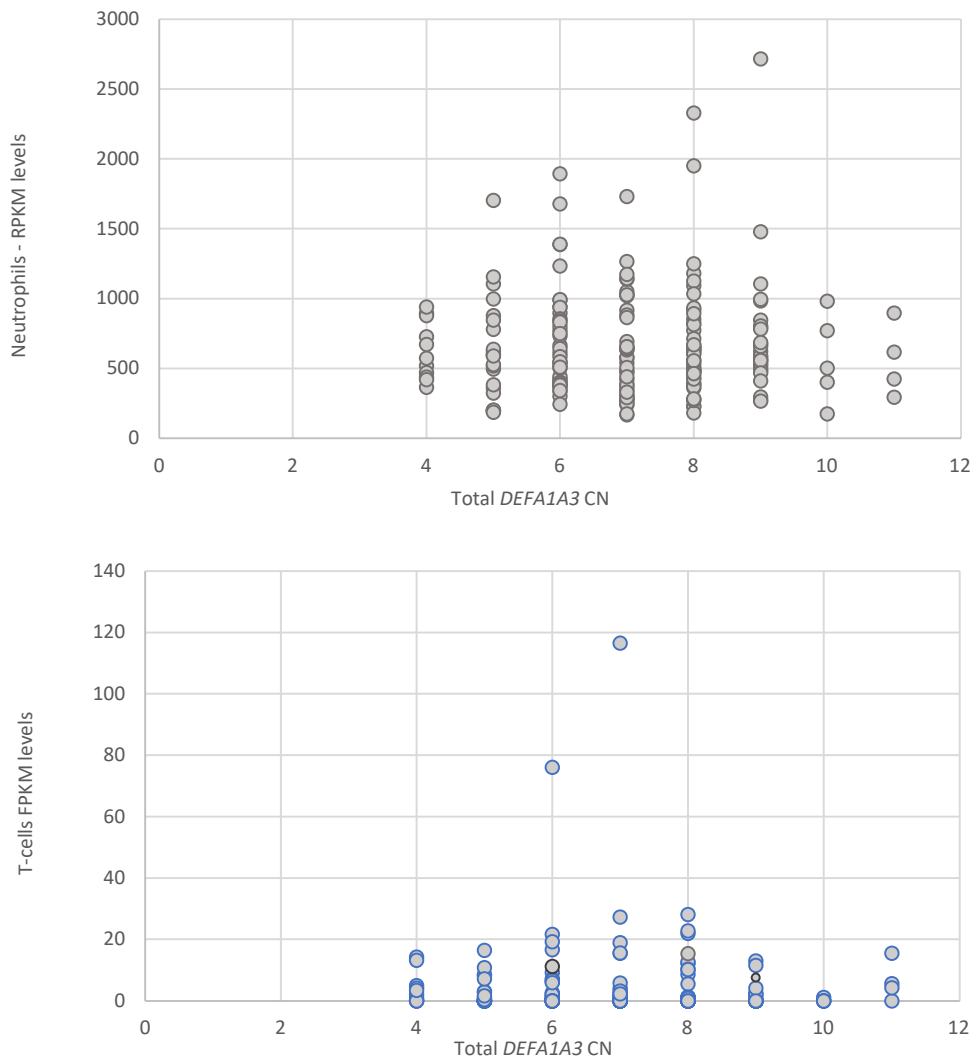


Figure 4.5. Association of total *DEFA1A3* copy number with gene expression in Neutrophils and T-cells. Note the very different scales on the y-axis for the two analyses.

4.4. Association of Neutrophils and T-cells' levels of transcripts

We tested the level of association between *DEFA1A3* transcript levels in Neutrophils and T-cells to check if there is evidence of parallel activation of expression or any cross-contamination of cell preparations [143]. There is no evidence of a significant correlation between both transcript levels ($r^2=0.001968$, $P\text{-value} = 0.5679$).

4.5. Association of relative ratios of DEFA1:DEFA3 RNA levels and equivalent gene numbers.

Here, we estimated the *DEFA1* and *DEFA3* proportion in the genomic DNA and RNA in neutrophils to test the association between relative *DEFA1:DEFA3* genomic ratios and their corresponding expression levels. We used two approaches to obtain such information. Prior to using the methods, we re-aligned the genomic and transcript reads to an artificial reference sequence (hg19:chr8:6790000-6850000) from which sequence had been omitted to leave a single *DEFA1A3* repeat.

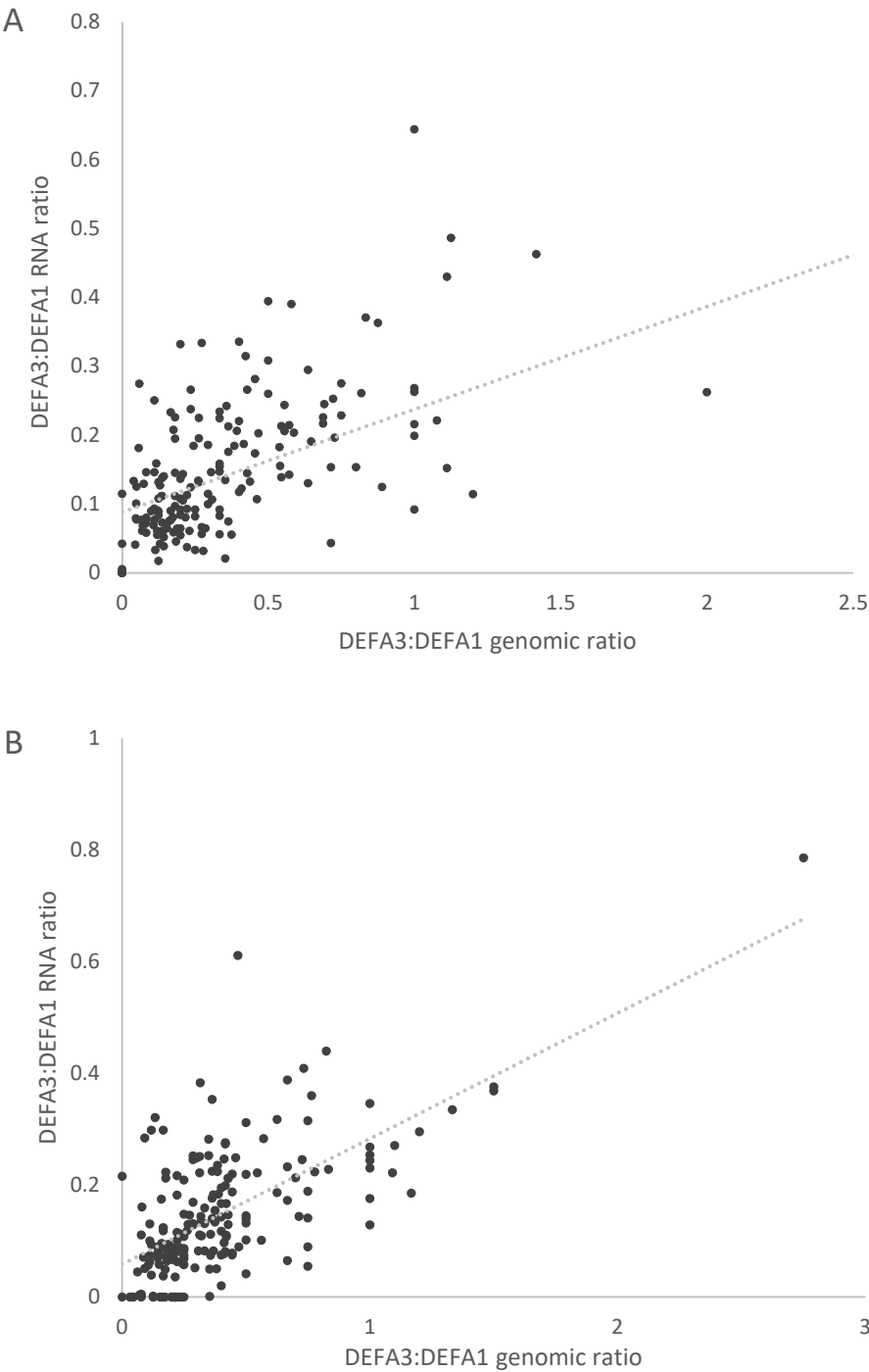
The first approach to test the association was to use a command syntax (`samtools+awk`) created by Masella *et al.* (see chapter 2, section 2.2) that filters out reads of interests based on a single nucleotide position in a sequence (see chapter 2, section 2.2, Figure 2.2). This approach is useful to sort reads from the *DEFA1(G)* allele and *DEFA3(T)* allele, and thereby count each allele to obtain the ratio.

The second approach is using the `bcftools mpileup` with allele depth information to obtain the values of *DEFA1(G)* and *DEFA3(T)* variants.

Additionally, we included two common expressed variants found within the *DEFA1A3* gene (chr8:6835194 – found on the third exon and chr8:6836028 – found on the second exon). In the RNA-seq files, using the `samtools+awk` method, the allele sorting was accurate as it separated the reads with *DEFA1(G)* alleles from *DEFA3(T)* alleles. While in the genomic files, we observed a mixture of both alleles even after sorting. This was also the case for the two common expressed variants. Table 4.1 summarises the correlation coefficient for both approaches to test the association of relative ratios of *DEFA1:DEFA3* RNA levels to corresponding gene numbers (Figure 4.6). Both methods indicated that there was a statistically significant association, although the correlation co-efficient ranged between 0.03 and 0.78. The strongest correlation was observed in the 6836028C/G variant whereas the weakest correlation was in the 6835194(G/A), which indicates that 62% and 10.5% of the variation in RNA levels are attributable directly to the genome copy number ratios, respectively. Additionally, 9.7% and 11.2% of individuals lack the *DEFA3* gene using the `samtools+awk` and `bcftools allele depth` method, respectively.

Table 4.1. Statistical correlation analysis for two methods used to test genomic *DEFA1:DEFA3* association to *DEFA1:DEFA3* transcripts. AD: allele depth

Variants	<code>samtools+awk</code> (R; <i>P</i> -value)	<code>Bcftools mpileup+AD</code> (R; <i>P</i> -value)
chr8:6835383(G/T)	0.641; <0.0001	0.565; <0.0001
chr8:6835194(G/A)	0.032; <0.0001	0.454; <0.0001
chr8:6836028(C/G)	0.406; <0.0001	0.789; <0.0001



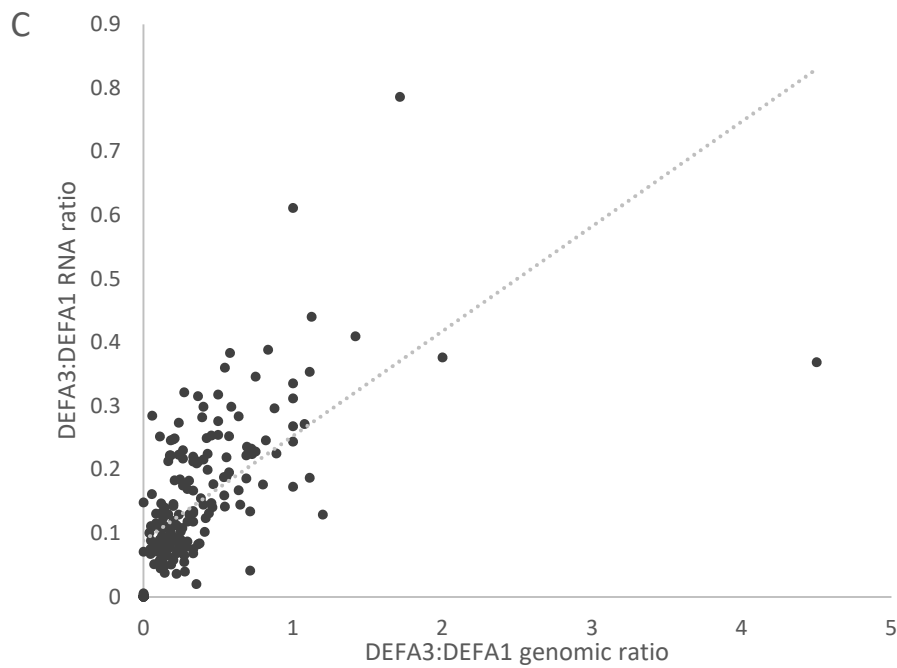


Figure 4.6. *DEFA3:DEFA1* ratio. A) Using *bcftools mpileup* + *AD* syntax to obtain the allele depth for *DEFA1* and *DEFA3* alleles; B) *samtools+awk* syntax command was used. C) Hybrid approach – *bcftools mpileup+AD* on x-axis for genomic and *samtools+awk* on y-axis for transcripts.

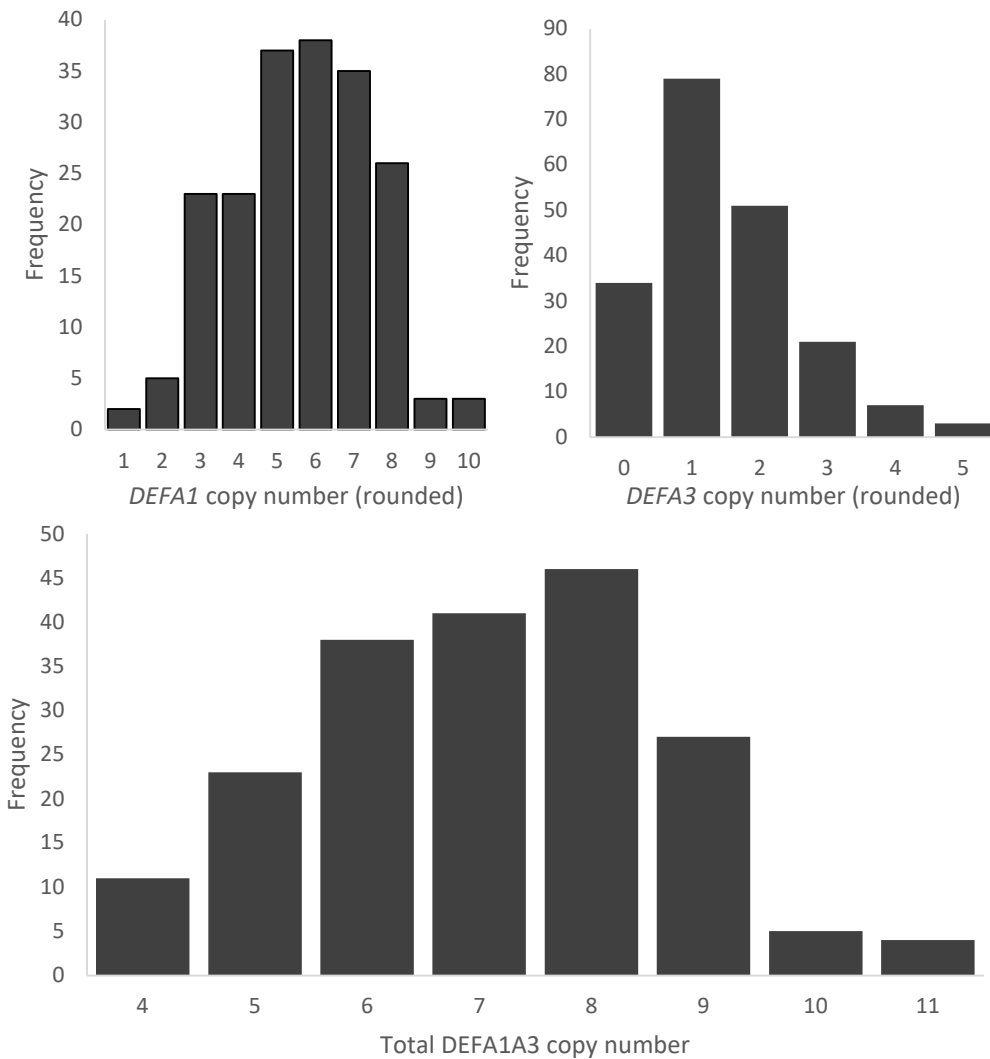


Figure 4.7. *DEFA1*, *DEFA3*, and total *DEFA1A3* copy number deduced from *bcftools mpileup* + *AD*.

In summary, the total copy number in this work ranged between 4 and 11 *DEFA1A3* copies, in which the copy number measurements were attained from the read-depth approach. We found that total *DEFA1A3* copy number was not associated with levels of transcripts in human immune cells (neutrophils, T-cells and monocytes). In addition, there was an association between *DEFA1:DEFA3* gene copy number ratio with equivalent RNA levels. This implies that other factors affect variation in the overall level of activation *DEFA1A3* expression. However once activated, all gene copies in the array are expressed approximately equally.

CHAPTER 5

5. Amplicon Nanopore Sequencing

We evaluated the efficacy of Oxford Nanopore Technologies here to reconstruct complex repeated haplotype structures. Two different devices were used (Flongle and MinION) for a long amplicon sequencing approach, which entails a tiling path technique. The aim here was to sequence long PCR products to include five amplicons of 5.5 - 7.5 kb long per *DEFA1A3* repeat with each amplicon consisting of overlapping regions to one another. As input for tiling path sequencing from *DEFA1A3* haplotypes, amplicons were generated with annealing temperature at 58 °C for PCR1, PCR2, PCR3 and PCR5 and at 64 °C for PCR4. After this the amplicons were subjected to nanopore sequencing using a flongle (2x) and a MinION sequencer (see Chapter 2, section 2.3 from Materials and Methods for more detail on samples used per experiment, the modified version of the protocol, and the principle behind the 5 amplicons per sample).

5.1. Flongle Amplicon Runs

Deducing the *DEFA1A3* copy number for samples NA12878 and NA12760 from the 1000 Genomes Project (see chapter 2, section 2.1) predicts that each sample should have 5 copies of the *DEFA1A3* region, as also determined in published work [2, 92] (data not shown). Here, two flongle experiments were conducted. The primary run involved pooling two samples (5 amplicons per sample) simultaneously, while the second flongle experiment comprised of sequencing each sample independently (information summarised on Tables 5.1 and 5.2). Each run lasted for 24 hours, and the real-time data analysis was displayed on the minotour webpage developed by DeepSeq [155]. The mean read lengths in both

runs varied between 2,182 bp and 3,163.2 bp. In addition, the mean quality of the reads for the former experiment was less than 6.2, while the latter was 11.3 and 11.4. It is recommended to only keep reads for downstream analysis with a minimum threshold of quality score above 7 [156]. Between the two experiments, it has resulted in producing varying numbers of reads. The secondary run generated 195.93 and 354.63 times more reads than the primary run for NA12878 and NA12760, respectively (Table 5.2). Similarly, the total yield in the second run was 135.15 and 312.8 times higher than the first one.

Table 5.1. Summary of the flongle runs. PCR1 primers: we used different primers for each run (see Chapter 2, section 2.3 from Materials and Methods). New PCR1 primers include essential centromeric tag SNPS found on DEFA1A3 haplotypes that are useful to classify five haplotypic structures in a European population [2].*

	Primary Run	Secondary Run
Pooled samples	Yes	No
Barcoding	Yes	No
No. of flongles used	1	2
PCR1 primers*	NanoPCR1_old	NanoPCR1_new
Duration of run hours	24	24
MinKNOW	3.0.13	20.10.6
MinKNOW core	3.1.8	4.1.2
Bream	3.1.0.17	6.1.4
Guppy	2.0.5	4.2.3

Table 5.2. Summary of sequencing statistics for samples NA12878 and NA12760 in the primary and secondary runs.

	NA12878		NA12760	
	1 st Run	2 nd Run	1 st Run	2 nd Run
Mean Read Length	3,163.2	2,182	2,759.2	2,433
Mean Read Quality	6.2	11.3	6.1	11.4
Number of Reads	1,611	315,645	1,265	448,617
N50	4,134	6,367	4,342	6,407
GC content (%)	45.16	46.99	44.94	47.02
Total Yield (bp)	5,095,947	688,735,837	3,490,370	1,091,466,999

Table 5.3 represents the number of reads derived per PCR bam file for each sample. The method used here was to split the original raw bam file into 5 unique PCR bam files corresponding to their bp-coordinates on the hg19 assembly (see Chapter 2, section 2.3 for the 5 PCR bp-coordinates). The design of the long amplicon experiment is that each PCR corresponds to a region in the *DEFA1A3* repeat – starting with PCR4 and then ending with PCR3 per repeat (except for the centromeric repeats that start with PCR4 and end with PCR1) (Check chapter 2, section 2.3, Figure 2.5). Moreover, there are overlapping regions between PCRs that are essential to support the characterization of unique repeats from one another, to assist in constructing the repeats together, and to identify which haplotype they belong to in the genome. The fold increase representation in the second phase for sample NA12878 ranged between 26.50-84.71, while it ranged between 57.33-167.78 for NA12760 (Table 5.3).

Table 5.3. Number of reads reported per PCR bam file per sample. Fold change between the primary and secondary run is reported for comparison.

Number of reads	NA12878			NA12760		
	1 st run	2 nd run	Fold Δ	1 st run	2 nd run	Fold Δ
PCR1	247	10143	41.06	109	18288	167.78
PCR2	176	10487	59.59	193	19444	100.75
PCR3	221	10174	46.04	196	24870	126.89
PCR4	265	22447	84.71	306	21340	69.74
PCR5	481	12748	26.50	341	19548	57.33

First Amplicon Experiment

NanoPCR1_old primers were used in the production of long amplicons for this experiment. These primers do not include the amplification of the four centromeric tag SNPs that classify the five *DEFA1A3* haplotype structures [2].

The pooled flongle run lasted for 24 hours and produced 10,844 raw reads with a yield of 26.76 Mb, of which 20.52% passed the quality threshold. The average read length was 2,459 bp (read lengths ranged from 2 bp to 14,309 bp). 1611 and 1265 raw sequences were generated for NA12878 and NA12760, respectively (Table 5.2). Minimap2 [148] was used to align the reads to an artificial reference sequence that consists of only one telomeric *DEFA1A3* repeat (hg19:chr8:6800001-6851705) – 1504 reads were mapped for NA12878 and 1019 for NA12760. The length of reads ranged from 500 bp to 7000 bp, and the expected read length for *DEFA1A3* amplicons is between 5.5 kb to 7.5 kb long.

Regarding sample NA12878, the highest number of reads (9.97%) had a read length of ~4.5 kb, while the lowest frequency (1.8%) was from reads that were ~7 kb. Similarly, sample NA12760 showed that 9.72% of reads were at ~6 kb, and 2.65% were at ~7 kb (Figure 5.1). Total full length reads of the expected amplicons for samples NA12878 and NA12760 were 386 (25.7%) and 267 reads (26.2%) only.

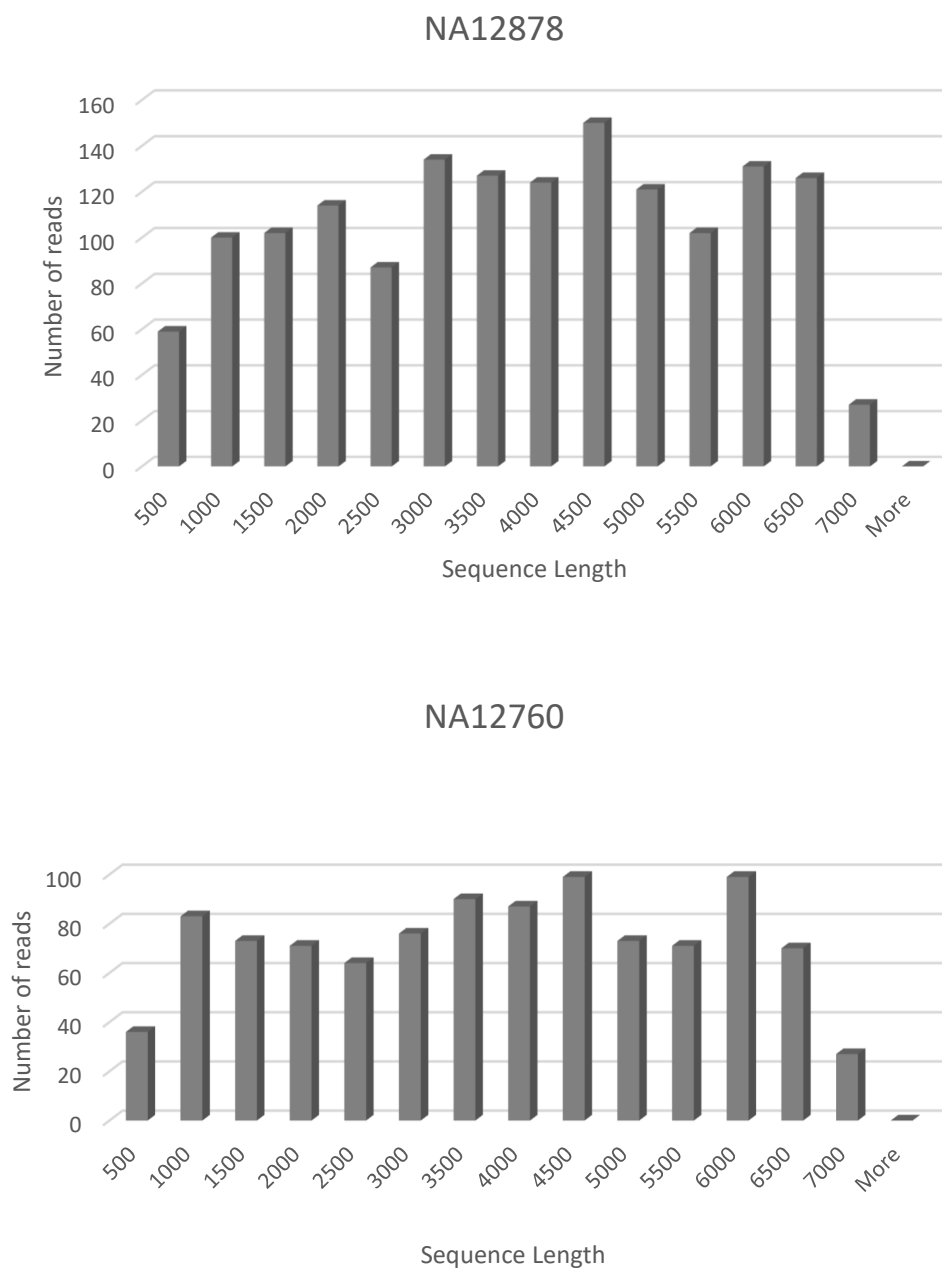


Figure 5.1. Histogram of read distribution for NA12878 (top) and NA12760 (bottom).

Repeat Separation Method

Prior to separating the repeats based on variable SNP positions, we evaluated the ratio of *DEFA1:DEFA3* from PCR1 and PCR5 as both amplicons

contain the *DEFA1*(G)/*DEFA3*(T) variant. All reads covering the base in question were used, therefore we have not selected full length reads due to the low coverage and low quality of reads. With prior knowledge, samples NA12878 and NA12760 have a deduced total *DEFA1A3* copies of 5 [2]. Similarly, from our read depth measurement using the 1000 Genomes Project sequencing (see Chapter 3, section 3.1), both samples should have 5 copies each. In addition, the *DEFA1:DEFA3* ratio was inferred previously [92] (NA12878 was 4:1 and NA12760 was 3:2).

Table 5.4. *DEFA1:DEFA3* ratio was determined by using the information of *DEFA1*(G) and *DEFA3*(T) alleles in the nanopore reads.

	NA12878		NA12760	
	PCR1	PCR5	PCR1	PCR5
<i>DEFA1:DEFA3</i> ratio	3.8:1	3.7:1	1.3:1	1.58:1
No. of reads with <i>DEFA1/DEFA3</i> variants	158	167	63	120

Firstly, we attempted to exploit AssociVar [157], a recently developed tool, using both the CRISPR cas9 and the primary amplicon sequencing data. It is a software that uses statistical associations between variant bases to distinguish between true variants and sequencing artifacts/errors, and was developed primarily to call variants in mixed populations of viral sequences. Because ONT sequencing suffers from high error rates, this could in principle be a useful tool to identify true mutations and infer haplotypes correctly. The tool depends on the assumption that true variants will demonstrate systematic patterns of correlation with other alleles, while sequencing artifacts will not display any associations. Following the protocol stated, no output was generated during the indexing of the artificial reference sequence. The steps are sequential and therefore the output is needed from each stage to follow through. It could be that AssociVar cannot handle a large reference sequence that is >10000 bp; our artificial reference sequence is approximately 53 kb for the amplicon sequencing and 400 kb for CRISPR-Cas9. The previous study that used AssociVar included a reference sequence of 1284 bp only [157].

Secondly, `samtools+awk` [146] was then used for the repeat separation on the basis of allelic status at specific positions (Figure 5.2). The syntax filters out reads based on the status of a single nucleotide (see Chapter 2 section 2.2 and Figure 2.2 for more information). The sorting of reads was successful, but completion of the tiling path was unsuccessful to construct the repeats based on the overlaps between PCRs, because we lack the most essential information in PCR1, the centromeric tag SNPs. Therefore, this work was not continued further, and the next steps used PCR1 primers that could anchor the centromeric flanking regions for each allele.

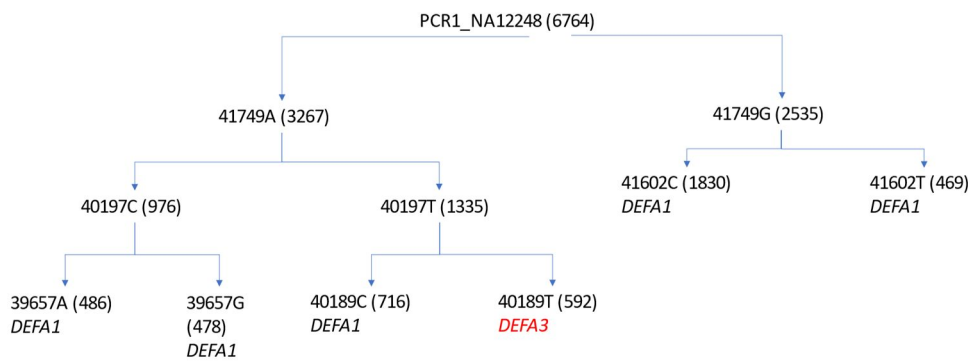


Figure 5.2. An example of repeat sorting example from the MinION amplicon experiment for sample NA12248 from PCR1 data. Six repeats from PCR1 were separated. The numbers in the figure represent the number of reads per step.

Second Amplicon Experiment

NanoPCR1_new primers were used in this experiment, these primers extend to approximately 1 kb more than the previous primers, so that they include the centromeric tag SNPs [2]. Two flongle experiments were processed. Each run lasted for 24 hours and it produced 367.36 K and 527.82 K raw reads (88% and 86.12% bases passed the quality threshold) for samples NA12878 and NA12760,

respectively. The mean read lengths were 2,182 bp and 2,433 bp with an N50 of 6,367 bp and 6,467 bp for NA12878 and NA12760, respectively (Table 5.2, Figure 5.3). The number of estimated bases for NA12878 was 877.29 Mb and 1.54 Gb for NA12760. In addition, the failed bases were 93.88 Mb (12%) for NA12878 and 174.07 Mb (13.88%) for NA12760. The minimum and maximum read lengths for each sample are as follows: a) NA12878: 69 bp and 11,851 bp, b) NA12760: 46 bp and 8985 bp. Minimap2 was also used here for alignment (chr8:6800001-6851705) – 65783 reads were mapped for NA12878 and 102916 for NA12760.

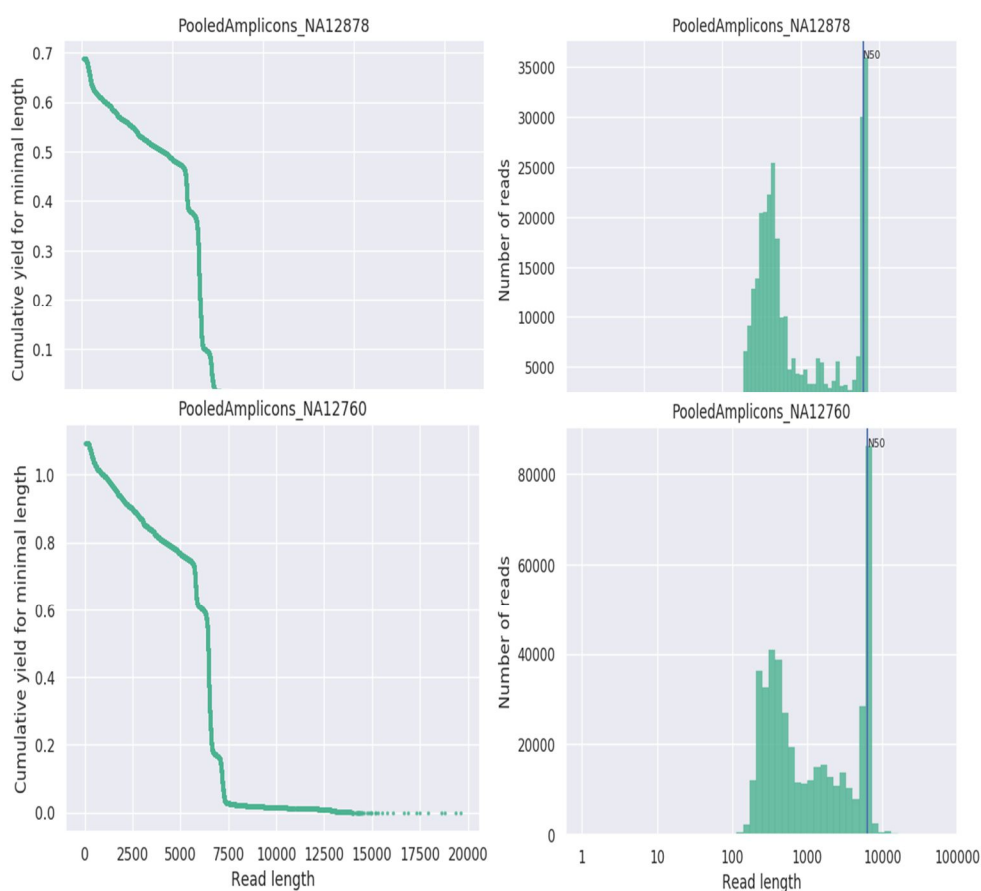


Figure 5.3. Raw read lengths plotted against the number of reads produced from the flongle sequencing run. Figures were created from a NanoPlot [1] python script.

Repeat Separation Method

Prior to the repeat sorting step, the relevant ratios were established to get an insight into roughly how many copies of *DEFA1A3* CNV we would predict from the run (Table 5.5). Firstly, the ratio of *DEFA1(G):DEFA3(T)* variants was determined from the reads acquired in PCR1 and PCR5 bam files. Between PCR1 and PCR5 is an overlapping region of ~1,520 bp long which includes the *DEFA1/DEFA3* variant and other informative SNPs which will be mentioned later in this section. Secondly, the ratio of the centromeric and non-centromeric (cent:non-cent) alleles was deduced from PCR1. For clarification, centromeric reads will always involve exactly two *DEFA1A3* repeats only, while the non-centromeric reads are the remainder that are not exclusively found at the centromeric end. Similarly, the telomeric and non-telomeric (tel:non-tel) ratio was also determined from PCR4. As shown on Table 5.5, the *DEFA1:DEFA3* ratio from PCRs 1 and 5 show approximately similar values, which demonstrates we were able to detect the copies of *DEFA1* and *DEFA3* from both samples.

Table 5.5. Observed ratios for a) *DEFA1(G):DEFA3(T)* alleles found in PCR1 and PCR5 reads, b) centromeric and non-centromeric alleles from PCR1 reads, and c) telomeric and non-telomeric alleles from PCR4 reads.

	NA12878	NA12760
<i>DEFA1:DEFA3</i> ratio in PCR1	4.1:1	1.59:1
<i>DEFA1:DEFA3</i> ratio in PCR5	4.2:1	1.78:1
Cent:Non-Cent ratio	4.74:1	2.2:1
Tel:Non-Tel Ratio	1.38:1	1.33:1

The `samtools+awk` command (see chapter 2, section 2.2 and Figure 2.2) was used to sort repeats according to the variant of interest. We firstly investigated the variants called by a variant-calling pipeline including `bcftools`. We then selected a variant to be the primary source of separation. The selection depended on several factors (in order): a) number of reads that encompass the variant of

interest, b) allele depth of the reference allele and the alternative allele, c) the quality of the variant (meaning that any variant that has more than one alternative allele must be excluded from the selection process). After the primary sorting, we continued the same process until we acquired the number of estimated repeats per sample from the established *DEFA1(G):DEFA3(T)* ratio. Tyson *et al.* identified 8 phased gene-specific SNPs that were used to construct sequence specific information per repeat within a haplotype for samples NA12878 and NA12760. Table 5.6 represents those 8 variants and are used for comparison with the previous findings from Tyson *et al.* variants matched what was found previously, except for those that are marked with a (^) symbol. The (*) represents a variant that was called by a variant calling tool (in this case, bcftools mpileup [147] command), and these variants were reported as a mixed position. Nevertheless, we reported in this situation the most common variant by evaluating the SNP information and the allele depth of the reference and alternative alleles.

Table 5.6. Eight phased SNPs from Tyson et al. study that were determined from emulsion haplotype fusion PCR. (*): variants originally called by bcftools as a mixed position, and the more frequent allele was reported here. (^): the variants reported here differ from the variants reported in the Tyson et al. study. Variant 35383G/T distinguishes *DEFA1*(G) from *DEFA3*(T). tel: telomeric repeat, mid: middle repeat, cent: centromeric repeat.

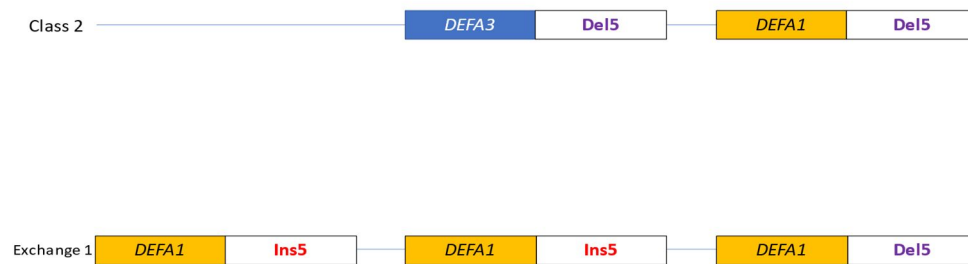
NA12878		3-copy			2-copy	
POS (hg19:chr8)	variant	tel	mid	cent	Tel	Cent
35383	G/T	G	G	G	T	G
35488	A/C	C	C	C	A	A*
35506	T/C	C	C*	C	T	T
35514	G/T	G	G*	G	G	T
35639	A/T	A	A	A	T	A
35645	T/G	T^	T^	T^	T	T
35780	C/T	T	T	T	T	T^
35821	G/T	G	G	G	G	G

NA12760		3-copy			2-copy	
POS (hg19:chr8)	variant	tel	mid	cent	tel	Cent
35383	G/T	G	G	G	T	T
35488	A/C	A*	A	C	A	A
35506	T/C	T	T	C*	T	T
35514	G/T	G	G	G	G*	T
35639	A/T	A	A	A	T	T
35645	T/G	G*	T	G*	T	T
35780	C/T	C	C	C	T	T
35821	G/T	T	T	T	G	G

Following the tiling path method - in which we connect each PCR to one another based on the SNP information within the overlaps – the repeat structures were characterized (Figure 5.4). As shown on the figure below, full reconstructed haplotypic sequences of NA12878 and NA12760 were achieved. Sample NA12760 is a RefSeq/Class1 sample that consists of three *DEFA1* copies and two *DEFA3* copies found within the same haplotype, respectively. Del5/Ins5 is an indel of 5bp long (ACAGA) found upstream of the *DEFA1/A3* gene before exon 1.

Additionally, sample NA12878 is a Class2/Exchange1, with 4 copies of *DEFA1* and 1 copy of *DEFA3*. In Figure 5.5 and Table 5.6, we report the number of reads output following the repeat separation from `samtools+awk` command. In both samples, repeats number 4 and 5 do not have PCR2 and PCR3 because both repeats are absent from the centromeric regions of the *DEFA1A3* CNV. Centromeric repeats only include (in order): PCR4, PCR5, and PCR1, which is a partial *DEFA1A3* repeat of approximately 10kb in length. By contrast, a full *DEFA1A3* repeat is 19 kb long and in this amplicon experiment all of the amplicons are included (in order): PCR4, PCR5, PCR1, PCR2, and PCR3.

NA12878



NA12760

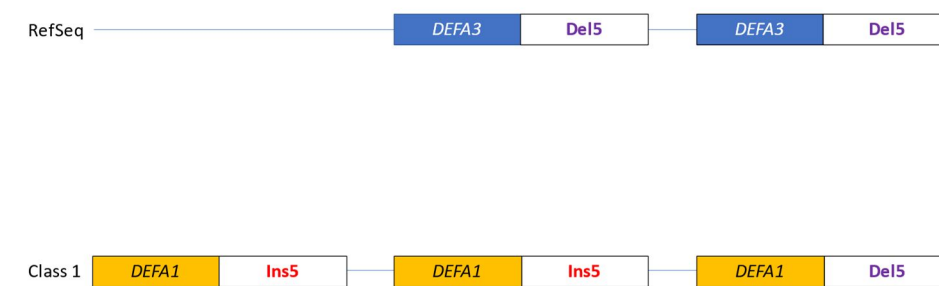


Figure 5.4. The repeat structures characterized by nanopore sequencing at the *DEFA1A3* locus following the tiling path of the PCRs to their corresponding repeats based on SNP information a) between PCR overlapping regions, b) centromeric/non-centromeric from PCR1, and c) telomeric/non-telomeric from PCR4. NA12878 (top) and NA12760 (bottom). Del5/Ins5: 5bp indel variant (ACAGA).



Figure 5.5. The figures here demonstrate the frequency of the reads outputted per DEFA1A3 repeat post-repeat separation via the samtools+awk command. NA12878 (top) and NA12760 (bottom) have 5 DEFA1A3 repeats each.

Table 5.7. Summary of number of mapped reads per repeat per PCR bam file. Repeats number 4 and 5 represent the centromeric repeats. These repeats do not include PCRs 2 and 3 which span the DEFT1P pseudogene that is only found within the telomeric and middle regions of any given DEFA1A3 haplotype.

Repeats	PCR4	PCR5	PCR1	PCR2	PCR3
1	1390	2399	1975	3866	447
2	627	730	2449	510	3318
3	1308	1080	963	1598	1789
4	2223	899	422	-	-
5	2428	1729	535	-	-

5.2. MinION Amplicon Run

NanoPCR1_new primers were used for the MinION experiment (see Chapter 2 – Materials and Methods, section 2.3); these primers extend approximately 1 kb further and include the centromeric tag SNPs [2]. One MinION device was utilized which included one pooled sample of amplicons from 5 barcoded samples (NA06993, NA07056, NA07347, NA12248, NA12249). The run lasted for 72 hours and produced 2,227,243 raw reads in total with a yield of 8,746,216,187 bases, of which 78.57% passed the quality threshold. The average read length was 3,927 bp, while the read lengths overall ranged from 9 bp to 163,112 bp with an estimated N50 of 6.52 kb (Table 5.7).

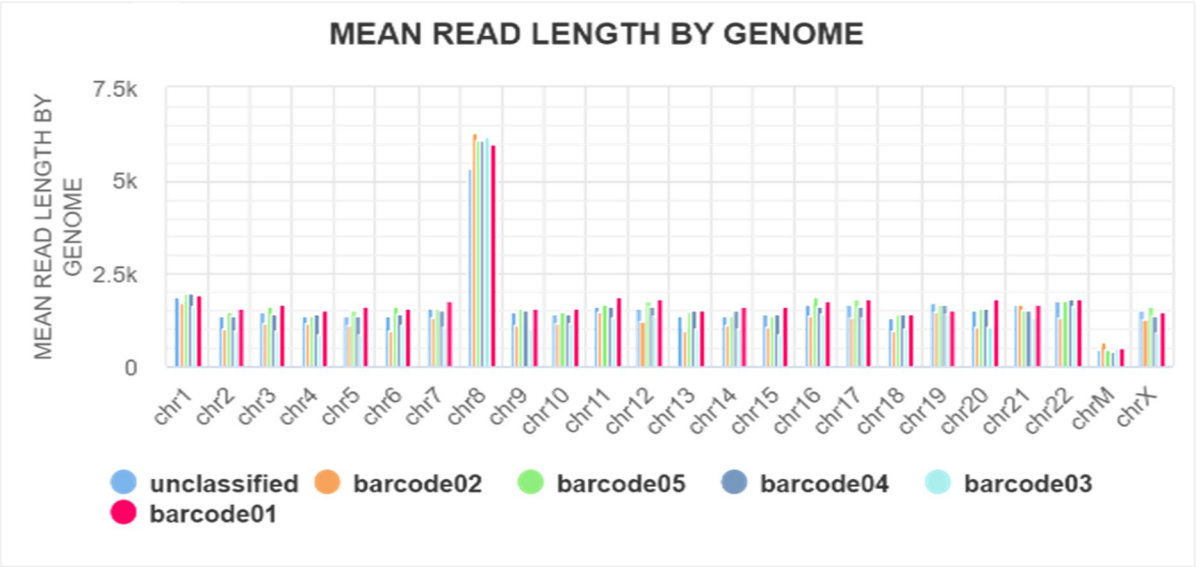
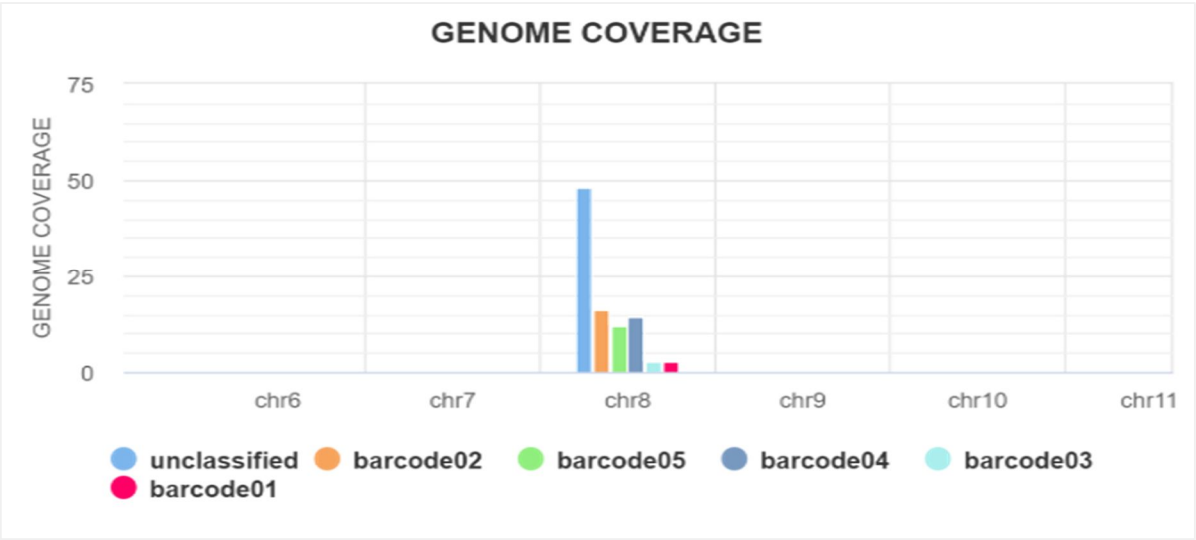
Table 5.8. Summary of sequencing statistics for one pooled sample made up of 5 barcoded samples. Samples sequenced in this study are from Northern and Western European ancestry (CEPH; Utah residents). Min: minimum, Max: maximum, No.: number, hrs: hours, M: million, Gb: gigabases, bp: base pair.

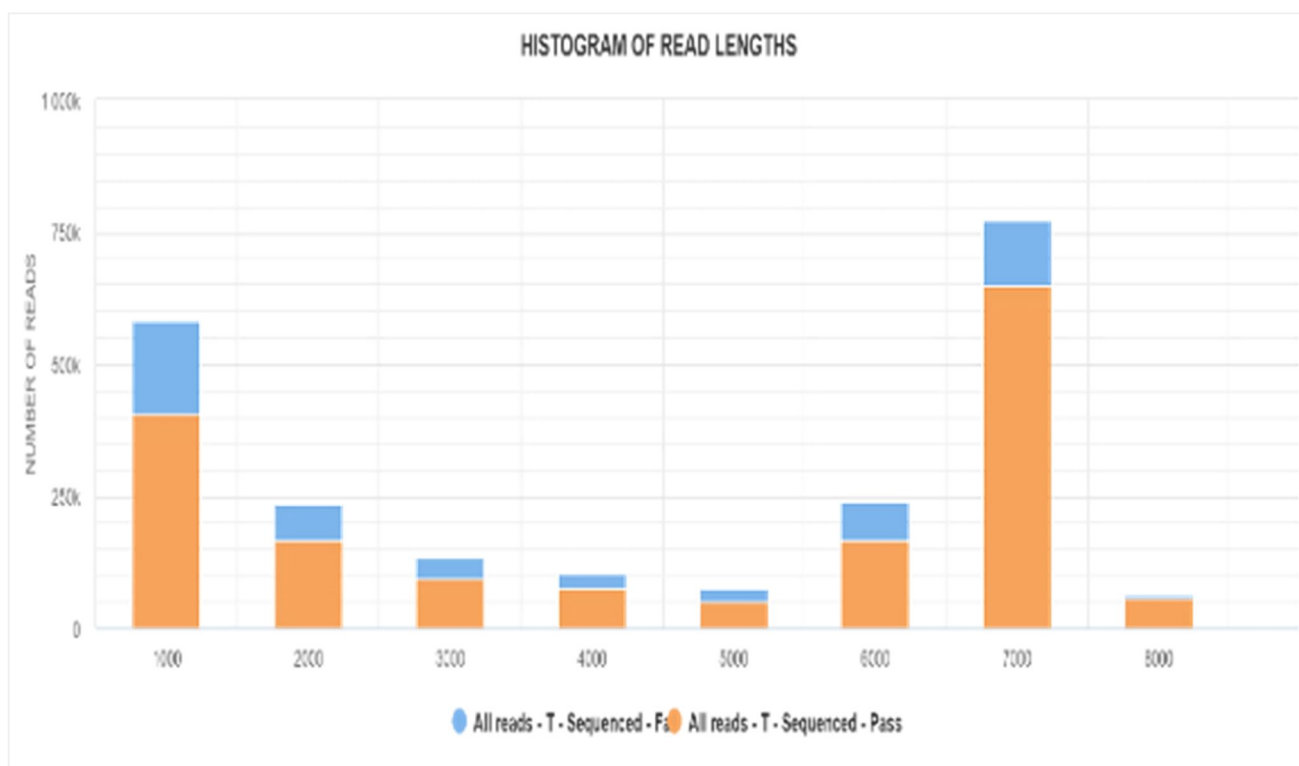
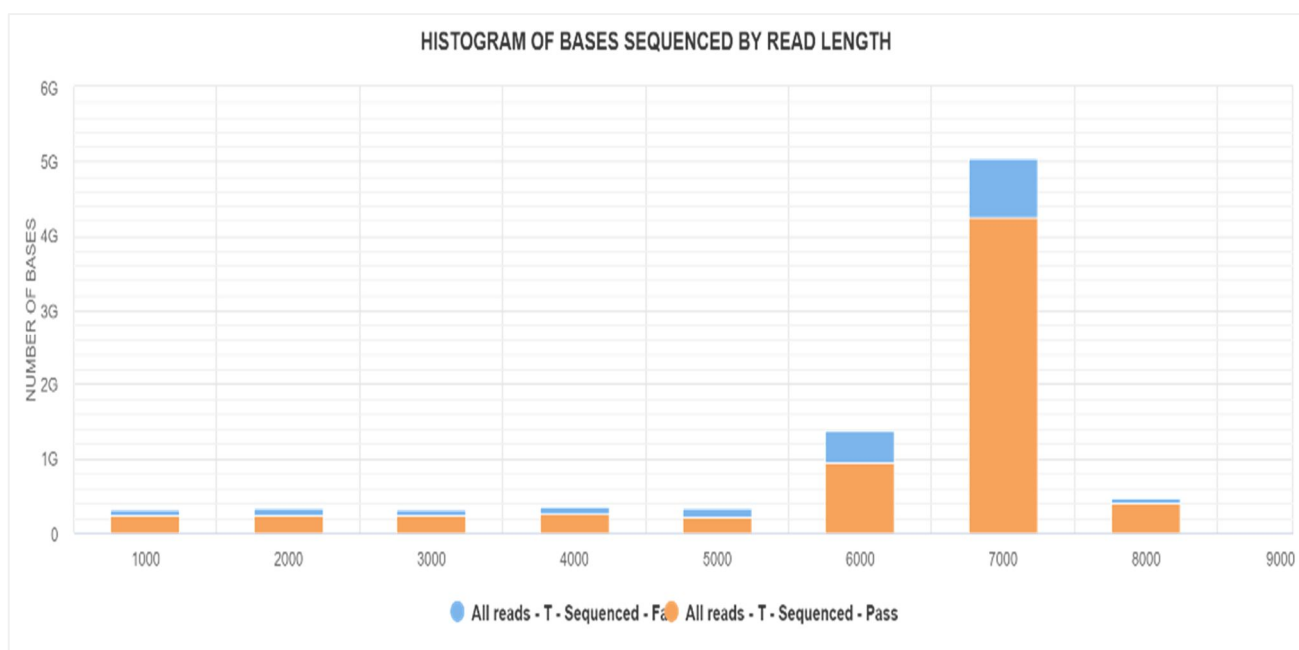
		Passed	Failed
No. of samples	5	-	-
Run time (hrs)	72	-	-
Reads generated (M)	2.23	1.67	0.56
Estimated bases (Gb)	9.11	6.87 (78.57%)	1.87 (21.43%)
Min read length (bp)	9	20	9
Average read length (bp)	3,927	4,119	3,354
Max read length (bp)	163,112	24,695	163,112
Average quality	11.3	12.9	6.8
MinKNOW		21.05.25	
MinKNOW core		4.3.12	
Bream		6.2.6	
Guppy		5.0.16	

The reads were grouped in to two categories: classified reads and unclassified reads. The classification of reads distinguishes them as either reads barcoded from one end only - which is the unclassified group - or reads barcoded from both ends, known as the classified or the barcoded group. The proportion of unclassified reads was 55%. During downstream analysis, the unclassified reads were not considered for repeat sorting. In addition, the classified group were then separated by barcode to produce individual bam files per barcoded sample for further analysis.

The mean read length mapped to chromosome 8 is approximately 6500 bp (the range of expected sizes for full-length amplicons is 5773bp - 7244bp) as shown on Figure 5.6A. The average genome coverage of reads over the full length

of chromosome 8 was between 5X and 15X for the barcoded samples (from the classified category), and approximately 50X for the unclassified reads (Figure 5.6B). The read lengths vary from ≤ 100 bp to ≥ 8000 bp. The most frequent read length was approximately between 6500 and 7000 bp long with a total number of more than 750,000 reads. The second most frequent read length category is below 1000 bp and the number of reads is more than 500,000 reads. However, the full-length reads of interest in this study are between 5000 bp and 7000 bp (Figure 5.6C and 5.6D). Figure 5.6E represents the total percentages of reads produced per barcoded sample, and the highest percentages are from barcoded samples 2, 4 and 5 (12.76%, 12.15%, and 12.01% respectively that include both the failed and passed reads).





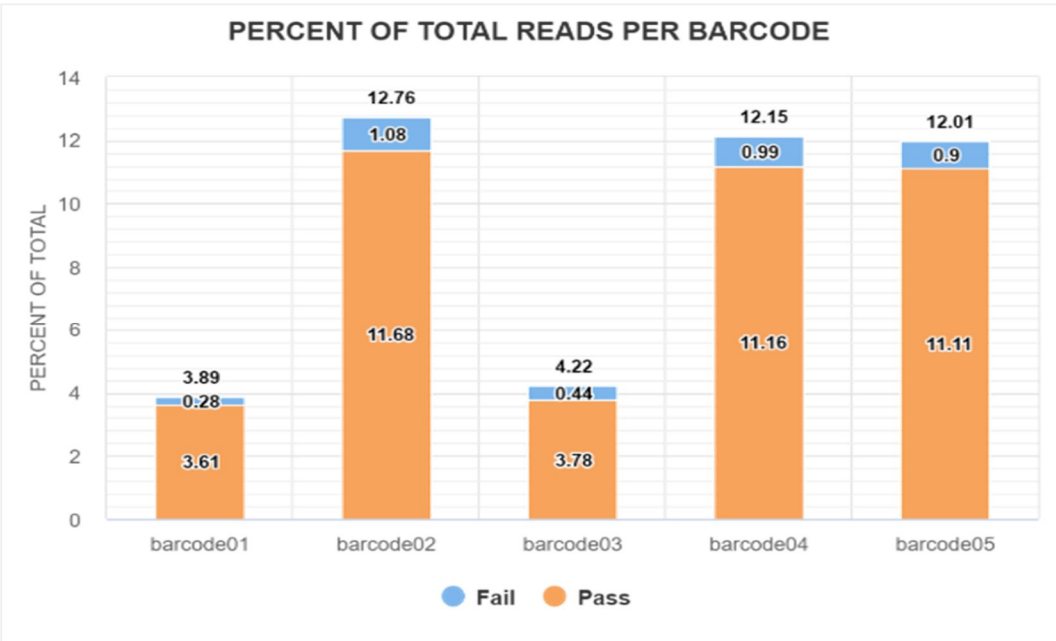


Figure 5.6. Summary of the 72-hour MinION experiment. Figures were obtained from the Minotour DeepSeq webpage (<http://nanoseq.nottingham.ac.uk:8000/login/>). A) Mean read length by genome mapped to chromosomes 1-22, X and M (human mitochondrial chromosome). B) Zoomed-in genome coverage focused on chromosome 8 from 5 barcoded samples and the unclassified group. C) Histogram of total number of reads by read lengths and D) total number of bases by read lengths. E) Percentage of total reads sequenced per barcoded sample. The orange and blue bars (found in Figures C, D, and E) represent reads that passed and failed the quality threshold respectively. Barcode 1: NA06993, Barcode 2: NA7056, Barcode 3: NA07347, Barcode 4: NA12248, Barcode 5: NA12249.

Table 5.8 shows an overview of the data documented from each barcoded sample. The mean read quality remains constant between samples and the lowest number of reads recorded were from samples NA06993 and NA07347 compared to other samples. The mean read length between samples ranged from 2,734.3 bp to 4,831.9 bp. Additionally, the N50 ranged from 6,537 bp to 6,612 bp. Total number of bases was also lowest from samples NA06993 (barcode 1) and NA07347 (barcode 3) compared to other samples. Subsequently, reads from each barcoded sample bam file were split into five different bam files based on their more precise chromosomal locations (see chapter 2, section 2.3 and Figure 2.5 for each PCR chromosomal location). Table 5.9 summarizes the number of reads produced per PCR bam file (PCRs 1, 2, 3, 4, and 5) from each sample. PCR4 has more reads than others and PCR1 has very few reads for NA07347 and NA12249.

Table 5.9. Summary statistics per barcoded sample. BC: barcoded sample.

	NA06993 (BC-01)	NA07056 (BC-02)	NA07347 (BC-03)	NA12248 (BC-04)	NA12249 (BC-05)
Mean read length (bp)	3,548.3	4,831.9	2,734.3	4,560.6	4,075.7
Mean read quality	13	13	13	13	13
GC content (%)	47.29	46.56	47.02	46.95	46.87
No. of reads	80,357	260,231	84,127	248,649	247,379
Read length N50 (bp)	6,537	6,612	6,553	6,547	6,545
Total bases (bp)	285,129,696	1,257,410,635	230,026,475	1,133,996,430	1,008,243,125

Table 5.10. Number of mapped reads per PCR bam file per sample.

	NA06993	NA07056	NA07347	NA12248	NA12249
PCR1	3209	13386	790	6764	462
PCR2	8315	17652	6861	36071	26049
PCR3	4719	20935	2454	25009	12809
PCR4	25018	116146	19253	65920	85835
PCR5	2546	14911	2210	19587	16553

NA06993

NA06993 is a male CEPH/CEU sample with 8 *DEFA1A3* copies and has a Class 1/RefSeq haplotype structure [2, 86] (see Chapter 1, section 1.4, Figure 1.1 for the different *DEFA1A3* haplotype classes). The *DEFA1(G):DEFA3(T)* ratio was determined from PCR1 and PCR5 reads as 8.88:1 and 11.93:1, respectively, suggesting that there is a single copy of *DEFA3*. However, the ratios from PCR1 and PCR5 reads are not consistent with one another, and indicate a total copy

number greater than 8. The centromeric to non-centromeric ratio (cent:non-cent) was also determined from PCR1 and it is 1.51:1, implying a total copy number of 5, and the PCR4 telomeric to non-telomeric ratio (tel:non-tel) is 2.91:1, suggesting a copy number of 8 (or other multiples of 4). There are three possible *DEFA1A3* haplotypic structures consistent with commonly observed patterns for class 1 and reference haplotypes [2] and they are presented on Figure 5.7. The most likely structure for an individual of 8 copies may either be a 6 + 2, 5 + 3, or 4 + 4 *DEFA1A3* copy structure, but no definite inference was achieved because the repeat sorting was incomplete for PCR3 and PCR5 reads due to lack of candidate SNP positions, and therefore tiling the amplicons to build the true haplotype structure was not achieved. However, it was still possible to partially tile non-centromeric PCR1 repeats with PCR2 repeats from unique position-specific variants from the overlaps (data not shown).

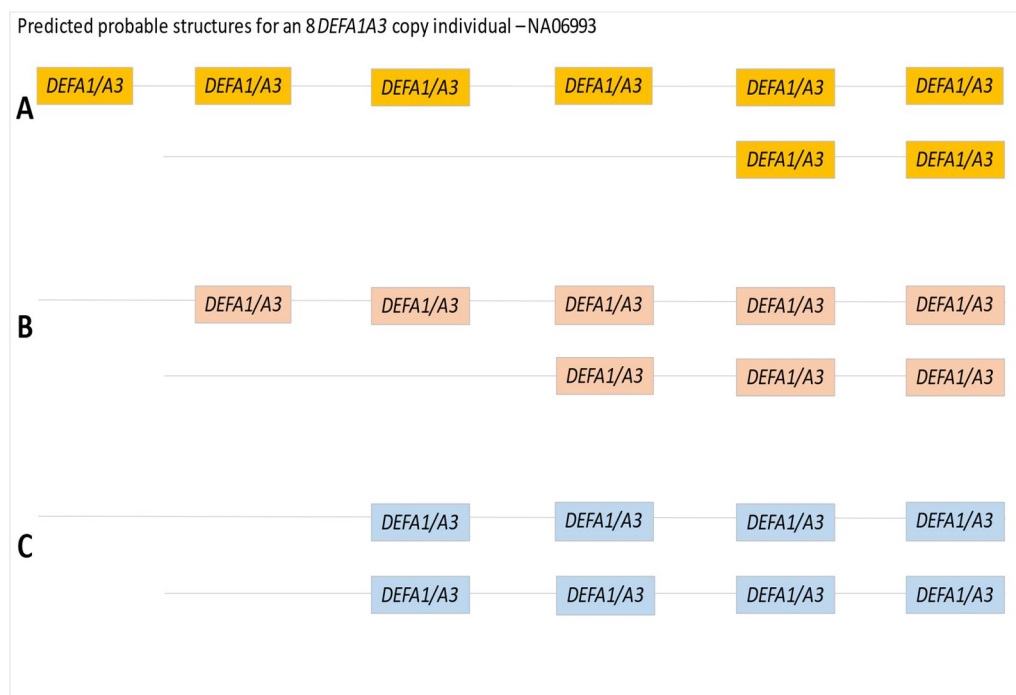


Figure 5.7. Possible general haplotype structures for NA06993. A) 6 + 2 copy haplotype, B) 5 + 3 copy, and C) 4 + 4 copy haplotype.

Figure 5.8 shows a screengrab from an IGV screen (Integrative Genomics Viewer) of bam files from PCR1 and PCR4 repeats. Both PCRs are important in

the amplicon tiling as they include useful SNP information that will distinguish repeats from either ends of the reads based on the location of the amplicon in the haplotype. For instance, in PCR1, we can differentiate the centromeric from the non-centromeric repeats by examining the block of SNPs found on the right side of reads (found in the red box). It is strongly noticeable that the block of SNPs in the centromeric repeats are distinct compared to the non-centromeric, and this phenomenon is consistent with other PCR1s across different samples. The centromeric SNP block seems to include more variation than the non-centromeric block, and this is due to the usage of one *DEFA1A3* repeat reference sequence for assembly (corresponding to a telomeric repeat). This also indicates that the centromeric repeat is distinct from the artificial reference sequence used and is therefore located centromerically (far right repeat of any given *DEFA1A3* haplotype).

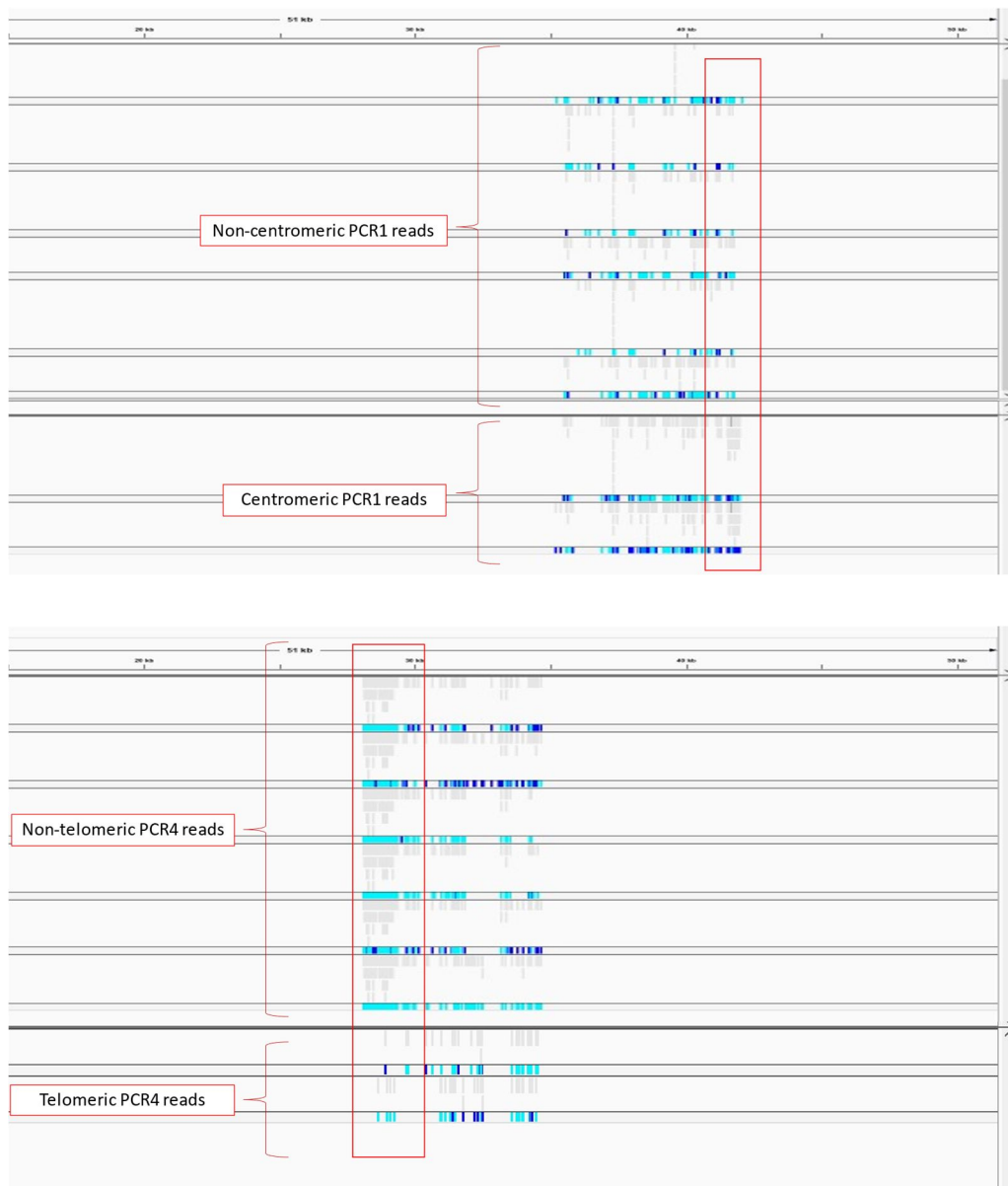


Figure 5.8. An IGV (version 2.5.2) snapshot of PCR1 (top) and PCR4 repeats (bottom) - in VCF format - from 8 repeats for sample NA06993 (zoomed out). The red rectangle represents a block of SNPs that differ in the centromeric and non-centromeric PCR1 repeats, and in the telomeric and non-telomeric PCR4 repeats. Light blue indicates SNPs different from the reference sequence, while dark blue indicates a mixed position relative to the reference sequence.

Based on Table 5.10, it is reasonable to say that there should be 7 copies of *DEFA1* and one copy of *DEFA3*. Thus, we would expect to have the *DEFA3* copy situated at the centromeric region of the *DEFA1A3* haplotype as would often be observed in the Reference Sequence class but not in Class 1 (see chapter 1, section

1.4, Figure 1.1). In Table 5.10, the last column (C_35383T repeat) shows that as expected, the *DEFA3* is in a centromeric position.

Table 5.11. Eight phased SNPs from Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). Each column represents a DEFA1 or DEFA3 repeat, and the first row displays labels given for each repeat sorted by the last variant position chosen for repeat separation. This table is from PCR1 reads only. PCR5 repeat separation was incomplete at this stage due to lack of candidate SNPs, although there are 8 variant positions that may be used for sorting, but there were no mixed positions identified in the reads. NC: non-centromeric repeat, C: centromeric repeat. These positions are found in the overlaps between PCR1 and PCR5 amplicons.

POS	NC_ 36556C	NC_ 36556T	NC_ 39657A	NC_ 41702T	NC_ 41199C	NC_ 41199T	C_ 35383G	C_ 35383T
35383	G	G	G	G	G	G	G	T
35488	A	A	A	A	A	A	C	A
35505	T	T	T	T	T	T	C	T
35514	G	G	G	T	T	T	G	T
35639	A	A	A	A	A	A	A	T
35645	T	G	G	T	T	T	G	T
35780	C	C	C	T	C	C	C	T
35821	T	T	T	T	T	T	T	G

NA07056

NA07056 is a female CEPH/CEU sample with 5 *DEFA1A3* copies and has Class 2 haplotypes [2] (see Chapter 1, section 1.4, Figure 1.1 for the different *DEFA1A3* haplotype classes). The *DEFA1(G):DEFA3(T)* ratio from PCR1 and PCR5 reads was determined as 1.5:1 and 1.49:1 respectively, consistent with three copies of *DEFA1* and two copies of *DEFA3*. The cent:non-cent ratio from PCR1 was 1.93:1, suggesting a total copy number estimate of 5.86, while the tel:non-tel ratio from PCR4 was 1.04:1, giving a total copy number estimate of 4.08. We were unable to construct the repeats of NA7056 in full (by tiling all PCRs together) due to high identity of repeats to one another, despite the success of sorting of the copies from PCR1 that was achieved; given that NA07056 has two class 2 haplotypes we might anticipate that the two haplotypes might show long patches

of sequence identity. However, the expected structure for NA7056 is a 3 by 2 copy composition as shown on Figure 5.9, as deduced from the telomeric placement of the *DEFA3* copies and the presence of *DEFA1* at the centromeric positions.

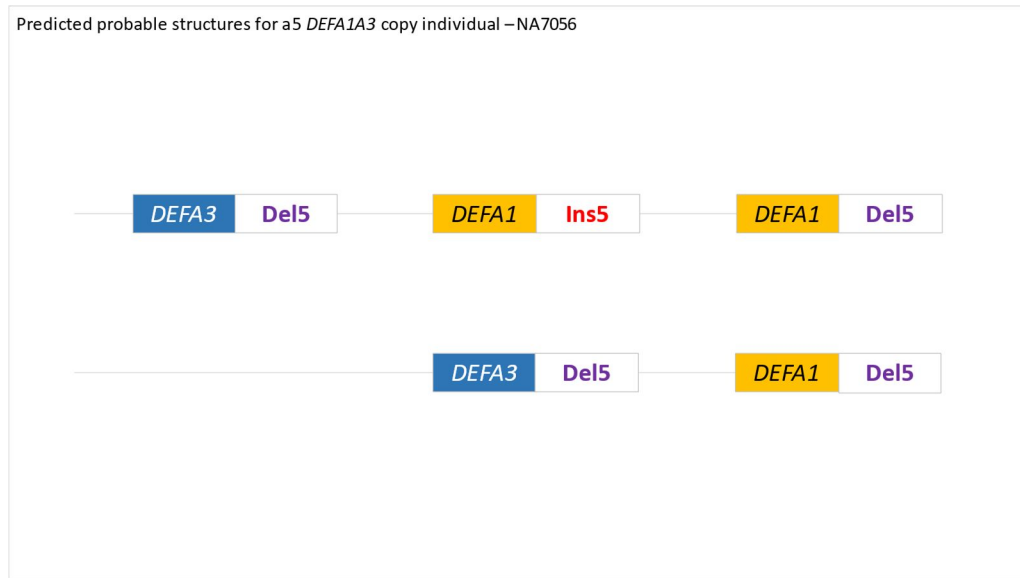


Figure 5.9. The expected Class 2 structure of NA07056 consisting of 5 copies of *DEFA1A3* (3 *DEFA1* copies + 2 *DEFA3* copies). The *DEFA3* copies are expected to be positioned in the telomeric region of the haplotype [2]. *Del5/Ins5* is a 5bp indel situated at position chr8:40301 in our sequences (ACAGA).

The expected construction of the structure was solely based on the information that was obtained from PCR1 reads from the centromeric and non-centromeric repeats and the SNP 35383G/T from Table 5.11 was also considered for predicting the structure. Additionally, we know that NA07056 is a class 2/class2 sample, indicating that *DEFA3* copies are likely to be at the telomeric position of any given haplotype. As expected, the *DEFA3* repeats are situated in the telomeric region of *DEFA1A3* haplotypes (NC_41691A and NC_39358T repeats). Both show 35383T variant which is indicative of a *DEFA3* gene. These were positively deduced from joining the telomeric PCR4 to *DEFA3* in PCR5 repeats.

Table 5.12. Eight phased SNPs from Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes *DEFA1(G)* from *DEFA3(T)*. Each column represents a *DEFA1* or *DEFA3* repeat, and the first row are labels given for each repeat sorted by the last variant position chosen for repeat separation. NC: non-centromeric repeat, C: centromeric repeat. MP: mixed position. These positions are found in the overlaps between PCR1 and PCR5 amplicons.

POS	NC_41691A	NC_39358T	NC_39358C	C_41691T	C_37233G
35383	T	T	G	MP	G
35488	A	A	A	A	A
35506	T	T	T	T	T
35514	G	G	G	G	T
35639	T	T	T	T	A
35645	T	T	T	T	T
35780	T	T	T	T	C
35821	G	G	T	G	G

NA07347

NA07347 is a male CEPH/CEU sample with 8 *DEFA1A3* copies and is a heterozygote for Class 1 / Reference Sequence haplotype classes [2] (see Chapter 1, section 1.4, Figure 1.1 for the *DEFA1A3* haplotype classes). The *DEFA1(G):DEFA3(T)* ratio was defined from PCR1 and PCR5 reads as 7.57:1 and 6.56:1, respectively. The cent:non-cent ratio from PCR1 was 1.23:1, while the tel:non-tel ratio from PCR4 was 2.52:1. Like sample NA06993, there are three possible general structures (Figure 5.7 from sample NA06993 section); for an individual with 8 *DEFA1A3* copies there may either be a 6 + 2, 5 + 3, or 4 + 4 copy structure. Although candidate positions for sorting were identified, determining which repeats were from the centromeric and non-centromeric regions was inconclusive due to the quality of reads in PCR1. Separating repeats from PCRs 2, 3 and 5 was successful except for PCR1 and PCR4. 5 repeats only were separated, and we were unable to distinguish the telomeric from non-telomeric repeats. Due to the quality of PCR1 reads, we were unable to tile the overlaps between PCR1

and PCR5 or PCR1 and PCR2 and therefore were unable to construct the full repeats to form the *DEFA1A3* haplotype of this sample.

Table 5.13. Eight phased SNPs from Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). Each column represents a DEFA1 or DEFA3 repeat, and the first row are labels given for each repeat sorted by the last variant position chosen for repeat separation. NC: non-centromeric repeat, C: centromeric repeat, (): not called by bcftools. These positions are found in the overlaps between PCR1 and PCR5 amplicons.*

PCR5	33384C	33384T	34396C	34396T	35194A	34236A	34236G	34179T
35383	G	G	G	G	T	G	G	G
35488	A^	A	A	A	A	A	A	A
35506	mixed	T	T	T	T	T	T	T
35514	G	G	T	T	T	G	G	T
35639	A	A	A	A	T	A	A	A
35645	Mixed	T	T	T	T	G*	G*	T
35780	C	C	T	C	T	C	C	C
35821	T	T	G	T	G	T	T	T

Table 5.14. Eight phased SNPs from Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). Each column represents a DEFA1 or DEFA3 repeat, and the first row are labels given for each repeat sorted by the last variant position chosen for repeat separation. NC: non-centromeric repeat, C: centromeric repeat. These positions are found in the overlaps between PCR1 and PCR5 amplicons.

PCR1	38034A	38034G	40197C	40197T	39496A	39496G	41765C	41765T
35383	MIXED	G	G	G	G	G	G	G
35488	A	A	A	A	A	A	A	A
35506	MIXED	MIXED	T	T	T	T	T	T
35514	MIXED	G	G	G	T	T	T	T
35639	MIXED	A	A	A	A	A	A	A
35645	T	MIXED	MIXED	MIXED	T	T	T	T
35780	MIXED	C	C	C	T	C	MIXED	MIXED
35821	MIXED	T	T	T	G	T	MIXED	MIXED

NA12248

NA12248 is a male CEPH/CEU sample determined by previous work to have a total of 6 *DEFA1A3* copies and has Class 1 and Class 2 haplotype structures [2] (see Chapter 1, section 1.4, Figure 1.1 for the different *DEFA1A3* haplotype classes). The *DEFA1:DEFA3* ratio from PCR1 and PCR5 reads was 4.85:1 and 5:1 respectively, consistent with a composition of five copies of *DEFA1* and one copy of *DEFA3*. The centromeric to non-centromeric ratio in PCR1 is 2.58:1, giving an estimated total copy number of 7.16, and the telomeric to non-telomeric ratio in PCR4 is 1.8:1 giving a total copy number estimate of 5.6. Assuming that the *DEFA3* copy is at the telomeric position, there are two possible *DEFA1A3* haplotypic structures shown on Figure 5.10. Constructing full haplotype sequences across the repeats was successful. All the amplicons were sorted efficiently into expected number of unique sequences (6 different sequences for PCR1, 4 and 5, and 4 distinct sequences for PCR2 and 3) and were then tiled together to construct the relevant repeats, with their relevant locations within a haplotype as shown in Figure 5.11. This is done by looking at the overlaps between each PCR (for example, by identifying the parallel identical SNPs from the overlaps between PCR1 with PCR2 or PCR5 with PCR1).

Table 5.15. Eight phased SNPs from the Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). Each column represents a DEFA1 or DEFA3 repeat, and the first row are labels given for each repeat sorted by the last variant position chosen for repeat separation. NC: non-centromeric repeat, C: centromeric repeat. These positions are found in the overlaps between PCR1 and PCR5 amplicons.

POS	NC_ 40189T	NC_ 39657A	NC_ 40189C	NC_ 39657G	C_ 41602C	C_ 41602T
35383	T	G	G	G	G	G
35488	A	A	A	A	A	A
35505	T	T	T	T	T	Mixed
35514	T	G	G	G	T	G
35639	T	T	A	A	A	A
35645	T	T	T	G	T	mixed
35780	T	T	C	C	C	C
35821	G	T	T	T	T	T

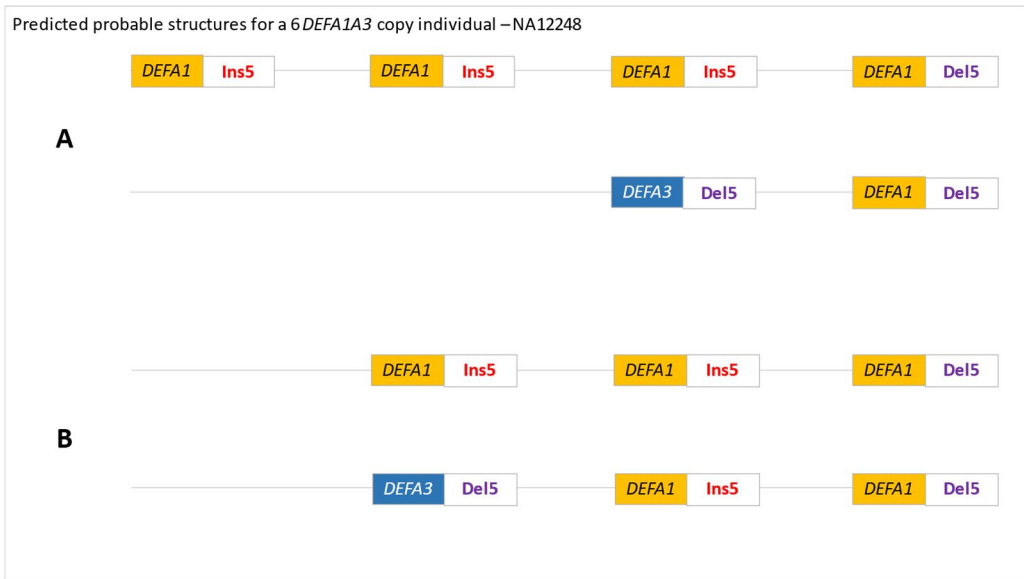


Figure 5.10. Possible haplotype structural organization for NA12248 based on other known structures of class 1 and class 2 alleles [2]. A) A 4 by 2 copy structure and a B) 3 by 3 copy structure. Del5/Ins5 is a 5bp indel situated at position chr8:40301 in our sequences (ACAGA). The *DEFA3* copy is shown in the telomeric position of the haplotype, as frequently observed in class 2 haplotypes [2].

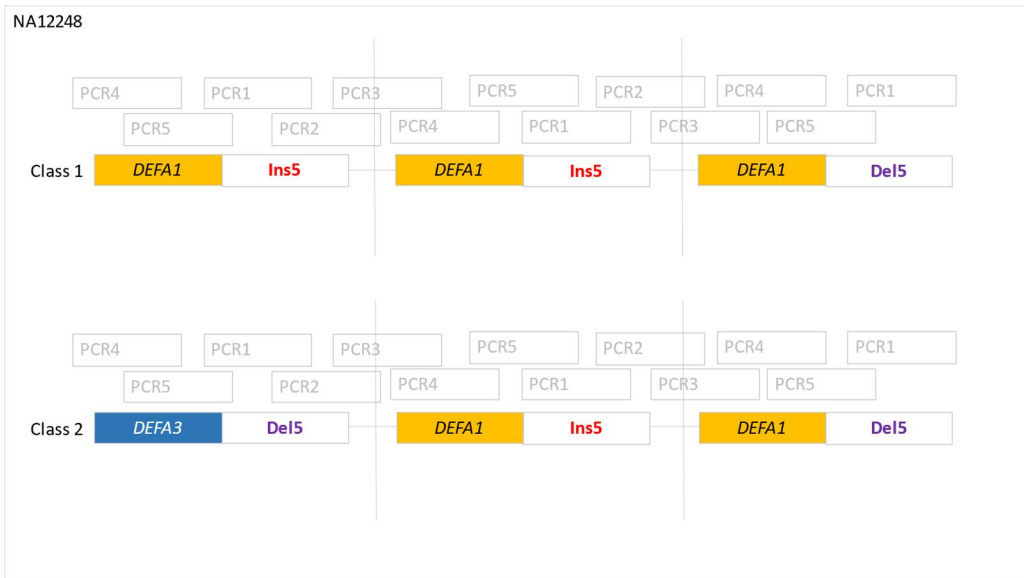


Figure 5.11. Full haplotype structure of NA12248 was constructed successfully for the *DEFA1A3* repeats. Del5/Ins5 is a 5bp indel situated at position chr8:40301 in our sequences (ACAGA).

Table 5.16. Summary of number of mapped reads per repeat per PCR bam file. Repeats number 5 and 6 represent the centromeric repeats. These repeats do not include PCRs 2 and 3 which span the DEFT1P pseudogene that is only found within the telomeric and middle regions of any given DEFA1A3 haplotype.

Repeats	PCR4	PCR5	PCR1	PCR2	PCR3
1	4275	965	486	4580	3148
2	5360	1516	592	5165	3108
3	3365	5557	478	4365	3075
4	3529	5941	716	5718	10133
5	7068	2398	469	-	-
6	3302	2000	1830	-	-

NA12249

NA12249 is a female CEPH/CEU sample with 9 *DEFA1A3* copies and has Exchange 1/Reference Sequence haplotype structures [2]. Table 5.16 shows a summary of the number of mapped reads per PCR bam file with their corresponding chromosomal positions.

Table 5.17. Summary of number of mapped reads per PCR bam file. The chromosomal locations for each PCR are demonstrated from the GRCh37/hg19 genome assembly. PCR1 and PCR5 includes the DEFA1(G)/DEFA3(T) gene.

Chromosomal location		
PCR1	chr8:6835034-6842185	462
	chr8:6854151-6861290	
	chr8:6873254-6880498	
PCR2	chr8:6837642-6844108	26049
	chr8:6856752-6863211	
PCR3	chr8:6842128-6848588	12809
	chr8:6861233-6867691	
PCR4	chr8:6828084-6834693	85835
	chr8:6847240-6853810	
	chr8:6866343-6872913	
PCR5	chr8:6830738-6836557	16553
	chr8:6849897-6855674	
	chr8:6869000-6874773	

The *DEFA1(G):DEFA3(T)* ratio was determined from PCR1 and PCR5 reads: 9.71:1 and 7.30:1 respectively. These ratios from both PCR reads are not consistent with each other (PCR1 and PCR5 includes a 1,530 bp overlap that contains the *DEFA1(G)/DEFA3(T)* position). PCR1 reads suggest there are approximately 10 copies of *DEFA1* and 1 copy of *DEFA3*, whilst PCR5 reads suggest there are 7 *DEFA1* copies and one *DEFA3* copy. The cent:non-cent ratio from PCR1 was 1.92:1, giving a copy number estimate of about 5.8, and the tel:non-tel ratio from PCR4 was 3.24:1, suggesting a total copy number estimate of 8.48. There are three possible general structures consistent with commonly observed patterns for individuals with Exchange 1 and Reference Sequence haplotypes and they are presented on Figure 5.12. Of these three, option C is most consistent with observation in European populations, in which the commonest copy number for a Reference haplotype is 5, and the commonest copy number for an Exchange 1 haplotype is 4 [2].

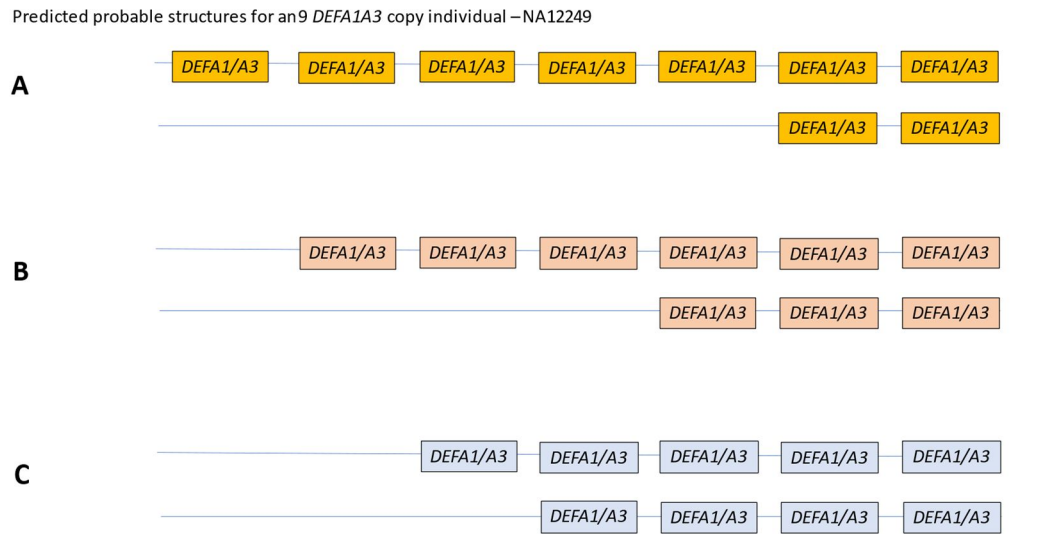


Figure 5.12. Possible general haplotype structures for NA12249. A) 7 + 2 copy and B) 6 + 3 copy and C) 5 + 4 copy haplotype structure.

Sorting of repeats from PCR1 was challenging, and only 7 repeats were extracted, of which two repeats consisted of only two (generated output file name: NA12249_PCR1_38458C_to_38826T_to_41970A.bam) and three reads (generated output file name:

NA12249_PCR1_38458C_to_38826T_to_41970T_to_36297C.bam). Candidate variant positions for sorting were restricted due to lack of adequate variable positions and the filtering process for PCR1 discarded numerous reads. Determining which repeats were from the centromeric and non-centromeric regions was inconclusive.

Although the proportion of *DEFA1(G)/DEFA3(T)* from PCR5 reads was 7.30:1, 10 repeats were extracted as shown on Table 5.17. 6 out of 10 repeats showed identical positions which causes ambiguity between the repeats and will affect how these will be sorted, but this also confirms that they are indeed identical repeats as was reported previously [2]. Those 8 positions are within the overlap between PCR1 and PCR5, so that it will be challenging to determine their unique structure and location within a haplotype. Although we were able construct the overlaps between the rest of the amplicons with one another for PCR2, 3 and 4 (but not PCR1 and PCR5), we were unable to construct the complete haplotype structures due to the ambiguity of the repeats.

Table 5.18. Eight phased SNPs from Tyson et al. [92] study that were determined by emulsion haplotype fusion PCR. Variant 35383G/T distinguishes DEFA1(G) from DEFA3(T). Each column represents a DEFA1 or DEFA3 repeat, and the first row displays labels given for each repeat sorted by the last variant position chosen for repeat separation. This table is from PCR5 reads only. PCR1 repeat separation was incomplete at this stage due to lack of candidate SNPs. POS: variant position. Variants highlighted in red correspond to unique difference compared to variants from the other repeats in a single reported position.

POS	33856G	31960A	31960G	33985A	33985G	31929G	34569A	31929A	35383G	35383T
35383	G	G	G	G	G	G	G	G	G	T
35488	C	C	C	C	C	C	A	C	C	A
35505	C	C	C	C	C	C	T	C	C	T
35514	G	G	G	G	G	G	G	G	G	T
35639	A	A	A	A	A	A	T	A	A	T
35645	T	T	T	T	T	T	T	T	T	T
35780	T	T	T	T	T	T	T	C	C	T
35821	G	G	G	G	G	G	T	T	T	G

In this section, we collated data from 8 SNP positions from *DEFA3* repeats for comparison among them and to evaluate the differences between the centromeric and the non-centromeric *DEFA3* repeats (Table 5.18). Although we were unable to define the location of the *DEFA3* repeats from samples NA07347 and NA12249 within their respective *DEFA1A3* haplotype, their SNP information is consistent to that of C_NA06993 and NC_NA12248. NC_NA07056_1 and NC_NA07056_2 repeats appear to have identical SNP information. Both repeats are found within the telomeric position. It appears that the position that distinguishes NC_NA07056_1 and NC_NA07056_2 from the other *DEFA3* repeats is 35514(G/T). We also collated the centromeric *DEFA1* sorted repeats for comparison (Table 5.19) but from three samples only and it seems the SNP positions among these repeats are variable with some positions recorded as mixed.

In summary, we were able to successfully reconstruct the full diploid *DEFA1A3* haplotypes for sample NA12248 only, although we were only able to sort several repeats from the rest of the samples with partial tiling. This is due to inadequate internal SNP information between the overlaps and the lack of candidate SNPs to use for repeat separation.

Table 5.19. *DEFA3* sorted repeats extracted using `samttools+awk` command across the five samples from the MinION experiment. Each column represents a single *DEFA3* repeat. SNP position 35383 distinguishes the *DEFA1* (35383G) from *DEFA3* (35383T). C: centromeric repeat, NC: non-centromeric repeat, *: unknown location within the *DEFA1A3* haplotype.

SNP position	C_NA06993	NC_NA12248	NA07347*	NA12249*	NC_NA07056_1	NC_NA07056_2
35383	T	T	T	T	T	T
35488	A	A	A	A	A	A
35505	T	T	T	T	T	T
35514	T	T	T	T	G	G
35639	T	T	T	T	T	T
35645	T	T	T	T	T	T
35780	T	T	T	T	T	T
35821	G	G	G	G	G	G

Table 5.20. Centromeric *DEFA1* sorted repeats that were defined from three samples only. Each column represents a single centromeric *DEFA1* repeat. MP: mixed position.

SNP position	NA06993 (35383G)	NA07056 (41691T)	NA07056 (37233G)	NA12248 (41602C)	NA12248 (41602T)
35383	G	MP	G	G	G
35488	C	A	A	A	A
35505	C	T	T	T	MP
35514	G	G	T	T	G
35639	A	T	A	A	A
35645	G	T	T	T	MP
35780	C	T	C	C	C
35821	T	G	G	T	T

CHAPTER 6

6. CRISPR-cas9 Sequencing

Firstly, gRNAs were designed (see Chapter 2, section 2.3, Table 2.6, and Figure 2.4) for the enrichment of *DEFA1A3* copy variable loci and neighbouring flanking regions surrounding the CNV and to sequence the run using Oxford Nanopore sequencing Technology using a MinION sequencing device. The sample used for sequencing is a gold standard genome sample, NA12878, in which high throughput analyses and sequencing were performed to obtain benchmark of accurate detection of small and large variations and to capture a robust consensus sequence [125, 158-161]. Sample NA12878 has a deduced total *DEFA1A3* copy number of 5 and consists of a 3 copy and a 2 copy *DEFA1A3* haplotypes [2].

6.1. Identifying the haplotypes of NA12878

DeepSeq – the sequencing facility at the University of Nottingham – extracted raw reads that were aligned to the *DEFA1A3* locus and the surrounding flanking region (hg19:chr8:6500000-7500000). In total, 449 reads were generated and aligned to the artificial reference sequence containing about 400 kb from chromosome 8 including the *DEFA1A3* repeat region (chr8:6700000-7100000 – the coordinates are from the GRCh37/hg19 genome assembly). Of those reads, 72.38% were within the *DEFA1A3* locus. The mean coverage of the raw reads was 28X comprising the entire region of *DEFA1A3* CNV plus the flanking regions. 57.91% of those reads are >10 kb in length, and only these long reads were included in the analysis. Figure 6.1 represents the mapped start and end positions of reads on chromosome 8 to a reduced reference genome (chr8:6750000-6910000). The red rectangle represents the actual positions of *DEFA1A3* locus and regions

flanking the gene, resulting in reads apparently derived from the 2-copy allele and the 3 copy allele. Approximately 260 reads are presented in the Figure, after reads that were shorter than 10 kb were excluded from further analysis. 13 reads only (2.89%) covered the entire *DEFA1A3* CNV including the flanking regions and were approximately 60 kb long (black lines in the figure) and they are presumably from the 3 copy haplotype. The rest of the reads were about 30 kb long and are apparently from the 2 copy haplotype. However, the 30 kb long reads (2 copy allele) were mapped to two different locations on the (3-copy allele) artificial reference sequence used (chr8:6828000-6860000 & chr8:6848000-6878000), presumably due to minimap2 [148] unable to distinguish between repeats as they are highly similar which causes the 2-copy allele to be misaligned. Using the 3 copy allele artificial reference sequence was challenging due to 2 copy reads aligning in both the middle and centromeric repeats (see Chapter 6, section 6.1 and 6.2, Figure 6.1).

Prior to the *DEFA1A3* construction, we investigated the nanopore reads further. 5 out of the 13 reads were on the forward strand in a fasta file, and the MEGAX software (version 10.0.5) was used to generate reverse complements and upload them into CLUSTAL omega for multiple alignment and generation of a consensus sequence. With Jalview 2.10.5, a consensus sequence was generated that was ~110kb in length, and nucleotides not within the *DEFA1A3* region from the consensus sequence were removed to generate the highest accuracy possible for consensus. After trimming the sequence, the length was 60.5kb in length of which corresponded to 4 contigs on the genome assembly (data not shown). The consensus sequence that was generated averaged 94.75% identity to the reference (chr8:6828284- 6879801) when multiple alignment tool (CLUSTAL omega) was used to compare both sequences. The three-copy allele reads were also used in de novo assembly on Canu, of which 12 reads were corrected and trimmed while 1 read was removed. Before correction, the total sequence was 0.74 Mb with 12.3x coverage and N50 of 0.053 Mb. After correction, results slightly changed with the removal of 1 read, resulting in the following: total sequence of 0.64 Mb, 10.6x

coverage, and N50 of 0.052 Mb. The final assembly with Canu is comprised of 1 contig sequence for the 3-copy allele with total length of 0.053 Mb after consensus generation. Multiple alignment was used on the contig to identify the identity of the sequence to the reference. Although the identity matrix of the consensus sequence (from the first method) to the reference was 94.75%, the contig generated by Canu was 99.06% identical to the reference. For the two-copy allele, 209 reads were chosen as input for Canu. However, 120 reads were only found by Canu with 3.15 Mb total number of sequences and 62.9x coverage and an N50 of 0.033 Mb for 94.17% of reads (reads with overlaps) and 0.017 Mb for 5.83% of reads without overlaps. After correction, 48.3% of reads were corrected with 2.02 Mb of total sequence, 40.4x coverage, and an N50 of 0.033 Mb. The final assembly with Canu is comprised of 1 contig for the 2-copy allele with total length of 0.072 Mb. Then, multiple alignment was used on the contig to identify the identity of the sequence to the reference. The contig was 99.07% identical to the reference. The 2-copy allele shows 2 different locations (chr8:6,828,000- 6,865,000 and chr8:6,848,000- 6,882,000) that were aligned to the reference sequence (data not shown). The ends of many sequences corresponded to the gRNAs designed and were precisely in their right positions after checking their corresponding locations on the assembly (DEFA1A3gRNA1 chr8:6,828,911-6,828,930 and DEFA1A3gRNA2 chr8:6,882,540- 6,882,559). One of the possible reasons could be that the reads were misaligned and the 2-copy alleles were shifted causing this complex feature. Multiple alignment on CLUSTAL omega was performed on both the 2-copy allele contig and the 3-copy allele contig (both contigs were produced by Canu) to identify the similarity of sequences, the 2-copy allele was 99.06% identical to the 3-copy allele. This suggests that the 2-copy allele should have been aligned correctly with the 3-copy allele but they were misaligned. Because the gRNAs are in a flanking region with overall sequence similarity to internal repeats, there is a tendency for one end of 2-copy sequences to be mismapped to the corresponding internal location on the (3-copy) reference assembly, thereby avoiding insertion of a 19kb gap. Further exploration of alignment parameters, and in particular

experimenting with gap penalties, might be useful in future work to give accurate (i.e. gapped) alignments of 2-copy sequences against a 3-copy assembly.

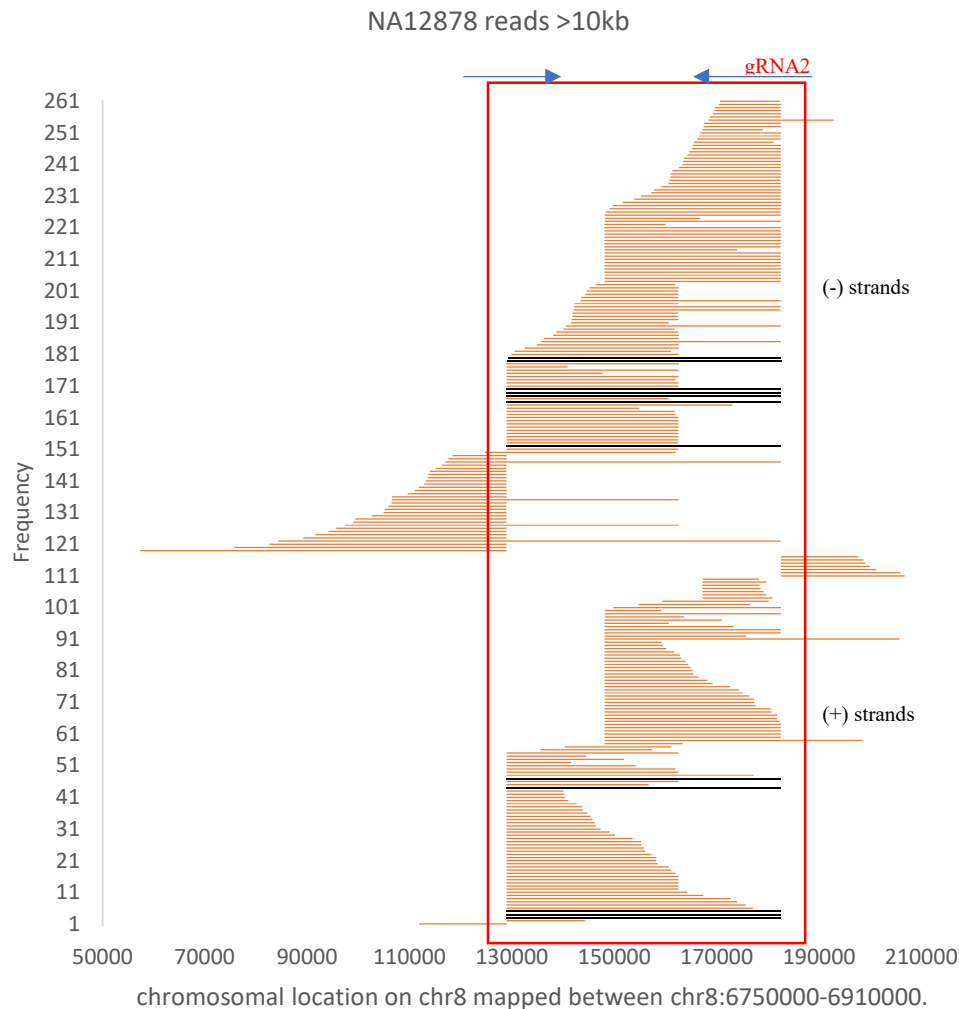


Figure 6.1. (-) and (+) strands of sample NA12878. Chromosomal location of reads on chromosome 8 mapped between chr8:6750000-6910000. The red rectangular box represents the location of *DEFA1A3* CNV and regions flanking the gene. Black lines represent the full length of the 3 copy allele reads. Total number of reads is approximately 260 reads.

6.2. Reconstructing the allelic structure

The mean coverage of the 3 copy allele reads is 7.3X (chr8:6700000-7100000), while it is 23X for the 2 copy allele (chr8:6700000-6851000 & chr8:6870300-7100000). The number of substitutions relative to the reference sequence identified in 3 copy and 2 copy alleles is 956 and 1890 variants,

respectively, while the number of indels called are 2,187 and 3,089 for 3 copy and 2 copy alleles, respectively. After identifying the length of each read, the 2 copy and 3 copy reads were separated into different sorted bam files to help in reconstructing the structure of the *DEFA1A3* gene in sample NA12878. For the 2 copy reads, an artificial truncated reference sequence (chr8:6700000-6851000 joined to chr8:6870300-7100000) was created to align the 2-copy allele to the first repeat of *DEFA1* (telomeric region) and the last repeat in the centromeric region (*DEFA3*). Using the original reference sequence was challenging due to 2 copy allele reads being aligned in both the middle and the centromeric repeat (as shown on Figure 6.1).

Figure 6.2 shows the haplotypic copy number and the structural arrangements for NA12878 that was reconstructed from the CRISPR-Cas9 reads using *freebayes* [162].

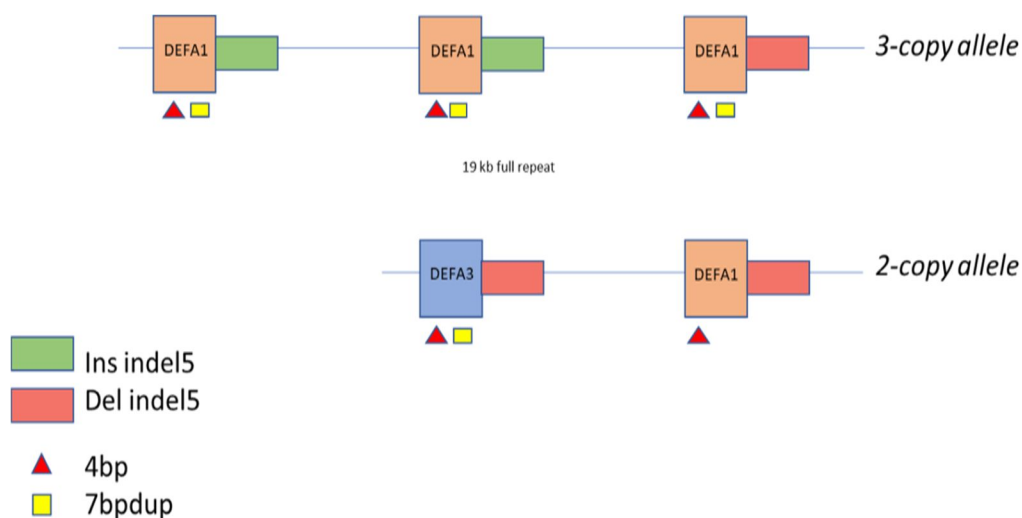


Figure 6.2. The reconstructed haplotypic structure of NA12878 from the 3 copy and 2 copy haplotypes using the CRISPR Cas9 sequencing.

Although *freebayes* called the *DEFA1/DEFA3* SNP and the internal indel variants correctly, the reads from the bam file showed variable SNPs/indels in relative positions. Table 6.1 represents the variation of each SNP/indel in reads. *Freebayes* calls variants and reports the most frequent SNPs found on each base-position as a consensus. But with bam files, one would expect that nanopore

sequencing would not call SNPs, especially indels, 100% accurately, due to the high error-rate compared with short-read sequencing methods like Illumina. For example, the indel5 for the second repeat in 3C-allele is not identical to the expected variant that is known from the literature. However, 11insA/12insGATG falls within the region of the variant “CCATA” on the reference sequence in which after including the insertion variants would appear something like this “CACAGATGTG”. The underlined variants are the inserted forms in the reads, but it shows that it contains the ACAGA variant. The same applies to the 4bp indel in the 3rd repeat, and the 7bp dup in the 2nd and 3rd repeat (data not shown), as well as the variants in the 2 copy reads (variants with an * in table 6.1).

Table 6.1. SNPs and indels in each read reported from 2 copy and the 3 copy alleles bam files. (*): variation in called SNPs and indels. The 3-copy allele consists of 13 reads only of ~60kb long. When the sequence is not perfectly accurate, all reads at each of these positions would not agree on a single allelic state. The artificial reference sequence used for mapping, the third repeat of the 3 copy allele and the second repeat of the 2 copy allele both have the 4bp deletion in exon 2, therefore relative to the reference sequence, the indel InsTAGA in the reads at those positions and GATA in other positions are all in the same allelic state.

3C			
	1 st repeat	2 nd repeat	3 rd repeat
DEFA1(G)/DEFA3(T)	12G/1delG	12G/1delG	9G/3T/1delT
Indel5 (ACAGA)	12A/13CAG/12A	11insA/12insGATG*	13delACAGA
4bp indel (GATA)	12GATA/1delGATA	13GATA	7insTAGA
7bp dup (TGTGACA)	13TGTGACA	1insT/insT*	insGACATGT
2C			
	1 st repeat	2 nd repeat	
DEFA1(G)/DEFA3(T)	46T/36G/6C/2A/3delG	74G/6T/2A/1C/17delT	
Indel5 (ACAGA)	delACAGA*	delACAGA	
4bp indel (GATA)	94GATA*/3delG/1delA	insTAGA	
7bp dup (TGTGACA)	TGTGACA*	insGACATGT (not reported by freebayes)	

6.3. Shared Variants in the 3 copy and the 2 copy alleles

Variants in the 3 copy allele showed 489 variants that are distinct from the reference genome, and 336 variants in the 2 copy allele (Figure 6.3). Among these variants, 60 of them are shared between the two alleles, and they are consistent with one another within the region. This shows that these variants are true differences from the reference sequence and are unlikely to be technical errors in sequencing.

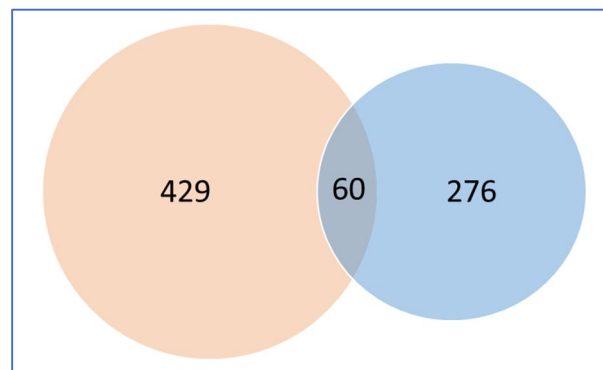


Figure 6.3. Shared variants between the 3C (red circle) and 2C (blue circle) haplotypes for sample NA12878.

6.4. Comparing the shared variants called by freebayes to the consensus sequence generated by Canu

We took some variants that are identical on both haplotypes in the 3 copy and the 2 copy alleles and compared them to the contig sequences generated by Canu [150]. Canu was used to *de novo* assemble the 3 copy and 2 copy reads in which contig sequences were produced. The variants are inconsistent in both cases suggesting that Canu does not distinguish between repeats due to high similarity in

sequence (Table 6.2). Among the 16 SNPs chosen arbitrarily, 73.33% of those variants from the 3 copy contig were identified as deletions and the remaining variants as substitutions. The 2 copy contig showed heterozygous positions, although some variants were identified correctly but there were also deletions identified in the same locations (73.33%). Although the consensus sequence produced by Canu showed ~99.06% identity to the reference sequence, this suggests that Canu is an unreliable method to deal with repetitive regions having structural rearrangements such as *DEFA1A3* CNV and this may lead to collapsing divergent repeats into one for which a specific phasing is needed to infer true haplotypes and resolve them.

Table 6.2. Variation of SNPs in both methods, freebayes and Canu in calling variants. 3 copy contigs and 2 copy contigs are produced by Canu. (SNPs: single nucleotide polymorphisms, BP: base-pair, 3C: 3 copy reads, 2C: 2 copy reads). 3C and 2C shared: are variants produced by freebayes in a VCF format at those specific locations telomerically. 3C and 2C contigs are produced by Canu.

SNPs	BP-position	3C & 2C (shared)	3C contig	2C contig
rs556597883	chr8:6833103	T	A/T	T/delT
rs143268564	chr8:6834179	G	delG	G/delG
rs4840655	chr8:6835514	G	T	G/delG
rs56186060	chr8:6834550	A	delA	A/delA
rs62486851	chr8:6849144	G	G/C	C/delG
rs2978908	chr8:6849289	C	delC	delC
rs138902863	chr8:6849820	T	delT	T/delT
rs141678943	chr8:6835569	A	delA	A/delA
rs62487469	chr8:6835821	G	delG	G/delG
rs59087747	chr8:6837440	G	delG	G/delG
rs750811579	chr8:6842895	G	delG	G/delG
rs2702923	chr8:6846454	T	delT	T/delT

6.5. Accuracy of base calling using known Exchange 1 and Class 2 flanking SNPs

Flanking SNPs reported previously [2] that are in LD are shown in Table 6.3. These SNPs determine the *DEFA1A3* haplotypic structure from the centromeric region. A comparison of flanking SNPs in the current analysis with those of previous studies confirms that their consensus states are consistent and that the 3 copy and 2 copy reads have the alleles expected for the known haplotypic classes shown via *freebayes* analysis.

Table 6.3. Flanking SNPs as reported by Black et al. compared to the 3 copy and 2 copy variants. (BP: base-pair, 3C: 3 copy reads, 2C: 2 copy reads).

Flanking SNPs	BP-position	Exchange 1	3C	Class 2	2C
rs7825750	chr8:6879803	T	13T	C	C: 76%
					T: 18%
					A: 5%
					G: 1%
rs62487514	chr8:6880094	A	12A/1C	C	C: 84%
					A: 14%
					T: 2%
rs4300027	chr8:6880575	T	13T	C	C: 69%
					T: 30%
					G: 1%
rs7826487	chr8:6880925	A	12A/delA	A	A: 96%
					delA: 4%

Nonetheless examining the frequency of each variant from the bam files, there are some inconsistencies with the number of different SNPs found. For example, although the expected variants in Class 2 haplotype are found at a higher frequency per flanking SNP on the table (ranging from 69% to 96%), there are still other nucleotides called (14% to 30%).

6.6. Comparing the current SNPs to a previously sequenced NA12878 using EHF-PCR

A previous MSci student (Emily White) in our lab used emulsion haplotype fusion PCR (EHF-PCR) [92] on the same sample of interest, which can accurately define the relative positions of sequence variants in the *DEFA1A3* gene across a haplotype. Emily produced two bed files that correspond to the 3 copy (NA12878_rs512398C) and the 2 copy (NA12878_rs512398T) haplotypes (unpublished data). Here, we compared the variants reported by *freebayes* from the CRISPR-Cas9 variants to Emily's. Approximately 65 variants combined from Emily's variants are found in our 3 copy and 2 copy variants. We found that those variants in both experiments are 100% identical, which suggests that using *freebayes* identifies the correct bases accurately. This also suggests that those variants called in both experiments are true variants and are not technical errors.

6.7. Comparing the CRISPR-cas9 nanopore sequences to the phased SNPs identified in Tyson et al.'s study

Tyson *et al.* [92] identified 8 phased SNPs for the *DEFA1A3* gene on the same sample of interest. We compared those phased SNPs to those in the CRISPR-cas9 data. We found that they are consistent with one another across haplotypes according to *freebayes*, but when exploring the bam files the alleles indicated for each SNP at their corresponding locations varied (Table 6.4). For example, rs77649030 has interesting findings in which in the 3 copy bam file shows that 100% of reads called the T variant in that position, but when using *freebayes*, it was a G variant. This is true for both the first and the second repeats in the 3 copy

reads. Similarly, with the 2 copy reads, the same variant was identified as G although the majority of reads in that position have T. A further factor to note is that the 2 copy allele might include some misattributed reads that are actually shortened reads from the 3 copy allele. For instance, rs145076681 on the first repeat shows that 46 reads had a T variant in that position while 36 other reads contain a G variant, suggesting that the T-carrying reads were shortened versions of the 3 copy haplotype. This SNP is important as it identifies the *DEFA1* or *DEFA3* gene.

Table 6.4. The SNPs listed are from Tyson et al. study [92] (haplotype emulsion fusion PCR was used in their experiment) and are compared to the current variants called from CRISPR-cas9 nanopore sequencing. The first repeat corresponds to the most telomeric end of the haplotype while the third repeat corresponds to the most centromeric end of the haplotype. The variants reported on the table are from the dbSNP build 135. The top SNP with an (*) is the non-synonymous variant that distinguishes DEFA1 (G) from DEFA3 (T) genes. (SNP: single nucleotide polymorphism; 3C: 3 copy reads, 2C: 2 copy reads). First row of each variant are results from this current study, and the second row is what Tyson et al. found.

SNP	Variant	3C			2C	
		1 st repeat	2 nd repeat	3 rd repeat	1 st repeat	2 nd repeat
rs145076681*	G/T	46T 12G 1delG	74G 12G 1delG	9G 3T 1delT	46T 36G 6C 2A 3delG	74G 6T 2A 1C 17delT
		G	G	G	T	G
rs148298380	A/C	6C 3A 2T 1G 1delA	6C 3A 2T 1G 1delA	7C 3A 3T	56A 25C 5T 2G 5delA	62A 21C 9T 3G 4delA
		C	C	C	A	A

rs139164630	T/C	10C 3delC	12C 1T	13C	55T 31C 1G 6delT	62T 31C 6delT
		C	C	C	T	T
rs147777993	G/T	6G 6T 1delT	8G 5T	6G 6T 1delT	57G 14T 7A 15delT	57T 15G 22delT
		G	G	G	G	T
rs77140339	A/T	13A	13A	13A	55T 33A 6delA	92A 3T 3delA
		A	A	A	T	A
rs77649030	T/G	13T* (called as G by freebayes)	13T* (called as G by freebayes)	12G 1T	87T 3A 1G 3delT* (called as G by freebayes)	68T 22G 2A 1C 6delT
		G	G	G	T	T

chr8:6835780; chr8:6845897; chr8:6873996	C/T	12T 1delT	12T 1delT	13T	93T 2delC	59C 36T 5delT
		T	T	T	T	C
rs77855524	G/T	11G 2T	11G 1A	13G	83G 7T 4A 1delT	94G 1T 1C 4delG
		G	G	G	G	G

6.8. Comparing the same phased SNPs from Tyson et al.'s study and the current study to that of the amplicon-sequencing method

We wanted to compare the variants from PCR1 reads produced with the amplicon sequencing method (see chapter 6, section 6.1, part A for *First Amplicon Experiment*). Similarly, *freebayes* was used. In this analysis, we just reported the percentage of these variants in PCR1 (Table 6.5).

Table 6.5. The variants reported on the table are from the dbSNP build 135. The top SNP with an () is the variant base that distinguishes the DEFA1 gene (G) from DEFA3 (T). The last column represents the predicted outcome ratio for the variants of interest from Tyson et al. [92] and what is reported in the current CRISPR analysis.*

PCR1	Variant	%	Predicted outcome of variants based on a previous study
rs145076681*	G/T	77/20 (A/C:1/1)	80:20
rs148298380	A/C	58/17 (T/G:23/1)	40:60
rs139164630	T/C	52/44 (A/G:3:1)	40:60
rs147777993	G/T	50/45 (A/C:4/2)	80:20
rs77140339	A/T	A	80:20
rs77649030	T/G	T	40:60
chr8:6835780; chr8:6845897; chr8:6873996	C/T	27/72 (A:1)	20:80
rs77855524	G/T	87/12	100:0

To summarize, a total of 449 reads were generated, of which 72.38% were within the *DEFA1A3* CNV region. Only reads that are > 10 kb in length (57.91%)

were included in the haplotype reconstruction. Among those reads, only 13 reads were from the 3 copy allele (2.89%). We furthermore compared our current SNPs to previously sequenced samples via EHF-PCR from the Tyson *et al.* paper and from unpublished data in our lab group, which suggests the CRISPR-cas9 method is reliable. The assembly of diploid haplotype of *DEFA1A3* CNV for sample NA12878 was successful, of which the structure is consistent with that of previous studies [2, 92]. Exchange 1 and Class 2 haplotypes were observed as expected, the 3-copy haplotype follows the Exchange 1 class, while the 2 copy haplotype is Class 2.

CHAPTER 7

7. Discussion

7.1. Aggressive Periodontitis Association Study

Aggressive periodontitis (AgP) is a destructive disorder, which is characterized by its acute and rapid nature, early onset (occurring in patients younger than 30 years), and periodontal tissue loss [106]. In 2013, a cohort of German samples was used for a GWAS study describing the association of the variant rs2738058 to chronic periodontitis (CP) [163]. Later, Dommisch *et al.* [164] demonstrated an increase in the HNP1-3 peptide concentration found in the patients, revealing a 15 fold increase in AgP patients and a 60 fold increase in CP patients. Recently, the largest GWAS periodontitis study was completed which included >800 AgP cases and >7000 controls from German and Dutch samples. Two independent variants downstream of the *DEFA1A3* locus showed evidence of strong association with AgP (rs2978951 and rs2738058). Their data validated the role of rs2738058 (P -value: 6.78×10^{-10}) from Teumer *et al.* Although these two variants were highly associated with periodontitis, no evidence shows how *DEFA1A3* CNV contributes to the disease, and therefore this study aims to investigate this association. To date, in human diseases, studies of SNVs and CNVs have been constructed on various approaches with differing hypotheses. SNPs and CNVs coexist in the genome, and they influence one another's measurement and therefore may influence the phenotype or disease individually and collectively [165].

The main summarized outcomes of this study are:

- 7 Lead SNPs showed association with AgP phenotype, the most significant variant being rs57688776 (P -value: 1.14×10^{-4} ; OR: 0.61 [95% CI: 0.48-0.79]).
- Lead SNPs showed no association with *DEFA1A3* copy number, suggesting *DEFA1A3* CNV has no influence on the AgP phenotype.
- rs2978951 and rs2738058 are associated with both AgP and *DEFA1A3* copy number. Both SNPs were associated with periodontitis in a previous GWAS study [102].
- rs4300027 showed strong positive association to *DEFA1A3* copy number (P -value: 5.20×10^{-28}), followed by rs7825750 (P -value: 1.50×10^{-16}) in AgP-GERx1kGP. Both are tag SNPs in characterizing the *DEFA1A3* haplotype classes.

Prior to case control association analysis, copy number measurements from the *DEFA1A3* locus were deduced using 1000 Genomes Project sequences. The copy number ranged from 3 to 8 with evidence of clustering around the integer values. However, there is some evidence of dispersion (less definition) around the integer values from copies higher than 8 (chapter 3, section 3.1, Figures 3.1 and 3.2). Remarkably, sample HG02554 (from an African-Caribbean ethnicity) demonstrated a total number of 28 *DEFA1A3* copies, which is out of the normal copy number range of any given population (data not shown). Sample HG02554 was then removed from the analysis as an outlier. This sample was also reported in Hughes *et al.* [166] where fibre-fish suggests a copy number of 26 (which includes a 24 copy haplotype). In addition, our EUR copy number frequency appears to agree with the EUR copy number distribution presented by a previous study [86]. However, the copy number measurements in this current analysis showed 4 to 14 copies for the EUR samples compared to 3 to 16 copies from Khan *et al.* This could be due to having different sample sizes in both studies. Khan *et*

al. included 600 EUR samples compared to this current study (504 EUR samples), which suggests that including more samples may reflect greater diversity in the European population.

This report provides clear evidence that the lead SNPs selected in this study are associated with AgP, and after controlling for *DEFA1A3* copy number, the association remains statistically significant (see chapter 3, section 3.4). Those lead SNPs found telomerically of the *DEFA1A3* CNV are novel and have not been reported previously prior to current analysis. Although no formal power calculation was undertaken, it is inevitable that the statistical power of our analysis is reduced compared with the original Munz *et al.* study, both because fewer samples and fewer SNPs are included. Our results rely on the analyses of Munz *et al.*, which originally included a GWAS of samples from Dutch, German, and Turkish cases and controls. In this reduced subset of SNPs, the case-control study focused on chr8:6700000-7100000 region, where the *DEFA1A3* locus is located (see chapter 2, section 2.1). The Dutch cases and controls, and German controls were excluded from the analysis since *DEFA1A3* copy number information was not available for these samples. We compared the two variants identified as suggestive risk loci for periodontitis from the Munz *et al.* study to our current analysis, and we found that those variants highlighted in Munz *et al.* are much more significant than in our current analysis, but that is due to the reduced number of samples and SNPs studied (Table 7.1).

Table 7.1. Comparison of two variants identified as suggestive risk loci for periodontitis from the Munz *et al* study. AgP-Gerx1kGP: 1000 Genomes Project European controls and German cases. Pooled: 1000 Genomes Project European controls and German cases, plus the Blueprint GBR controls. SNP: single nucleotide polymorphism, A2: minor allele, A1: major allele, ADJ: adjusted, OR: odds ratio, CI: confidence interval, P: P-value.

SNP	r ²	BP	A2	A1	ADJ	AgP-Gerx1kGP		Pooled		Munz <i>et al.</i>	
						OR [95% CI]	P	OR [95% CI]	P	OR [95% CI]	
rs297851		chr8:6965773	A	G	-	1.24 [1.04-1.47]	1.6x10 ⁻²	1.19 [1.01-1.4]	3.5x10 ⁻²	1.32 [1.18-1.48]	2.1x10 ⁻⁶
					SEX	1.24 [1.04-1.48]	1.5x10 ⁻²	1.19 [1.01-1.4]	3.7x10 ⁻²	1.32 [1.18-1.48]	2.4x10 ⁻⁶
rs2738058	0.27	chr8:6964095	T	C	-	1.24 [1.05– 1.48]	1.4x10 ⁻²	1.23 [1.05-1.45]	1.2x10 ⁻²	1.29 [1.15-1.45]	1.3x10 ⁻⁵
					SEX	1.25 [1.05– 1.48]	1.3x10 ⁻²	1.23 [1.04-1.44]	1.5x10 ⁻²	1.29 [1.15-1.45]	1.7x10 ⁻⁵

The lead SNPs showed no evidence of association to *DEFA1A3* copy number except for rs2978951 and rs2738058. Similarly, two centromeric tag SNPs (rs4300027 and rs7825750) that characterize the *DEFA1A3* haplotype classes are associated with the copy number, but not rs7826487 (which is one of the top SNPs in the centromeric flanking region that is associated with AgP). rs7826487G (protective allele) is characteristic of Class 1 haplotypes, and rs7825750C (protective allele) defines Class 2. It was reported that Class 2 is the second most frequent class in the EUR population (about 30%) as well as in ASN and AMR populations [2]. While Class 1 is the most common haplotype in AFR populations, it is less common (about 12%) in European populations and this might suggest the reason why we found no evidence of association of rs7826487G to *DEFA1A3* copy number in the current study population.

SNP rs4300027 is the most strongly associated variant with *DEFA1A3* copy number in this analysis and was identified as a tag SNP for copy number in a previous study in EUR-populations [86]. The minor allele for rs4300027 is T and was previously associated with high *DEFA1A3* copy number (4-5 copies per haplotype) [86].

Finally, we also show no evidence of association of *DEFA1A3* copy number to AgP. This is interesting as it contradicts the findings of a previous study by an MRes student (Eric Moise) in our lab (data not published). There could be various explanations for this. The previous study used an R package called CNVassoc [167], which allows association testing between copy and phenotypes using copy number probabilities. This means that it accounts for the uncertainty of the copy number calling, and therefore includes the mean and standard deviation of each inferred copy number. CNVassoc also considers batch effect as a factor in the analysis and therefore helps in reducing false positives. By contrast, in the current study PLINK was used, and it does not account for copy number uncertainty. As a result, it considers the copy number values included as estimates with no error. It may also be that population stratification in our case-control

analysis alters the results due to comparing the Germans with the 1000 Genomes Project samples.

The number of samples included that had copy number information was only about 60% of the total number of cases, so this could also reduce the statistical power of the outcome. In our study, there is an overrepresentation of GBR samples in controls after the inclusion of BP-GBR (197 samples from the Blueprint study, UK-based population) in the pooled analysis, but this does not explain the agreement in results from both the AgP-GERx1kGP-EUR (German cases plus European controls from the 1000 Genomes Project) and the pooled analysis. Although there are discrepancies in results between our current *DEFA1A3* CNV analysis to AgP and Eric Moise's, this shows that validations still need to be done to confirm the results obtained.

In conclusion, there is no clear indication of direct association of *DEFA1A3* CNV to AgP or lead SNPs. Our analysis shows different outcomes to what was found in a previous study (Eric Moise's analysis) in terms of *DEFA1A3* association to AgP, and while the same samples were used in both analysis, different methods were performed. This suggests that more investigation needs to be done with a more accurate matching of controls to cases to avoid subpopulation stratification. Despite using a European population for controls from the 1kGP, the controls do not include any German samples to accurately represent the dataset of German cases.

7.2. Blueprint Study

Multi-allelic CNVs are poorly studied due to difficulty in measuring gene copy numbers in sequence-based methods accurately. As the copy number increases, it becomes more challenging to accurately pinpoint the correct integer because high copy numbers are most error prone. This can influence replication studies when genotyping sequences with high identity in CNV regions. CNVs are usually collapsed into one region after mapping back to genome assemblies due to high

similarity in sequences. *DEFA1A3* CNV is one example that exhibits high identity between repeats. The total genomic copy number in a European population has been reported to range from 3 to 16 *DEFA1A3* copies [52, 86]. In this study, we have shown that the total *DEFA1A3* copy number ranges from 4 to 11 copies in the UK population studied (197 samples). This finding agrees with previous studies on European samples [52, 86] and was also replicated in the 1000 Genomes Project European samples in our study (see Chapter 3, section 3.1 and 3.2). This has been achieved by using the read-depth method [51]. We also show that the measurement to determine the copy number was accurate by comparing two independent approaches that estimates the *DEFA1A3* copy number.

We confirmed that total *DEFA1A3* copy number had no detectable correlation with the level of mRNA expression in blood cells, and this validates the findings of Aldred *et al.* [52]. This indicates that other heritable factors such as variation in elements controlling signalling or transcription, or non-heritable factors like immune activation, may influence the observed level of expression. Although no correlation was found, it is important to note that the levels of *DEFA1A3* mRNA expression in neutrophils is approximately 150x greater than that of T-cells, while monocytes showed no detectable expression levels. These observations confirm the experimental findings of a previous MRes student Ignacio Rodriguez (data not published) where 9 samples were analysed with experimental measurements (RT-qPCR) but showed similar outcomes for relative expression of *DEFA1A3* in blood cell subsets. We also observed a significant but weak correlation between *DEFA1A3* and *ELANE* expression levels in neutrophils, which may reflect immune cell activation. *ELANE* (elastase, neutrophil expressed) gene maps to chromosome 19; its key role is in the innate host defence and to modify the functions of natural killer cells, monocytes, and granulocytes. Previous studies have reported that the *ELANE* gene played a role in periodontitis development, but the sample size was small, and a larger study sample is needed for replication [168, 169].

Aldred *et al.* also investigated the mRNA levels in human white blood cells, and there was evidence of significant positive correlation of the relative proportion of *DEFA1:DEFA3* mRNA levels to their equivalent gene number (r^2 : 0.713, $P < 0.01$). However, the sample size consisted of only 17 individuals and was done with semi-quantitative PCR methods. In this current study, the sample size consisted of 197 individuals from a UK-based population. Two different computational methods (samtools+awk [146] and bcftools mpileup+AD [147]) were used to evaluate the *DEFA1:DEFA3* RNA ratios to those of the genomic. The results suggest that there is a clear correlation between the relative proportion of *DEFA1:DEFA3* RNA levels and equivalent gene copy number in both methods. Similarly, there was a significant correlation detected for two other common variants that are found within the transcribed part of the *DEFA1A3* gene on the second and the third exon. However, the weakest correlation was from the 6835194(G/A) using the samtools+awk method, while the strongest correlation was observed from the 6836028(C/G) using the bcftools mpileup+AD method (R : 0.789, P -value: <0.0001) (see chapter 4, section 4.5, Table 4.1). This demonstrates that each repeat is approximately equally expressed regardless of haplotype state or position of repeats. However it is difficult to explain the different levels of correlations between different variants in the same samples, at least some which has to come from variation in the accuracy of the read calling. This study also confirms that 9.7%-11.2% of samples lack the *DEFA3* gene which is concordant with a previous study [52].

One of the drawbacks in our current findings with samtools+awk and bcftools mpileup+AD (although a positive significant correlation was detected) is the variation in correlation between both methods. An issue with the samtools+awk and bcftools mpileup+AD approaches is, when re-aligning the reads into a one repeat reference sequence, this may introduce reference bias. This may also affect the accuracy of allele sorting with samtools+awk. Reference bias is when reads that differ sufficiently from the reference will not correctly map and are either mapped to an incorrect location in the genome or remain unmapped. This will lead

to inaccurate analysis of reads and will produce false negatives/positives in variant calling. For example, the *HLA* loci (human leukocyte antigen) is highly polymorphic in the human genome and mapping bias may occur as was reported by Brandt et al. [160]. Additionally the levels of read counts covering the *DEFA1A3* region in the WGS reads were low and therefore not informative enough to deduce an accurate outcome for genomic copy number. This may have also played a role in the variation of levels of correlation in both methods, but should still apply equally between samples. Having high coverage reads surrounding CNV regions such as *DEFA1A3* will further improve the outcome for future studies.

7.3. Nanopore Sequencing

7.3.1 Amplicon Sequencing

Unlike the current CRISPR-cas9 sequencing study (see chapter 6), the principle used here was a novel long amplicon sequencing approach that focused on amplifying regions between 5.5 kb to 7.5 kb long based on the primers used and their chromosomal locations (see chapter 2.3, Table 2.7 and Figure 2.5). The regions amplified harbour several factors that supported sorting and classifying where each *DEFA1A3* repeat sits in any given haplotype: 1) amplicons from PCR1 and PCR5 per sample included the *DEFA1*(G) and *DEFA3*(T) SNP that manifested in inferring the *DEFA1:DEFA3* ratio, which supported estimation of the copy number. Similarly, PCR1 and PCR4 amplicons included unique centromeric and telomeric SNPs, respectively, to obtain the relevant ratios, 2) overlapping regions between amplicons that included crucial unique SNP information to assist in joining the amplicons to form the *DEFA1A3* repeats, 3) unique SNP positions that distinguished the telomeric from the non-telomeric amplicons in PCR4, and 4) similarly for PCR1, unique SNP information that identified the centromeric amplicons from the non-centromeric ones. For either telomeric_PCR4 and centromeric_PCR1, there will only be two copies each in any haplotype studied. This approach was implemented in initial work by a previous PhD student from

our lab group (Xiao Xu, unpublished). Nevertheless, it was an unsuccessful attempt due to 1) PCR-mediated recombination affecting the amplification step (amplicon sizes ranged between 8 kb to 10.5 kb), and 2) the low accuracy obtained from the MinION sequencing device due to an older version of flow cell used (R7 flow cell).

In this study, we used two sequencing devices (for individual experiments, see chapter 2, section 2.3) from Oxford Nanopore Technologies, a Flongle and a MinION. This enables sequencing long reads uninterruptedly. Both devices are applicable to studying the *DEFA1A3* locus (a full repeat is ~19 kb), which enables the evaluation of highly repeated regions that would surround a large territory of repeat unit and simplify the *de novo* assembly selection.

For the primary experiment, the number of total raw reads generated per sample were 1,611 and 1,265 reads for NA12878 and NA12760 respectively. Although those are sufficient reads to acquire a consensus sequence and to further analyse the variations within a given reference sequence, it is not ideal for a region with a phenomenon such as CNVs or repeats. Additionally, the current amplicon sequencing method consists of producing 5 amplicons per sample. Therefore, assuming equal number of reads per amplicon were produced, we would expect to obtain between 300.8 and 203.8 reads per PCR bam file (for NA12878 and NA12760, respectively). Given that both samples consist of 5 copies of *DEFA1A3*, this indicates we would expect to obtain 5 repeats from PCR1, PCR4 and PCR5 (between 60.16 and 40.76 reads per PCR bam file). However, this was not the case and fewer reads were generated, due to several factors: the quality of the reads was below the expected value, and this may be due to limitations from the PCR yield during library preparations (see Chapter 5, section 5.1, Table 5.2 and 5.3) and the number of reads extracted during the repeat sorting were low. Prior to repeat sorting, reads were aligned to a custom-made artificial reference sequence that consists of only one *DEFA1A3* repeat (found telomerically in the GRCh37/hg19 genome assembly; chr8:6800001-6851705). This approach was adopted because reads may misalign due to the high similarity in sequence and CNVs are normally

collapsed into one region. Normally in *Homo sapiens*, two homologous copies are found in every chromosome. Yet in conventional de novo assembly procedures, the two copies are collapsed and are represented as a haploid sequence in the reconstruction of genome sequences. This may result in assembly errors and may overlook regions with high diversity between haplotypes [170]. This will then affect the ability to capture true repeats that may represent a unique feature of a haplotype. In our methods, using a one repeat reference sequence (instead of three repeats as seen from hg19 assembly) helps extract repeats by looking at single variable SNP positions and then sorting them accordingly even when the repeats are highly identical. We attempted to use two publicly available tools for repeat sorting, AssociVar [157] (see chapter 6, section 6.1) and samtools+awk command [146] (see chapter 2, section 2.2, Figure 2.2). Although AssociVar appears to be a useful tool, it was excluded at a later stage due to limitations of the tool in our analysis, and this may be due to AssociVar unable to handle large reference sequences. The unsuccessful use of the tool was also observed in the study of mycoviruses and cowpea mosaic virus [171, 172]. In contrast, samtools+awk command successfully sorted reads into separate files based on variant status at specific positions. In this first experiment, the assembly of repeats to one another into an allelic contig was not successful based on the overlaps between PCRs due to the low coverage of the centromeric tag SNPs from PCR1. In addition, the number of the nanopore reads were not sufficient regardless of the centromeric tag SNP information.

For the next experiment, two flongles were used (one for each sample individually). As a consequence, the pooling and barcoding steps were eliminated. This run generated 195.93 and 354.63 times more reads than the primary run for NA12878 and NA12760, respectively (see chapter 5, section 5.1, Table 5.2). Similarly, the total yield in the second run was 135.15 and 312.8 times higher. This suggests that the repeat sorting of alleles will not be challenging to process for various reasons: we have the liberty to lose multiple reads without it affecting obtaining true repeats, and this will not result in the loss of spatial SNP information

between overlaps and in centromeric and telomeric regions of the *DEFA1A3* locus. In addition, the primers used are from NanoPCR1_new (see chapter 2, section 2.3, Table 2.7). These primers amplify a region that further extends to include four tag SNPs in the centromeric *DEFA1A3* repeat. This is specifically important for the tiling construction of the repeats per *DEFA1A3* haplotype. Like the first run, the reads were aligned to reference sequence that consists of only one *DEFA1A3* repeat. We attempted to use the samtools+awk command for repeat sorting, and we successfully separated repeats from each PCR. This has led into the successful construction of the *DEFA1A3* haplotypes for both NA12878 and NA12760 (see chapter 5, section 5.1, part B). Despite a few SNP positions being reported differently compared to Tyson *et al.* in sample NA12878 (1 out of 8 phased SNPs were different in 4 repeats, see chapter 5, section 5.1, Table 5.6), overall called SNPs matched those previously reported [92].

Next, we attempted to use a MinION device to sequence 16 new samples (sample concentrations: 10-20 ng/μL). The sequencing run consisted of a pooled experiment from barcoded samples (out of 16 samples initially planned, only 5 were sequenced). The run took 72 hours as opposed to the flongle runs which were 24 hours only. The software versions for MinKNOW and guppy were different for the MinION experiment compared to the flongle runs (see chapter 5.1, Table 5.1 and chapter 5.2, Table 5.7). Out of the five samples sequenced, only sample NA12248 was successfully constructed for the full *DEFA1A3* haplotype (see chapter 5, section 5.2, Figure 5.11; Class1/Class2 haplotype). Although the haplotypic structure of sample NA12248 only was complete, we were still able to sort several repeats from other samples sequenced and partial tiling was performed. Complete tiling was unsuccessful for the other samples for several factors: 1) incomplete repeat sorting from some PCR bam files due to lack of good candidate SNP positions, 2) the inability to distinguish centromeric/telomeric reads from non-centromeric/non-telomeric reads, 3) high identity of repeats to one another despite obtaining unique copies from PCR1, but SNP information in the overlapping regions is identical to all copies (e.g. Sample NA07056, Class2/Class2

haplotype). The most problematic PCR bam file to sort was from PCR1 that harbours the *DEFA1(G)/DEFA3(T)* SNP, and it is unclear why this occurs. It may potentially be due to the high identity of the *DEFA1/A3* repeats to one another and it is not easily detectable by nanopore sequencing as true variation. This may possibly be caused by various factors that could influence the error rate, such as the quality of the base-calling used to interpret the signals, the DNA translocation speed through the nanopore membranes, or when the nucleotides are misidentified due to the interference with the electric signals.

In addition, for some samples in this study, the number of PCR1 reads was lower compared to other PCR reads. Samples NA06993 and NA07347 documented the lowest number of reads compared to other samples (see chapter 5, section 5.2, Tables 5.8 and 5.9), due to the PCR yield during library preparation. Although we used 10 ng/uL for all samples (see chapter 2, section 2.3, Table 2.8), potentially increasing the concentration may possibly improve the sequencing yield. Despite their lower yields compared to other samples, the outcome from the MinION experiment is undeniably better than our first flongle run (see chapter 5, section 5.1, part A).

The inability to classify centromeric from non-centromeric repeats will affect the entire tiling path process of *DEFA1A3* haplotype construction, and this is also true for the other repeats from any PCR. However, the *DEFA3* allele can commonly be either located centromerically or telomerically on a single haplotype and this permits the identification of which *DEFA1A3* class the sample comes from. For instance, in Table 5.18, samples NA07347 and NA12249 have unidentified *DEFA3* positions but we were still able to distinguish it from *DEFA1* repeats. It is also important to note that, while comparing *DEFA3* repeats between samples, we observed that position 35514(G/T) was the only position that was variable. This was also the case for samples NA12760 and NA12878 in our flongle experiments and from Tyson *et al.* [92].

7.3.2 CRISPR-Cas9 Sequencing

The principle of a targeted Cas9 enrichment method is to dephosphorylate the ends of the adaptor DNA and target cleavage with Cas9 (liberating a 5' phosphorylated end) to ligate adaptors specifically to the region of interest [139]. This will improve in the targeted analysis of molecular events such as SNVs, structural variants, and DNA methylation much more clearly. The advantages of using this method include cost effectiveness, short library preparation, use of approximately 3 µg of genomic DNA and preservation of base modifications. The design of the two gRNAs (see Chapter 2, section 2.3, Table 2.6, and Figure 2.4) heavily relied on the on-target and off-target scores, and therefore the most suitable gRNAs were chosen. The on-target score represents the Cas9 cleavage efficiency, while the Cas9 off-target score represents the off-target binding probability.

Figure 6.1 (see Chapter 6, section 6.1) is an illustration of reads mapped and aligned to chromosome 8 specifically surrounding the *DEFA1A3* region. Although most of the reads are within the *DEFA1A3* region, a subset of reads were mapped outside of that region. This suggests that the ends of reads appear to be targeted to their gRNA sites as expected. Although there are others that do not cover the expected region of interest (*DEFA1A3* CNV and its flanking regions) they start at the expected gRNA sites but go in the opposite direction to the expected location (see Chapter 6, section 6.1 and 6.2). This observation was also seen from MCF10A (human breast epithelial cells) cell line reads (18 kb long) aligned to *GSTP1* locus [139].

The construction of the *DEFA1A3* haplotypes of sample NA12878 was successful (see Chapter 6, section 6.2, Figure 6.2). The diploid *DEFA1A3* haplotypic structure is consistent with that of previous studies [2, 92]. The 3-copy allele follows the expectation of the Exchange 1 class, as it consists of 3 copies of the *DEFA1* gene with a deletion of indel5 at the centromeric (rightmost) repeat. It is also important to note that the 3 repeats contain the insertion of a 4bp variant (TATC), found within the second intron of *DEFA1A3*. In addition, the three

DEFA1 repeats contain the 7bp duplication in intron 1. The 2-copy allele follows the expectations of a Class 2 haplotype structure, and consists of 2 copies, with one copy each of *DEFA3* (telomeric repeat) and *DEFA1* (centromeric repeat). Both contain the 4bp insertion variant, while the 7bpdup variant is only found in the *DEFA3* gene. The 4bp and 7bpdup status of spatially reconstructed alleles has not been reported in previous studies, and this data adds spatial information to improve the determination of *DEFA1A3* haplotypic structures.

In conclusion, two sequencing approaches were successfully applied to investigate and construct complex tandemly repeated regions such as the *DEFA1A3* locus. This provides potential that the method can be used for other repeated regions that are highly identical and are not straightforward to assemble. CRISPR-cas9 enrichment and the long PCR sequencing are both cost-effective. However, the consequence of the long PCR sequencing is the duration of library preparation, the number of amplicons to produce per sample, and the duration of downstream analysis including the repeat sorting (`samtools+awk` command). For the repeat sorting analysis, `samtools+awk` was successfully used to separate repeats although it was not effective for some reads either due to low yield or lack of candidate SNPs to use for separation (due to high identity of repeats). Nevertheless, this can be used in future studies to automate the command to simultaneously sort repeats of each PCR into separate compartments per sample to reduce the analysis time. This may also improve the command in finding potential SNPs that may have been overlooked in this study which may possibly enhance the repeat sorting and construct the haplotype structures accurately, especially for haplotypes containing highly similar copies that are difficult to characterize such as Exchange 1 and Exchange 2 classes. Additionally, the CRISPR cas9 method seems to be the best approach in sequencing the repeated regions and preserving the true sequence without fragmentation of repeats, which will then capture the true haplotype structure and avoid the prolonged steps of repeat sorting.

Although both sequencing approaches are considered targeted and selective of the regions we are interested in, an alternative approach to selective sequencing is the ‘Read Until’ technology facilitated by Oxford Nanopore Technologies. The only difference between the Read-Until and the CRISPR cas9 enrichment approach, is that with Read-Until, you can eliminate the CRISPR step and sequence your sample immediately. The Read Until method allows selective rejection of individual reads from the flow cell prior to the completion of the read navigation through the pore. This allows the completion of the sequencing run with time being reduced massively. The approach was attempted using bacteriophage lambda DNA through target enrichment and amplicon sequencing [173]. The MinION sequencing rate when this was executed was 70 bases/s, but it is currently at 450 bases/s. In addition, this was also assessed on sample NA12878 to test its performance [174]. However, a better approach for variable regions such as the *DEFA1A3*, would be to combine the Read-Until with CRISPR-cas9 enrichment as it will be a powerful approach to targeting the region of interest. Therefore, it would be interesting to investigate how the method might affect the study of the *DEFA1A3* locus and if it improves the accuracy of the run, especially with the ongoing advances and improvements done from Oxford Nanopore Technologies.

7.4 Summary and Implications of PhD Thesis

The key findings and conclusions of this project are **A)** no association was established between total gene copy number of *DEFA1A3* and AgP from German cases vs European controls, although there was evidence of correlation between telomeric local SNPs and AgP. Perhaps, employing matched cases and controls, we may observe a different outcome. In addition, sequencing the *DEFA1A3* copy number from the German controls and Dutch samples may support the analysis. The original dataset consisted of both German and Dutch samples [102]. However, the Dutch samples were excluded due to lack of copy number information. It would also be interesting to use the CNVassoc program [167] on the current dataset to understand any discrepancies that may arise. A previous MRes student in our lab

used the software which observed a correlation with the disease. We also tested the haplotype-based analysis from centromeric and telomeric SNPs with AgP and it is unclear which position (either centromeric or telomeric haplotypes) will most likely drive the association with either AgP or *DEFA1A3* copy number. Therefore, further analysis and the inclusion of additional samples are needed to understand this phenomenon. **B)** no association was detected between total *DEFA1A3* copy number and RNA levels in Neutrophils, T-cells, or Monocytes from a UK-based population. On the contrary, a correlation was observed between relative ratios of RNA levels (*DEFA1:DEFA3*) and corresponding gene numbers. We were able to obtain an estimated copy number of *DEFA1* and *DEFA3* copies individually by using the `samtools+awk` [146] command in genomic and RNA sequencing data, and this method has enabled us to obtain the relative ratios for comparison. These current findings replicate what was previously stated in Aldred *et al.* study and from an unpublished data from our lab group. **C)** Lastly, reliable haplotype resolution can be observed by utilizing the nanopore sequencing technology, more specifically the CRISPR-Cas9 enrichment and long amplicon (tiling path) sequencing method. However, it appears that the CRISPR-cas9 approach to be the most appropriate method in classifying repeated complex structures as it requires reduced repeat sorting and reconstruction. The repeat sorting was done by utilizing the same command (`samtools+awk`) that was used in the previous study to obtain individual gene copy numbers of *DEFA1* and *DEFA3*. Separating alleles from one another was achieved to determine the likelihood of where each repeat sits in a haplotype. The estimated *DEFA1:DEFA3*, centromeric:non-centromeric and telomeric:nontelomeric ratios also helped to determine the estimated copy number. This highlights the importance of understanding tandemly repeated structures and to identify the correct construction of those alleles as they may play a role on disease or phenotype progression. It would also be interesting to test this novel method with other copy variable regions and assess its feasibility in reconstructing those regions.

REFERENCES

1. De Coster, W., et al., *NanoPack: visualizing and processing long-read sequencing data*. Bioinformatics, 2018. **34**(15): p. 2666-2669.
2. Black, H.A., et al., *Inferring mechanisms of copy number change from haplotype structures at the human DEFA1A3 locus*. BMC Genomics, 2014. **15**(1): p. 614.
3. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
4. Gibbs, R.A., et al., *The International HapMap Project*. Nature, 2003. **426**(6968): p. 789-796.
5. The Genomes Project, C., et al., *A global reference for human genetic variation*. Nature, 2015. **526**: p. 68.
6. Munz, M., et al., *A genome-wide association study identifies nucleotide variants at SIGLEC5 and DEFA1A3 as risk loci for periodontitis*. Human Molecular Genetics, 2018. **27**(5): p. 941-942.
7. Sebat, J., et al., *Strong association of de novo copy number mutations with autism*. Science, 2007. **316**(5823): p. 445-9.
8. Munz, M., et al., *Meta-analysis of genome-wide association studies of aggressive and chronic periodontitis identifies two novel risk loci*. European Journal of Human Genetics, 2019. **27**(1): p. 102-113.
9. Ai, Z., et al., *Low α -defensin gene copy number increases the risk for IgA nephropathy and renal dysfunction*. Science Translational Medicine, 2016. **8**(345): p. 345ra88-345ra88.
10. Kiryluk, K., et al., *Discovery of new risk loci for IgA nephropathy implicates genes involved in immunity against intestinal pathogens*. Nature Genetics, 2014. **46**: p. 1187.
11. Conrad, D.F., et al., *Origins and functional impact of copy number variation in the human genome*. Nature, 2010. **464**(7289): p. 704-12.
12. Sudmant, P.H., et al., *Diversity of human copy number variation and multicopy genes*. Science, 2010. **330**(6004): p. 641-6.

13. Wain, L.V., J.A. Armour, and M.D. Tobin, *Genomic copy number variation, human health, and disease*. Lancet, 2009. **374**(9686): p. 340-50.
14. Soave, A., et al., *Copy number variations in primary tumor, serum and lymph node metastasis of bladder cancer patients treated with radical cystectomy*. Sci Rep, 2020. **10**(1): p. 21562.
15. Zarrei, M., et al., *A copy number variation map of the human genome*. Nat Rev Genet, 2015. **16**(3): p. 172-83.
16. Zhao, M., et al., *Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives*. BMC Bioinformatics, 2013. **14**(11): p. S1.
17. Handsaker, R.E., et al., *Large multiallelic copy number variations in humans*. Nature Genetics, 2015. **47**(3): p. 296-303.
18. Dopman, E.B. and D.L. Hartl, *A portrait of copy-number polymorphism in Drosophila melanogaster*. Proceedings of the National Academy of Sciences, 2007. **104**(50): p. 19920-19925.
19. Emerson, J.J., et al., *Natural Selection Shapes Genome-Wide Patterns of Copy-Number Polymorphism in Drosophila melanogaster*. Science, 2008. **320**(5883): p. 1629-1631.
20. Nguyen, D.Q., et al., *Reduced purifying selection prevails over positive selection in human copy number variant evolution*. Genome Res, 2008. **18**(11): p. 1711-23.
21. Nozawa, M. and M. Nei, *Genomic drift and copy number variation of chemosensory receptor genes in humans and mice*. Cytogenetic and genome research, 2008. **123**(1-4): p. 263-269.
22. Perry, G.H., et al., *Diet and the evolution of human amylase gene copy number variation*. Nature genetics, 2007. **39**(10): p. 1256-1260.
23. Xue, Y., et al., *Adaptive Evolution of UGT2B17 Copy-Number Variation*. The American Journal of Human Genetics, 2008. **83**(3): p. 337-346.
24. Usher, C.L. and S.A. McCarroll, *Complex and multi-allelic copy number variation in human disease*. Briefings in functional genomics, 2015. **14**(5): p. 329-338.
25. Zhang, F., et al., *Copy number variation in human health, disease, and evolution*. Annu Rev Genomics Hum Genet, 2009. **10**: p. 451-81.
26. Craddock, N., et al., *Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls*. Nature, 2010. **464**(7289): p. 713-20.

27. Zhang, Y., et al., *Child development and structural variation in the human genome*. Child Dev, 2013. **84**(1): p. 34-48.
28. Weischenfeldt, J., et al., *Phenotypic impact of genomic structural variation: insights from and for human disease*. Nat Rev Genet, 2013. **14**(2): p. 125-38.
29. Lupski, J.R., et al., *Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A*. Nat Genet, 1992. **1**(1): p. 29-33.
30. Stankiewicz, P. and J.R. Lupski, *Structural variation in the human genome and its role in disease*. Annu Rev Med, 2010. **61**: p. 437-55.
31. Sekar, A., et al., *Schizophrenia risk from complex variation of complement component 4*. Nature, 2016. **530**(7589): p. 177-83.
32. Cantsilieris, S. and S.J. White, *Correlating multiallelic copy number polymorphisms with disease susceptibility*. Hum Mutat, 2013. **34**(1): p. 1-13.
33. Aitman, T.J., et al., *Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans*. Nature, 2006. **439**(7078): p. 851-5.
34. Fanciulli, M., et al., *FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity*. Nature Genetics, 2007. **39**(6): p. 721-723.
35. Willcocks, L.C., et al., *Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake*. J Exp Med, 2008. **205**(7): p. 1573-82.
36. Mueller, M., et al., *Genomic Pathology of SLE-Associated Copy-Number Variation at the FCGR2C/FCGR3B/FCGR2B Locus*. The American Journal of Human Genetics, 2013. **92**(1): p. 28-40.
37. Hollox, E.J., et al., *Psoriasis is associated with increased beta-defensin genomic copy number*. Nat Genet, 2008. **40**(1): p. 23-5.
38. Stuart, P.E., et al., *Association of β -defensin copy number and psoriasis in three cohorts of European origin*. J Invest Dermatol, 2012. **132**(10): p. 2407-2413.
39. Bentley, R.W., et al., *Association of Higher DEFB4 Genomic Copy Number With Crohn's Disease*. Official journal of the American College of Gastroenterology | ACG, 2010. **105**(2): p. 354-359.
40. Fellermann, K., et al., *A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon*. Am J Hum Genet, 2006. **79**(3): p. 439-48.

41. Aldhous, M.C., et al., *Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease*. Human Molecular Genetics, 2010. **19**(24): p. 4930-4938.
42. Gonzalez, E., et al., *The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility*. Science, 2005. **307**(5714): p. 1434-40.
43. McKinney, C., et al., *Evidence for an influence of chemokine ligand 3-like 1 CCL3L1 gene copy number on susceptibility to rheumatoid arthritis*. Annals of the Rheumatic Diseases, 2008. **67**(3): p. 409-413.
44. Aklillu, E., et al., *CCL3L1 copy number, HIV load, and immune reconstitution in sub-Saharan Africans*. BMC infectious diseases, 2013. **13**: p. 536-536.
45. Field, S.F., et al., *Experimental aspects of copy number variant assays at CCL3L1*. Nature medicine, 2009. **15**(10): p. 1115-1117.
46. Carpenter, D., et al., *Accuracy and differential bias in copy number measurement of CCL3L1 in association studies with three auto-immune disorders*. BMC genomics, 2011. **12**: p. 418-418.
47. Tang, Y.C. and A. Amon, *Gene copy-number alterations: a cost-benefit analysis*. Cell, 2013. **152**(3): p. 394-405.
48. Hollox, E.J., *The challenges of studying complex and dynamic regions of the human genome*. Methods Mol Biol, 2012. **838**: p. 187-207.
49. Cantsilieris, S., P.N. Baird, and S.J. White, *Molecular methods for genotyping complex copy number polymorphisms*. Genomics, 2013. **101**(2): p. 86-93.
50. Mazaika, E. and J. Homsy, *Digital Droplet PCR: CNV Analysis and Other Applications*. Current Protocols in Human Genetics, 2014. **82**(1): p. 7.24.1-7.24.13.
51. Carpenter, D., et al., *Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes*. Human Molecular Genetics, 2015. **24**(12): p. 3472-3480.
52. Aldred, P.M.R., E.J. Hollox, and J.A.L. Armour, *Copy number polymorphism and expression level variation of the human α -defensin genes DEFA1 and DEFA3*. Human Molecular Genetics, 2005. **14**(14): p. 2045-2052.
53. Falchi, M., et al., *Low copy number of the salivary amylase gene predisposes to obesity*. Nat Genet, 2014. **46**(5): p. 492-7.

54. Cantsilieris, S., et al., *Technical considerations for genotyping multi-allelic copy number variation (CNV), in regions of segmental duplication*. BMC genomics, 2014. **15**(1): p. 329-329.
55. Hollox, E.J., J.A. Armour, and J.C. Barber, *Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster*. Am J Hum Genet, 2003. **73**(3): p. 591-600.
56. Cantsilieris, S. and S.J. White, *Correlating Multiallelic Copy Number Polymorphisms with Disease Susceptibility*. Human Mutation, 2013. **34**(1): p. 1-13.
57. Armour, J.A.L., et al., *Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats*. Nucleic acids research, 2007. **35**(3): p. e19-e19.
58. Walker, S., S. Janyakhantikul, and J.A.L. Armour, *Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs*. Genomics, 2009. **93**(1): p. 98-103.
59. Brook, M., et al., *Neutrophil-derived alpha defensins control inflammation by inhibiting macrophage mRNA translation*. Proceedings of the National Academy of Sciences of the United States of America, 2016. **113**(16): p. 4350-4355.
60. Lehrer, R.I. and T. Ganz, *Cathelicidins: a family of endogenous antimicrobial peptides*. Curr Opin Hematol, 2002. **9**(1): p. 18-22.
61. White, S.H., W.C. Wimley, and M.E. Selsted, *Structure, function, and membrane integration of defensins*. Current Opinion in Structural Biology, 1995. **5**(4): p. 521-527.
62. Yang, D., et al., *Human neutrophil defensins selectively chemoattract naive T and immature dendritic cells*. J Leukoc Biol, 2000. **68**(1): p. 9-14.
63. Wimley, W.C., M.E. Selsted, and S.H. White, *Interactions between human defensins and lipid bilayers: evidence for formation of multimeric pores*. Protein science : a publication of the Protein Society, 1994. **3**(9): p. 1362-1373.
64. Kagan, B.L., et al., *Antimicrobial defensin peptides form voltage-dependent ion-permeable channels in planar lipid bilayer membranes*. Proc Natl Acad Sci U S A, 1990. **87**(1): p. 210-4.
65. Rice, W.G., et al., *Defensin-rich dense granules of human neutrophils*. Blood, 1987. **70**(3): p. 757-65.
66. Fujii, G., M.E. Selsted, and D. Eisenberg, *Defensins promote fusion and lysis of negatively charged membranes*. Protein Sci, 1993. **2**(8): p. 1301-12.

67. Schibli, D.J., et al., *The solution structures of the human beta-defensins lead to a better understanding of the potent bactericidal activity of HBD3 against Staphylococcus aureus*. J Biol Chem, 2002. **277**(10): p. 8279-89.
68. Lehrer, R.I., et al., *Modulation of the in vitro candidacidal activity of human neutrophil defensins by target cell metabolism and divalent cations*. The Journal of clinical investigation, 1988. **81**(6): p. 1829-1835.
69. Daher, K.A., M.E. Selsted, and R.I. Lehrer, *Direct inactivation of viruses by human granulocyte defensins*. J Virol, 1986. **60**(3): p. 1068-74.
70. Ericksen, B., et al., *Antibacterial activity and specificity of the six human {alpha}-defensins*. Antimicrob Agents Chemother, 2005. **49**(1): p. 269-75.
71. Ganz, T., et al., *Defensins. Natural peptide antibiotics of human neutrophils*. The Journal of clinical investigation, 1985. **76**(4): p. 1427-1435.
72. Miles, K., et al., *Dying and necrotic neutrophils are anti-inflammatory secondary to the release of alpha-defensins*. J Immunol, 2009. **183**(3): p. 2122-32.
73. Kohlgraf, K.G., et al., *Defensins as anti-inflammatory compounds and mucosal adjuvants*. Future microbiology, 2010. **5**(1): p. 99-113.
74. Greer, A., C. Zenobia, and R.P. Darveau, *Defensins and LL-37: a review of function in the gingival epithelium*. Periodontology 2000, 2013. **63**(1): p. 67-79.
75. Yang, D., et al., *Multiple Roles of Antimicrobial Defensins, Cathelicidins, and Eosinophil-Derived Neurotoxin in Host Defense*. Annual Review of Immunology, 2004. **22**(1): p. 181-215.
76. Yeung, A.T., S.L. Gellatly, and R.E. Hancock, *Multifunctional cationic host defence peptides and their clinical applications*. Cell Mol Life Sci, 2011. **68**(13): p. 2161-76.
77. De Smet, K. and R. Contreras, *Human Antimicrobial Peptides: Defensins, Cathelicidins and Histatins*. Biotechnology Letters, 2005. **27**(18): p. 1337-1347.
78. Linzmeier, R., et al., *A 450-kb contig of defensin genes on human chromosome 8p23*. Gene, 1999. **233**(1): p. 205-211.
79. Tran, D., et al., *Microbicidal properties and cytotoxic selectivity of rhesus macaque theta defensins*. Antimicrobial agents and chemotherapy, 2008. **52**(3): p. 944-953.

80. Garcia, A.E., et al., *Isolation, synthesis, and antimicrobial activities of naturally occurring theta-defensin isoforms from baboon leukocytes*. Infect Immun, 2008. **76**(12): p. 5883-91.
81. Tang, Y.Q., et al., *A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated alpha-defensins*. Science, 1999. **286**(5439): p. 498-502.
82. Lehrer, R.I. and W. Lu, *α -Defensins in human innate immunity*. Immunological Reviews, 2012. **245**(1): p. 84-112.
83. Ganz, T., *Defensins and Host Defense*. Science, 1999. **286**(5439): p. 420-421.
84. Amid, C., et al., *Manual annotation and analysis of the defensin gene cluster in the C57BL/6J mouse reference genome*. BMC Genomics, 2009. **10**(1): p. 606.
85. Liu, L., et al., *The human beta-defensin-1 and alpha-defensins are encoded by adjacent genes: two peptide families with differing disulfide topology share a common ancestry*. Genomics, 1997. **43**(3): p. 316-20.
86. Khan, F.F., et al., *Accurate measurement of gene copy number for human alpha-defensin DEFA1A3*. BMC Genomics, 2013. **14**(1): p. 719.
87. Ballana, E., et al., *Inter-population variability of DEFA3 gene absence: correlation with haplotype structure and population variability*. BMC Genomics, 2007. **8**(1): p. 14.
88. Daher, K.A., et al., *Isolation and characterization of human defensin cDNA clones*. Proceedings of the National Academy of Sciences, 1988. **85**(19): p. 7327.
89. Arnljots, K., et al., *Timing, targeting and sorting of azurophil granule proteins in human myeloid cells*. Leukemia, 1998. **12**(11): p. 1789-95.
90. Linzmeier, R.M. and T. Ganz, *Human defensin gene copy number polymorphisms: Comprehensive analysis of independent variation in α - and β -defensin regions at 8p22-p23*. Genomics, 2005. **86**(4): p. 423-430.
91. Jespersgaard, C., et al., *Alpha-Defensin DEFA1A3 Gene Copy Number Elevation in Danish Crohn's Disease Patients*. Digestive Diseases and Sciences, 2011. **56**(12): p. 3517-3524.
92. Tyson, J. and J.A.L. Armour, *Determination of haplotypes at structurally complex regions using emulsion haplotype fusion PCR*. BMC Genomics, 2012. **13**(1): p. 693.

93. Chen, Q., et al., *Increased genomic copy number of DEFA1/DEFA3 is associated with susceptibility to severe sepsis in Chinese Han population*. Anesthesiology, 2010. **112**(6): p. 1428-34.
94. Yu, X.-Q., et al., *A genome-wide association study in Han Chinese identifies multiple susceptibility loci for IgA nephropathy*. Vol. 44. 2011. 178-82.
95. Munz, M., et al., *A genome-wide association study identifies nucleotide variants at SIGLEC5 and DEFA1A3 as risk loci for periodontitis*. Hum Mol Genet, 2017. **26**(13): p. 2577-2588.
96. Nazir, M.A., *Prevalence of periodontal disease, its association with systemic diseases and prevention*. International journal of health sciences, 2017. **11**(2): p. 72-80.
97. Berger, J. and N. Hinglais, *[Intercapillary deposits of IgA-IgG]*. J Urol Nephrol (Paris), 1968. **74**(9): p. 694-5.
98. Sanders, J.T. and R.J. Wyatt, *IgA nephropathy and Henoch-Schönlein purpura nephritis*. Curr Opin Pediatr, 2008. **20**(2): p. 163-70.
99. Storrar, J., et al., *The epidemiology and evolution of IgA nephropathy over two decades: A single centre experience*. PLOS ONE, 2022. **17**(9): p. e0268421.
100. D'Amico, G., *Natural history of idiopathic IgA nephropathy and factors predictive of disease outcome*. Semin Nephrol, 2004. **24**(3): p. 179-96.
101. Kiryluk, K., et al., *Geographic differences in genetic susceptibility to IgA nephropathy: GWAS replication study and geospatial risk analysis*. PLoS Genet, 2012. **8**(6): p. e1002765.
102. Munz, M., et al., *A genome-wide association study identifies nucleotide variants at SIGLEC5 and DEFA1A3 as risk loci for periodontitis*. Human Molecular Genetics, 2017. **26**(13): p. 2577-2588.
103. Sanz, M., *European workshop in periodontal health and cardiovascular disease*. European Heart Journal Supplements, 2010. **12**(suppl_B): p. B2-B2.
104. Lovegrove, J.M., *Dental plaque revisited: bacteria associated with periodontal disease*. J N Z Soc Periodontol, 2004(87): p. 7-21.
105. Tsuchida, S., et al., *Ubiquitination in Periodontal Disease: A Review*. International journal of molecular sciences, 2017. **18**(7): p. 1476.
106. Masumoto, R., et al., *Identification of genetic risk factors of aggressive periodontitis using genomewide association studies in association with those of chronic periodontitis*. J Periodontal Res, 2018.

107. Schaefer, A.S., et al., *A genome-wide association study identifies GLT6D1 as a susceptibility locus for periodontitis*. Hum Mol Genet, 2010. **19**(3): p. 553-62.
108. De Maio, N., et al., *Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes*. Microbial Genomics, 2019. **5**(9).
109. Porubsky, D., et al., *Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads*. Nature Biotechnology, 2021. **39**(3): p. 302-308.
110. Amarasinghe, S.L., et al., *Opportunities and challenges in long-read sequencing data analysis*. Genome Biology, 2020. **21**(1): p. 30.
111. Cui, J., et al., *Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome*. Plant Methods, 2020. **16**(1): p. 85.
112. Lang, D., et al., *Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore*. GigaScience, 2020. **9**(12).
113. Jain, M., et al., *The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community*. Genome Biology, 2016. **17**(1): p. 239.
114. Garalde, D.R., et al., *Highly parallel direct RNA sequencing on an array of nanopores*. Nature Methods, 2018. **15**: p. 201.
115. Reuter, J.A., D.V. Spacek, and M.P. Snyder, *High-throughput sequencing technologies*. Mol Cell, 2015. **58**(4): p. 586-97.
116. Deamer, D., M. Akeson, and D. Branton, *Three decades of nanopore sequencing*. Nature Biotechnology, 2016. **34**: p. 518.
117. Deamer, D.W. and M. Akeson, *Nanopores and nucleic acids: prospects for ultrarapid sequencing*. Trends in Biotechnology, 2000. **18**(4): p. 147-151.
118. Kasianowicz, J.J., et al., *Characterization of individual polynucleotide molecules using a membrane channel*. Proceedings of the National Academy of Sciences, 1996. **93**(24): p. 13770-13773.
119. Feng, Y., et al., *Nanopore-based Fourth-generation DNA Sequencing Technology*. Genomics, Proteomics & Bioinformatics, 2015. **13**(1): p. 4-16.
120. Branton, D., et al., *The potential and challenges of nanopore sequencing*. Nature biotechnology, 2008. **26**(10): p. 1146-1153.

121. Stubbs, S., et al., *A Nanopore-based method for generating complete coding region sequences of dengue virus in resource-limited settings*. bioRxiv, 2018: p. 499111.
122. Johnson, S.S., et al., *Real-Time DNA Sequencing in the Antarctic Dry Valleys Using the Oxford Nanopore Sequencer*. Journal of biomolecular techniques : JBT, 2017. **28**(1): p. 2-7.
123. Quick, J., et al., *Real-time, portable genome sequencing for Ebola surveillance*. Nature, 2016. **530**: p. 228.
124. Faria, N.R., et al., *Mobile real-time surveillance of Zika virus in Brazil*. Genome medicine, 2016. **8**(1): p. 97-97.
125. Jain, M., et al., *Nanopore sequencing and assembly of a human genome with ultra-long reads*. Nature Biotechnology, 2018. **36**: p. 338.
126. Payne, A., et al., *Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files*. bioRxiv, 2018: p. 312256.
127. Simpson, J.T., et al., *Detecting DNA cytosine methylation using nanopore sequencing*. Nature Methods, 2017. **14**: p. 407.
128. Gabrieli, T., et al., *Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH)*. Nucleic Acids Research, 2018. **46**(14): p. e87-e87.
129. Norris, A.L., et al., *Nanopore sequencing detects structural variants in cancer*. Cancer biology & therapy, 2016. **17**(3): p. 246-253.
130. Kuderna, L.F.K., et al., *Selective single molecule sequencing and assembly of a human Y chromosome of African origin*. Nature Communications, 2019. **10**(1): p. 4.
131. Goodwin, S., et al., *Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome*. Genome research, 2015. **25**(11): p. 1750-1756.
132. Laver, T., et al., *Assessing the performance of the Oxford Nanopore Technologies MinION*. Biomolecular Detection and Quantification, 2015. **3**: p. 1-8.
133. Cretu Stancu, M., et al., *Mapping and phasing of structural variation in patient genomes using nanopore sequencing*. Nature Communications, 2017. **8**(1): p. 1326.
134. Ip, C., et al., *MinION Analysis and Reference Consortium: Phase 1 data release and analysis [version 1; referees: 2 approved]*. F1000Research, 2015. **4**(1075).

135. Tyson, J.R., et al., *Whole genome sequencing and assembly of a *Caenorhabditis elegans* genome with complex genomic rearrangements using the MinION sequencing device*. bioRxiv, 2017: p. 099143.
136. Gilpatrick, T., et al., *Targeted nanopore sequencing with Cas9-guided adapter ligation*. Nature Biotechnology, 2020. **38**(4): p. 433-438.
137. Jeck, W.R., et al., *A Nanopore Sequencing-Based Assay for Rapid Detection of Gene Fusions*. The Journal of Molecular Diagnostics, 2019. **21**(1): p. 58-69.
138. Grädel, C., et al., *Rapid and Cost-Efficient Enterovirus Genotyping from Clinical Samples Using Flongle Flow Cells*. Genes (Basel), 2019. **10**(9).
139. Gilpatrick, T., et al., *Targeted Nanopore Sequencing with Cas9 for studies of methylation, structural variants and mutations*. bioRxiv, 2019: p. 604173.
140. Kwak, W., et al., *Complete Genome of *Lactobacillus iners* KY Using Flongle Provides Insight Into the Genetic Background of Optimal Adaption to Vaginal Ecniche*. Frontiers in Microbiology, 2020. **11**.
141. Levy, S., et al., *The diploid genome sequence of an individual human*. PLoS Biol, 2007. **5**(10): p. e254.
142. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions*. Nature reviews. Genetics, 2011. **13**(1): p. 36-46.
143. Chen, L., et al., *Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells*. Cell, 2016. **167**(5): p. 1398-1414.e24.
144. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses*. Am J Hum Genet, 2007. **81**(3): p. 559-75.
145. Koch, C.M., et al., *A Beginner's Guide to Analysis of RNA Sequencing Data*. American journal of respiratory cell and molecular biology, 2018. **59**(2): p. 145-157.
146. Masella, A.P., et al., *BAMQL: a query language for extracting reads from BAM files*. BMC Bioinformatics, 2016. **17**(1): p. 305.
147. Danecek, P., et al., *Twelve years of SAMtools and BCFtools*. Gigascience, 2021. **10**(2).
148. Li, H., *Minimap2: pairwise alignment for nucleotide sequences*. Bioinformatics, 2018. **34**(18): p. 3094-3100.
149. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics, 2010. **26**(6): p. 841-2.

150. Koren, S., et al., *Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation*. Genome Res, 2017. **27**(5): p. 722-736.
151. https://community.nanoporetech.com/protocols/1d-native-barcoding-genomic-dna/checklist_example.pdf.
152. Pruim, R.J., et al., *LocusZoom: regional visualization of genome-wide association scan results*. Bioinformatics, 2010. **26**(18): p. 2336-2337.
153. Li, M., et al., *Identification of new susceptibility loci for IgA nephropathy in Han Chinese*. Nat Commun, 2015. **6**: p. 7270.
154. Kato, M., Y. Nakamura, and T. Tsunoda, *MOCSphaser: a haplotype inference tool from a mixture of copy number variation and single nucleotide polymorphism data*. Bioinformatics, 2008. **24**(14): p. 1645-6.
155. Munro, R., et al., *minoTour, real-time monitoring and analysis for nanopore sequencers*. Bioinformatics, 2021. **38**(4): p. 1133-1135.
156. Delahaye, C. and J. Nicolas, *Sequencing DNA with nanopores: Troubles and biases*. PLOS ONE, 2021. **16**(10): p. e0257521.
157. Harel, N., et al., *Direct sequencing of RNA with MinION Nanopore: detecting mutations based on associations*. Nucleic Acids Research, 2019. **47**(22): p. e148-e148.
158. Zook, J.M., et al., *A robust benchmark for detection of germline large deletions and insertions*. Nature Biotechnology, 2020. **38**(11): p. 1347-1355.
159. Zook, J.M., et al., *An open resource for accurately benchmarking small variant and reference calls*. Nature Biotechnology, 2019. **37**(5): p. 561-566.
160. Brandt, D.Y.C., et al., *Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data*. G3 (Bethesda, Md.), 2015. **5**(5): p. 931-941.
161. Zook, J.M., et al., *Extensive sequencing of seven human genomes to characterize benchmark reference materials*. Scientific Data, 2016. **3**(1): p. 160025.
162. Garrison E, M.G., *Haplotype-based variant detection from short-read sequencing*. arXiv:1207.3907 [q-bio.GN], 2012.
163. Teumer, A., et al., *Genome-wide association study of chronic periodontitis in a general German population*. J Clin Periodontol, 2013. **40**(11): p. 977-85.

164. Dommisch, H. and S. Jepsen, *Diverse functions of defensins and other antimicrobial peptides in periodontal tissues*. Periodontology 2000, 2015. **69**(1): p. 96-110.
165. Korn, J.M., et al., *Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs*. Nature Genetics, 2008. **40**(10): p. 1253-1260.
166. Hughes, T., et al., *Runaway multi-allelic copy number variation at the α -defensin locus in African and Asian populations*. Scientific Reports, 2020. **10**(1): p. 9101.
167. Subirana, I., et al., *CNVassoc: Association analysis of CNV data using R*. BMC Medical Genomics, 2011. **4**(1): p. 47.
168. Aota, K., et al., *Management of tooth extraction in a patient with ELANE gene mutation-induced cyclic neutropenia: A case report*. Medicine, 2019. **98**(39): p. e17372.
169. Ye, Y., et al., *Mutations in the ELANE gene are associated with development of periodontitis in patients with severe congenital neutropenia*. Journal of clinical immunology, 2011. **31**(6): p. 936-945.
170. Garg, S., et al., *Chromosome-scale, haplotype-resolved assembly of human genomes*. Nature Biotechnology, 2021. **39**(3): p. 309-312.
171. Jones, N.R., *Cowpea Mosaic Virus (CPMV) Sequence Analysis over the Course of an Infection in Cowpea Plant*. Point Loma Nazarene University: Wesleyan-Holiness Digital Library.
172. Leigh, D.M., C. Schefer, and C. Cornejo, *Determining the Suitability of MinION's Direct RNA and DNA Amplicon Sequencing for Viral Subtype Identification*. Viruses, 2020. **12**(8): p. 801.
173. Loose, M., S. Malla, and M. Stout, *Real-time selective sequencing using nanopore technology*. Nature Methods, 2016. **13**(9): p. 751-754.
174. Payne, A., et al., *Readfish enables targeted nanopore sequencing of gigabase-sized genomes*. Biotechnol, 2021. **39**(4): p. 442-450.