Evolution of *CCL3L1/CCL4L1* haplotypes

Somwang Janyakhantikul

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

July 2011

Abstract

CCL3L1 and *CCL4L1* are chemokine genes, located on chromosome 17q12. They are copy number variable genes which share 95% sequence identity with their non-copy number variable paralogues *CCL3* and *CCL4*. The copy number of these genes varies between populations and has been reported to be associated with phenotypes such as susceptibility to HIV infection, hepatitis C virus infection, Kawasaki disease and SLE.

The aim of this study is to understand the evolutionary history of variation at the *CCL3L1/CCL4L1* cluster. To accomplish this goal, several approaches including typing microsatellites, single nucleotide polymorphisms (SNPs) and *CCL3L1/CCL4L1* sequence haplotypes were used to investigate the association with *CCL3L1* and *CCL4L1* copy number. However, the results showed that there is no strong association between a single-copy marker and *CCL3L1* and *CCL4L1* copy number, but there is evidence of recombination. Therefore, this may suggest that *CCL3L1/CCL4L1* is a complex region and one plausible hypothesis is that there is a high rate of recombination in this region. This study of the evolution of *CCL3L1/CCL4L1* haplotypes showed that a major one-copy *CCL3L1/CCL4L1* haplotype (about 70% haplotype frequency) identified in humans, represents the ancestral state, as inferred from comparison with chimpanzee.

Acknowledgements

First of all, I would like to thank my supervisor, Prof. John A.L Armour, for giving me a chance to do a research with him, and for his valuable guidance and patience during my research. Moreover, I would like to thank him again for being a role model in an academic career. I would also like to thank Dr. Jess Tyson, who is always willing to give advice, help and support me throughout my study, and of course this is not only in the academic field but also helping me to adapt and understand UK's life; Dr. Danielle Carpenter and Dr. Susan Walker for their technical help and sharing research results. My thank also goes to all members of the C10 laboratory, Dr. Tamsin Majerus, Raquel Pala, Fayeza Khan, Sughanda Dhar, Holly Black, Dibo Pughikumo, Omniah Mansouri and the former member Ladas Ionnis for their friendship and making the lab a good working atmosphere. I would like to give my special thank to all my Thai friends as well for giving me a colourful life.
I would like to thank the Faculty of Pharmaceutical Sciences, UbonRatchathani University and Ministry of Science and Technology for funding me to study here and allowed me to gain invaluable experiences.

Lastly, I would like to thank my family for their understanding, cheering and supporting me with endless love, especially my mum who never enrolled in the education system but always convinced me to recognise the importance of the education. Definitely, she is also my best teacher.

Table of Contents

Figures

Tables

Appendix (Tables)

Abbreviations

AIDS	Acquired Immune Deficiency Syndrome
Array CGH	Array-Based Comparative Genomic Hybridisation
BSA	Bovine Serum Albumin
CCL3	Chemokine (C-C motif) Ligand 3
CCL3L1	Chemokine (C-C motif) Ligand 3-Like 1
CCL4	Chemokine (C-C motif) Ligand 4
CCL4L1	Chemokine (C-C motif) Ligand 4-Like 1
CNV	Copy Number Variation
dNTPs	Deoxyribonucleotide Triphosphates
EDTA	Ethylenediaminetetraacetic acid
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
10x "LD" buffer	10x "Low dNTP" buffer
LD	Linkage Disequilibrium
MIP	Macrophage Inflammatory Protein
PCR	Polymerase Chain Reaction
PRT	Paralogue Ratio Test
RFLP	Restriction Fragment Length Polymorphism
SNP	Single Nucleotide Polymorphism
TB	Tuberculosis
TBE	Tris Borate EDTA

Chapter 1: Introduction

1.1 Genetic variation

The success of the human genome project (International Human Genome Sequencing Consortium 2001; Venter *et al.* 2001) and the revolution in biomolecular technology has had a significant role in the discovery of human genes and human genetic diversity. Genetic variation seems to contribute to many complex human diseases such as bipolar disorder, Crohn's disease, type 1 and 2 diabetes, coronary artery disease, and rheumatoid arthritis (The Wellcome Trust Case Control Consortium 2007, 2010). Interestingly, recent studies have revealed that genetic variation in the *CCR5* gene, which encodes a protein co-receptor for the entry of human immunodeficiency virus into human immune cells, and the *CCL3L1* gene, which encodes a chemokine, might be involved in protection against HIV/AIDS (Gonzalez *et al.* 2005). These proteins function by decreasing HIV infection through their effects on both cell mediated immunity and (other) viral entry–independent mechanisms (Dolan *et al.* 2007).

Consequently, studying genetic variations that predispose or increase risk of these common diseases might lead to the development of new interventions that could have an enormous effect on medical therapy and benefit to public health.

Genetic variation can be present in many forms such as SNP, insertion/deletion variant (InDel), microsatellite, minisatellite and variable numbers of tandem repeats, multisite variant, copy number variation (CNV), inversion, translocation and unbalanced rearrangements (Feuk *et al.* 2006). However, the forms that have interested researchers most in recent years are microsatellites, single nucleotide polymorphisms (SNPs) and copy number variations (CNVs). Those due to SNPs are the prevalent form of genetic variation and contribute to much normal phenotypic variation (The International SNP Map Working Group 2001; The International HapMap Consortium 2005, 2007; The International HapMap 3 Consortium 2010). After Iafrate *et al.* (2004) and Sebat *et al.* (2004) had reported the widespread presence of copy number variation in normal individuals, several reports showed association between copy number variation and common disorders (reviewed in Estivill and Armengol (2007) and Zhang *et al.* (2009)). Although it seems microsatellites are currently less popular than SNPs and CNVs, using combined markers may provide more advantage than using either marker alone, such as increasing detail of linkage disequilibrium (LD) (Beckmann *et al.* 2007; Payseur *et al.* 2008).

1.1.1 Single nucleotide polymorphisms (SNPs)

Single nucleotide polymorphisms (SNPs) are genetic variations that occur when a single nucleotide in the genome sequence is altered, and the variant is found in more than 1% of the population (Brookes 1999).

They make up about 90% of known sequence variants in humans and are the most abundant type of genetic variation (Collins *et al.* 1998). At present, the International HapMap project has genotyped common and rare SNPs and reported the data for public use at http://www.hapmap.org (more detail is provided in section 1.2). In addition, in the near future, SNP information will be even more comprehensive on data release from the 1000 genomes project (http://www.1000genomes.org/). This is the ongoing project with the aim of providing a public resource of human genetic variation using sequencing data from 2500 samples in 25 populations from around the world. Information will be included for SNPs and structural variants which have frequencies ≥ 1% in the population studied. All SNPs validated by experimental methods from both projects are also recorded in the SNP database at http://www.ncbi.nlm.nih.gov/snp/.

Generally, SNPs have been used as genetic markers in case-control studies to find genetic associations with diseases or traits. In these studies large-scale SNP genotyping is performed in a patient group and a healthy-matched control population, and their genotypes will be compared to look for significant differences. Then, the relationship between a specific genotype and a phenotype will be identified further for susceptibility genes that are associated with a disease (Kim and Misra 2007). Some examples of association between SNP alleles and diseases are given in section 4.2.1.

1.1.2 Copy number variations (CNVs)

Recent discoveries have revealed that large segments of DNA (>100 kb) can vary in copy-number (Iafrate *et al.* 2004; Sebat *et al.* 2004). For example, genes that were thought to always occur in two copies per diploid genome have now been found to sometimes be present in one, three, or more than three copies (Figure 1.1).

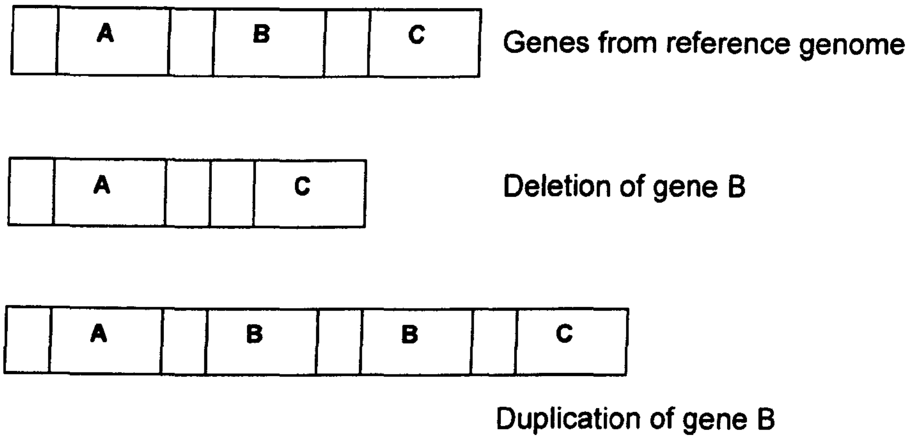


Figure 1.1: Copy number variation in the human genome.

Copy number variations (CNVs) are defined as a segment of DNA that is ≥ 1kb and shows variable copy number compared to a reference genome, including insertions, deletions and duplications (Feuk *et al.* 2006).

To date (Nov 02, 2010), it is estimated that there are 66,741 CNVs in the human genome (http://projects.tcag.ca/variation/). Recently, several studies have reported that CNVs are associated with Mendelian diseases such as mental retardation and alpha/beta thalassemia, and complex/common disorders such as HIV susceptibility, systemic autoimmune disease and schizophrenia (reviewed in Zhang *et al.* 2009).

However, although many studies of the association between CNVs and diseases have been published, and the availability of several techniques that can be used to genotype CNVs, such as array comparative genomic hybridization (arrayCGH), quantitative PCR (qPCR), fluorescent *in situ* hybridization (FISH) and next generation sequencing (NGS), a robust measurement of CNVs is still required because different techniques can provide different copy number in a CNV assessment and each technique has its own limitations for CNV genotyping (Alkan *et al.* 2011).

1.1.3 Microsatellites

Microsatellites, also known as short tandem repeats (STRs) and simple sequence repeats (SSRs) are short DNA sequences which contain tandem repeat structures of 1-4 motifs. For example, "AAAAAAA" which is referred to as $(A)_7$ and "GTGTGTGTGTGT" which is referred to as $(GT)_6$ or $G(TG)_5T$. Terms like mono-, di-, tri or tetranucleotide are often used, and these are the main types of microsatellite. However, the repeats of five (penta-) or six (hexa-) nucleotides are also classified as microsatellites (Ellegren 2004).

Microsatellites are abundantly distributed across genomes both in coding and non-coding DNA regions; they are estimated to constitute 3% of the human genome (Ellegren 2004). Microsatellites are also highly variable in the sizes of their alleles. They are also found to vary extensively in between African, Asian and European populations (Jorde *et al.* 1997).

Microsatellites are ubiquitous in the genome and show a high degree of polymorphism (informativeness). Moreover, they are convenient to detect using the polymerase chain reaction (PCR). As a result, microsatellites are popular markers for use in linkage mapping to identify susceptibility loci involved in human diseases. The diseases that microsatellites cause directly include Huntington's disease, fragile X syndrome, and myotonic dystrophy (reviewed in Cummings and Zoghbi (2000)). Microsatellite analysis is also one approach for studying complex diseases such as cancer (Ginzinger *et al.* 2000; Popat *et al.* 2005) and schizophrenia (Virgos *et al.* 2001; Bailer *et al.* 2002). In addition to their applications in genome mapping, these markers have been applied in a variety of fields such as personal identification (Thomson *et al.* 1999; Staiti *et al.* 2004), population genetic analysis and the construction of human evolutionary trees (Zhivotovsky *et al.* 2003).

1.2 The International HapMap Project

1.2.1 Background and concept

To date, there are approximately 19.7 million SNPs currently recorded in the human genome (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi), meaning if all SNPs were tested, it would be extremely expensive, laborious and time consuming.

Nevertheless, SNPs are generally found close together in blocks, called haplotype blocks (Gabriel *et al.* 2002) and the haplotypes in an individual DNA can be identified by tag-SNPs (Johnson *et al.* 2001). As a result, tag-SNPs are exploited to determine the collection of haplotypes present in each subject instead of examination all SNPs in an individual DNA. An example of the relationship between SNPs, haplotypes and tag-SNPs is shown in Figure 1.2.

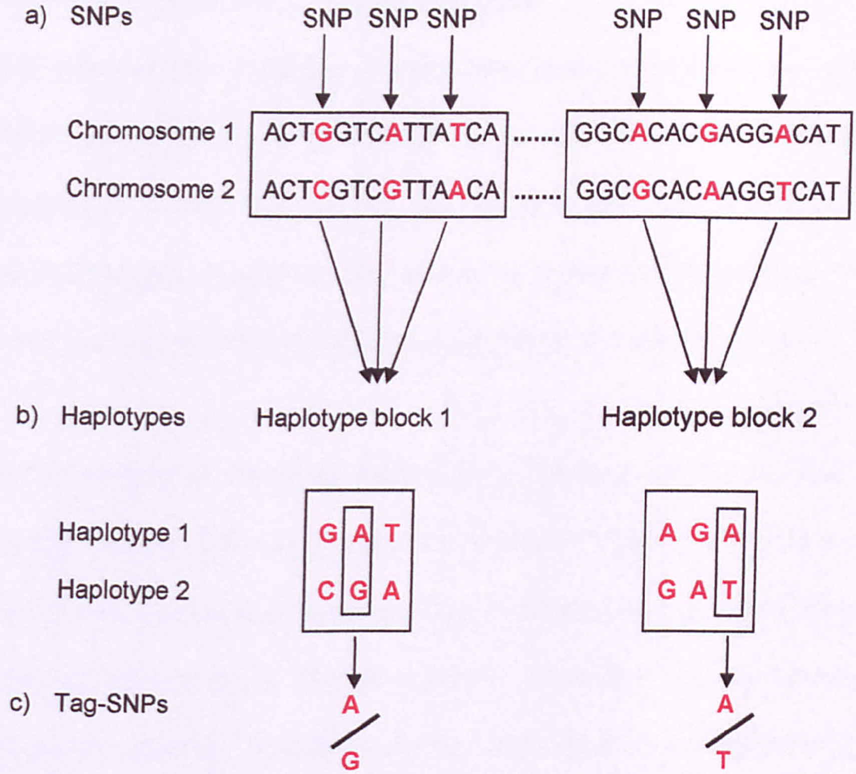


Figure 1.2: An example of SNPs, haplotypes and tag-SNPs.

a. SNPs: Most of the DNA sequences between two chromosomes are identical, but there are variant bases that have arisen in three loci between these two chromosomes.

b. Haplotypes: A set of SNPs on the chromosome that are transmitted together, and these SNPs can be found together in a block, called a haplotype block.

c. Tag-SNPs: haplotype blocks can be tagged by genotyping a SNP located in the blocks.

1.2.2 The International HapMap Project

The International HapMap Project was encouraged by two studies, Reich *et al.* (2001) and Gabriel *et al.* (2002), which showed that haplotypes across populations are shared, and only a few common haplotypes are observed. So, in theory it should be possible to use these haplotypes as a powerful tool to discover disease genes.

The International HapMap Project was founded by the collaboration among organizations in Japan, the United Kingdom, Canada, China, the United States and Nigeria. The objective was to determine the patterns of common human genetic variation, by characterizing sequence variants, their frequencies, and correlations between them. This has been a useful research tool for researchers to identify genetic factors that contribute to health, disease and drug response in human (The International HapMap Consortium 2003).

The International HapMap Project started with 269 samples from four populations: (1) 90 individuals (30 parent-offspring trios) from the Yoruba in Ibadan, Nigeria (Abbreviation YRI); (2) 90 individuals (30 trios) from Utah, USA, selected from the Centre d'Etude du Polymorphisme Humain collection (abbreviation CEU); (3) 45 Han Chinese in Beijing, China (Abbreviation CHB); (4) 44 Japanese in Tokyo, Japan (Abbreviation JPT). The sampling of human populations was selected on the basis that these populations will represent most genetic variation found in all populations throughout the world (based on their allele frequency distributions and similarity in haplotype patterns in the pilot studies) (The International HapMap Consortium 2003). The project was divided into two phases: Phase I and Phase II. However, the data from HapMap Phase I and II have limitations; for example, they do not provide information on rare variants and structural variants (Manolio and Collins 2009). As a result, The International HapMap Project continued to expand the HapMap Phase I and II data, calling this expansion HapMap 3, with the aim of providing information on rarer variants and copy number polymorphisms in populations with a wide range of ancestry. These data were just published in 2010.

1.2.2.1 The Phase I HapMap

Phase I of the International HapMap Project consisted of 2 sections: first, to genotype at least one common SNP with a minor allele frequency of 0.05 or greater every 5kb across the human genome in each of 269 DNA samples from four populations, including YRI, CEU, CHB and JPT; and second, to compare genotyping data with ten selected 500-kb regions from the ENCODE (Encyclopedia of DNA Elements) project (http://www.genome.gov/10005107).

The data from phase I HapMap showed that it contained 1,007,329 SNPs, and 17,944 SNPs discovered in the ENCODE regions (one per 279 bp). The International HapMap project also identified recombination hotspots and the block-like structure of linkage disequilibrium. Moreover, analysis of the phase I map showed that between 260,000 and 474,000 tag-SNPs are required to capture all common SNPs in the phase I data set (The International HapMap Consortium 2005).

1.2.2.2 The Phase II HapMap

A further 2.1 million SNPs were genotyped on the same individuals that were used in phase I. The data showed that there are 1.14 genotyped polymorphic SNPs per kilobase, and 0.5-1.09 million SNPs are required to capture all common Phase II SNPs with $r^2 \geq 0.8$ (The International HapMap Consortium 2007). r^2 is one of the measures of linkage disequilibrium (LD). It is the correlation coefficient of alleles at the two loci. It will range between 0 and 1; $r^2 = 1$ is perfect LD. This indicates that there is no evidence of recombination between the pair of SNPs, and also shows that the SNPs have the same allele frequency (Ardlie *et al.* 2002).

1.2.2.3 HapMap 3

HapMap 3 is an expansion of the HapMap Phase I and II resources. It combines data sets of common and rare alleles, including both SNPs and copy number polymorphisms (CNPs) by genotyping 1.6 million SNPs in 1,184 DNA samples from 11 populations and sequencing ten 100-kb regions in 692 of these DNA samples. The samples include the same four populations from HapMap Phase I and II and additional samples from seven populations: (1) Luhya in Webuye, Kenya

(Abbreviation LWK); (2) Maasai in Kinyawa, Kenya (Abbreviation MKK); (3) Tuscans in Italy (Abbreviation TSI); (4) Gujarati Indian in Houston, Texas, USA (Abbreviation GIH); (5) Denver (Colorado) metropolitan Chinese community (Abbreviation CHD); (6) people of Mexican origin in Los Angeles, California, USA (Abbreviation MXL); (7) people with African ancestry in the southwestern United States (Abbreviation ASW). In summary, HapMap 3 found that rare alleles are less shared across populations of study, even in the closely related populations (The International HapMap 3 Consortium 2010).

1.3 Linkage disequilibrium (LD) concept and haplotype blocks

1.3.1 The patterns and factors that influence linkage disequilibrium

Linkage disequilibrium (LD) is the association of alleles at different loci being found together more often than expected by chance (Figure 1.3).

Region 1 ← 100kb → Region 2

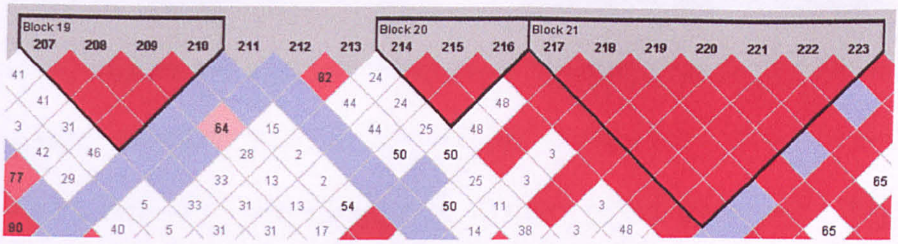
SNP A	SNP B	SNP C		SNP X	SNP Y	SNP Z
A	G	C		A	C	A
A	G	T		A	G	A
A	G	T		A	G	T
T	C	T		T	C	T
T	C	C		T	C	A
A	G	C		A	G	T
A	G	T		A	C	A
T	C	C		T	G	A
A	G	C		A	C	T
T	C	T		T	G	T

Figure 1.3: An example of linkage disequilibrium.

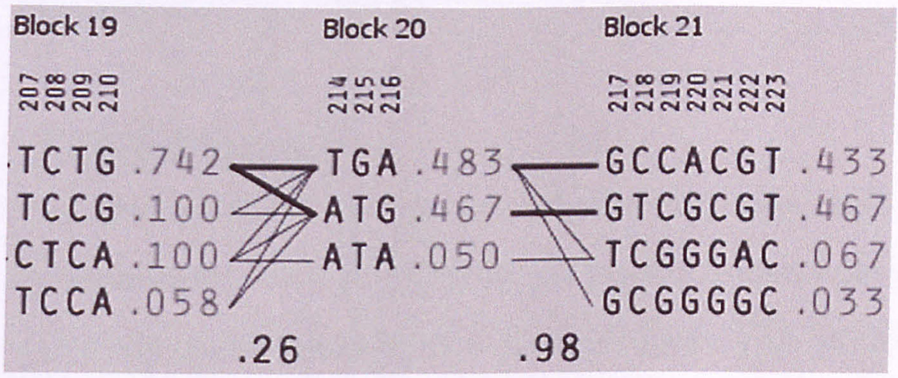
SNP A at region 1 has a perfect LD with its neighbour, SNP B, and with SNP X at region 2, although they are located a distance of 100kb apart.

However, LD will not be maintained over indefinitely long time periods. The pattern of LD will gradually break down through recombination, mostly at hot spots which are found about every 1 in 50kb across the human genome (Jeffreys *et al.* 2001; Myers *et al.* 2005). Recombination hot spots will thus separate SNPs into blocks, called haplotype blocks (Figure 1.4) (Daly *et al.* 2001; Reich *et al.* 2001; Gabriel *et al.* 2002). Although LD decays rapidly between the blocks, it will be still maintained within the block. Johnson *et al.* (2001) demonstrated that using LD mapping, which is a method to determine the non-random association of alleles at two more different loci on a chromosome, can reduce significantly the number of SNPs which are required for genotyping; in their study, common haplotypes in 9 genes of European populations could be identified with 34 SNPs from a total of 122 SNPs.

Based on the characteristics of LD, therefore, to save the cost and time of genotyping all SNPs in the block without the loss of SNP information, genotyping a set of selected SNPs (*i.e.* tag-SNPs) in the block should provide sufficient SNP information to use to identify all haplotypes in the block precisely. A further example is shown in Figure 1.4.



a)



b)

Figure 1.4: Examples of haplotype blocks, revealed by Haploview, in the CEPH-HapMap population at Chr.17: 31,889,664-31,936,690.

a) The haplotype blocks are shown outlined. The colours indicate LD in terms of D' and their statistically significant evidence (log odds; LOD) between pairs of SNPs. In brief, the white colour indicates low LD (D' <1) with low statistical significance (LOD <2), the blue colour shows perfect LD (D'=1) with low statistical significance (LOD <2), and red colour refers to perfect LD (D'=1) with high statistical significance (LOD >2), but if the LD is low (D' <1) shades of pink will be used instead. D' is a type of statistics for measuring LD. Similar to r^2, the range of D' is between 0 and 1; D' = 1 is perfect LD and indicates that two SNPs have not been separated by recombination.

b) Haplotypes and their frequencies in each block.

In block 21, for example there are 4 distinct haplotypes which were generated from 7 SNPs. Instead of typing all 7 SNPs to identify these haplotypes, genotyping just 2 SNPs, SNP 218 and SNP 219, can reveal two major haplotypes (89% of the population), and if SNP 217 is included, all haplotypes will be disclosed.

However, besides recombination which has an important impact on LD, many factors are also can influence to the pattern of LD. For example, natural selection and genetic drift will increase LD between closely linked loci, and population subdivision, population bottlenecks and inbreeding can affect LD across the genome, in general resulting in increased LD (Slatkin 2008).

1.3.2 Applications of LD

The characteristics of LD can be used to map disease genes (or traits/phenotypes). This is because disease-causing alleles might also have a non-random association with a genetic marker such as a SNP. Many studies have now shown that LD can be applied to identify disease genes and trait-associated alleles. Examples of such successful studies are the discovery of a suspect region for the diastrophic dysplasia gene (Hästbacka *et al.* 1992) and finding a candidate region in Huntington's disease (MacDonald *et al.* 1992).

Therefore, it is proposed that LD could be of benefit in genome-wide LD mapping to find common disease alleles and in studying population history by using a dense map of common SNP markers.

Currently, with the availability of massive numbers of SNP markers from the HapMap (more details are described in 1.2), LD has contributed an important tool in genome wide association studies (GWAS) to find common alleles or genes which underlie common diseases, with the primary aim of the HapMap to select a set of optimal tag-SNPs (as described above). The examples of LD mapping to find common disease genes in GWAS include risk factors for type 2 diabetes in Finns (Scott *et al.* 2007), GWAS in Parkinson's disease (Satake *et al.* 2009; Simón-Sánchez *et al.* 2009) and GWAS to identify genes for biomarkers of cardiovascular disease (Wallace *et al.* 2008).

1.4 *CCL3L1* and *CCL4L1*

1.4.1 The characteristics of *CCL3, CCL4, CCL3L1 and CCL4L1*

CCL3, CCL4, CCL3L1 and *CCL4L1* are chemokine genes located at chromosome 17q12. *CCL3* and *CCL4* are the common ancestral genes, and in the reference genome assembly both have a second non-allelic copy, *CCL3L1* and *CCL4L1*, originating by segmental duplication. Next to them, can also be found a *CCL3L1* 5′-truncated pseudogene (Hirashima *et al.* 1992; Modi 2004). The map of these genes is shown in Figure 1.5.

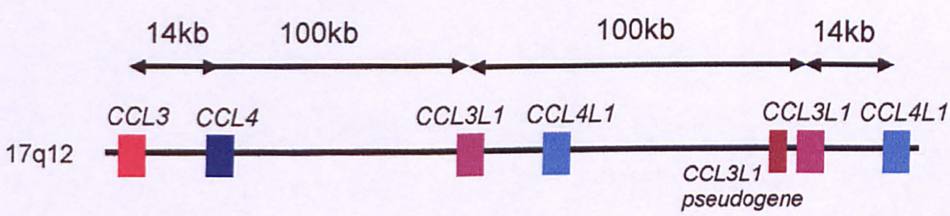


Figure 1.5: Map of *CCL3, CCL4, CCL3L1, CCL4L1* and *CCL3L1* pseudogene at chromosome 17q12 [UCSC:Mar.2006 (NCBI36/HG18)].

CCL3L1 and *CCL4L1* are copy number variable genes. They are highly similar to their non-copy number variable paralogues *CCL3* and *CCL4* with each pair sharing 95% sequence identity at both the genomic and the amino acid levels (Nakao *et al.* 1990; Modi 2004).

CCL3 and CCL4 are the abbreviated forms of chemokine (C-C motif) ligand 3, and chemokine (C-C motif) ligand 4, respectively. Similar to CCL3 and CCL4, CCL3L1 is the abbreviated form of chemokine (C-C motif) ligand 3-like 1, and CCL4L1 is the abbreviated form of chemokine (C-C motif) ligand 4-like 1. Lastly, CCL3L1 pseudogene or CCL3L2 is the abbreviated form of chemokine (C-C motif) ligand 3-like 2. These chemokines also have alternative names as shown in Table 1.1.

Table 1.1: The alternative names of CCL3, CCL4, CCL3L1, CCL4L1 and CCL3L2.

Chemokine	Alternative name
CCL3	MIP1A, SCYA3, G0S19-1, LD78ALPHA and MIP-1-alpha
CCL4	ACT2, G-26, LAG1, MIP1B, SCYA2, SCYA4, AT744.1, MGC104418, MGC126025, MGC126026 and MIP-1-beta
CCL3L1	LD78, 464.2, CCL3L3, MIP1AP, SCYA3L, G0S19-2, CYA3L1, D17S1718, LD78BETA, MGC12815, MGC104178 and MGC182017
CCL4L1	LAG1, CCL4L, LAG-1, CCL4L2, SCYA4L and AT744.2
CCL3L2	SCYA3L2 and LD78gamma

1.4.2 The association between genetic variation at *CCL3, CCL4* and diseases

Many studies have indicated that polymorphisms in *CCL3* and *CCL4* might contribute to diseases. For example, a diversity of polymorphisms of *CCL3* and *CCL4* have been found to be associated with HIV infection susceptibility and progression (Modi *et al.* 2006). Furthermore, Vyshkina and Kalman (2006) found association of haplotype 278A-277T with multiple sclerosis in the *CCL3* gene.

1.4.3 The association between genetic variation at *CCL3L1, CCL4L1* and diseases

Since the variation of *CCL3L1* copy number was found, the relationship between this gene and diseases in which the gene might play a role has been examined. One of the most interesting diseases is HIV infection; *CCL3L1* copy number has been now proposed to be a genetic determinant of HIV infection (Gonzalez *et al.* 2005; Shostakovich-Koretskaya *et al.* 2009). However, there are still some controversial issues related to the association between *CCL3L1* and HIV infection such as how to measure *CCL3L1* copy number accurately. More details are described in section 1.4.4.

In addition, the other diseases in which copy number variation of *CCL3L1* has been implicated are systemic lupus erythematosus (SLE) (Mamtani *et al.* 2008), Kawasaki disease (Burns *et al.* 2005; Mamtani *et al.* 2010) and chronic hepatitis C (Grünhage *et al.* 2010).

Like *CCL3L1*, *CCL4L1* has been found to be associated with HIV infection (Colobran *et al.* 2005). Recently, it has been found that high copy number of *CCL4L1* is associated with acute rejection in lung transplantation (Colobran *et al.* 2009).

1.4.4 The association between *CCL3L1* and HIV

1.4.4.1 HIV background

The identification of genes that are associated with HIV-1 infection and AIDS progression has had a lot of attention after "elite controllers" and "long-term non progressors" were characterized. These patients show resistance to HIV/AIDS; they have low viral load (less than 50 copies mL^{-1} in elite controllers and less than 10,000 copies mL^{-1} in long-term non progressors) despite the absence of antiretroviral therapy (Deeks and Walker 2007; Piacentini *et al.* 2009). It is these patients that show low susceptibility to HIV-1 infection and AIDS progression due to inter-individual variability. Consequently, numerous studies in genetic association analyses of AIDS have shown that there are several genetic factors associated with HIV-1 infection and AIDS progression, such as the HLA system (HLA-B*27 and HLA-B*57 alleles (Goulder *et al.* 1997; Migueles *et al.* 2000; Bailey *et al.* 2008), HLA-C allele (Fellay *et al.* 2009; Thomas *et al.* 2009)), chemokines and chemokine receptors (CCL3L1(Gonzalez *et al.* 2005; Nakajima *et al.* 2007; Shostakovich-Koretskaya *et al.* 2009), CCR5 (Martin *et al.* 1998; Mummidi *et al.* 1998)) and others (Javanbakht *et al.* 2006; Sobti *et al.* 2010).

1.4.4.2 CCL3L1 and CCR5 receptor

CCR5 is a chemokine receptor which plays a role in the immune system and has been found to be important in AIDS, transplantation and cancer (Allen *et al.* 2007). Many genetic variants within the *CCR5* region have been studied, since this receptor was discovered as a major co-receptor for HIV-1 entry to the host cell (Berger *et al.* 1999; Lederman *et al.* 2006). However, the most investigated genetic variant of *CCR5* is the loss of function allele *CCR5*-Δ32, in which 32 bp of the coding region of *CCR5* are deleted, resulting in a truncated structure that no longer functions as a chemokine receptor on the cell surface (Silva and Stumpf 2004; Arenzana-Seisdedos and Parmentier 2006). This allele, in particular in its homozygous form, is a rare variant in the world population (*i.e.* Homozygosity was found at a 1% frequency in the European population only and was not found in Asian and African) (Martinson *et al.* 1997). The homozygous carriers of this mutant allele show HIV-1 protection, although they are not completely resistant to infection (Carrington *et al.* 1999; Sheppard *et al.* 2002). However, there is a patient, who has acute myeloid leukemia and HIV-1 infection, who was transplanted with stem cells from a *CCR5*-Δ32 homozygous donor. This patient has still not shown viral rebound for 20 months after transplantation, despite discontinuation of antiretroviral therapy (Hutter *et al.* 2009). The heterozygous carriers of this allele also show slower progression to AIDS (Carrington *et al.* 1999). In addition, polymorphism in the *CCR5* promoter region has been found to be associated with HIV-1 infection and progression of AIDS, but many of these variants are rare and restricted to specific populations (Arenzana-Seisdedos and Parmentier 2006).

CCR5 is bound and activated by several ligands, including CCL3, CCL4, CCL5, CCL8, CCL11, CCL14 and CCL16, also known as MIP-1α, MIP-1β, RANTES, MCP-2, Eotaxin, HCC-1, HCC-4, respectively (Allen *et al.* 2007). However, the most potent ligand which may also have an important role in HIV-1 infection/AIDS is CCL3L1 (Nibbs *et al.* 1999). As a result of its copy number variation, the dose of CCL3L1 binding to CCR5 might have an effect on the susceptibility of HIV-1 infection and AIDS progression.

1.4.4.3 *CCL3L1* copy number variation and HIV/AIDS susceptibility

Townson *et al.* (2002) initially showed that there are variable numbers of *CCL3L1* and *CCL4L1* in the human genome. Moreover, they hypothesized that copy number variation in *CCL3L1* may affect the susceptibility to, or the progression or severity of, disease in which this chemokine plays a role. This hypothesis is supported by later studies. For example, Gonzalez *et al.* (2005) demonstrated that there is copy number variation of *CCL3L1* between individuals and between populations. They also investigated the relationship between copy number variation of *CCL3L1* and susceptibility to HIV/AIDS. The study showed that possession of a *CCL3L1* copy number lower than the population average is associated with increasing susceptibility to HIV-1 infection and development of AIDS. Nakajima *et al.* (2007) performed a study investigating *CCL3L1* copy number in long term HIV-1 infected individuals with haemophilia. They found that the average copy number of *CCL3L1* in HIV-1 infected patients was significantly lower than in non-HIV-1 infected subjects. However, they did not find the relationship between the variation of *CCL3L1* copy number and AIDS progression. Shostakovich-Koretskaya *et al.* (2009) studied the relationship between *CCL3L1* and *CCL4L1* copy number and HIV/AIDS susceptibility in Ukrainian children. They also confirmed that the low copy number of *CCL3L1* is related to HIV-1 infection and AIDS progression. Moreover, they demonstrated that *CCL4L1* plays a role in HIV-AIDS susceptibility, in addition to *CCL3L1*, and suggested that the variation in combined copy number of *CCL3L1* and *CCL4L1* may determine the susceptibility to HIV/AIDS.

The variation of copy number of *CCL3L1* is not restricted to humans, but it also has been found in macaques. Moreover, the variation of *CCL3L1* copy number in macaques has been discovered to be associated with SIV progression rate, which is similar to the relationship between *CCL3L1* and HIV-1 infection/AIDS progression in humans, and monkeys with lower *CCL3L1* copy number have a more rapid progression of simian-AIDS than those with higher *CCL3L1* copy number (n=57) (Degenhardt *et al.* 2009).

The relationship and mechanism between CCL3L1 and CCR5's ability to inhibit HIV-1 infection and/or to slow the progression of AIDS is still unclear but many proposals have been put forward (Mackay 2005).

One possible hypothesis is that CCL3L1 binds to CCR5 and blocks the binding site for HIV-1. The second possibility is CCL3L1 might affect CCR5 expression by down regulation of CCR5 from leukocytes. Lastly, CCL3L1 might cause quantitative or qualitative changes in leukocyte recruitment. However, Dolan *et al.* (2007) demonstrated that CCL3L1 and CCR5 variations influence cell-mediated immunity (CMI) and affect HIV pathogenesis via viral entry-independent mechanisms.

Although the exact mechanism by which *CCL3L1* copy number and *CCR5* genotype respectively may act against HIV-1 infection and AIDS progression is still unknown, Gonzalez *et al.* (2005) investigated risk of HIV-1 infection and rates of AIDS progression by grouping individuals based on *CCR5* genotype and *CCL3L1* copy number. They created three genetic risk groups; high risk group (low *CCL3L1* copy number and detrimental *CCR5* variations), low risk group (high *CCL3L1* copy number and non-detrimental *CCR5* variations) and moderate risk group (high *CCL3L1* and detrimental *CCR5* variations or low *CCL3L1* copy number and non-detrimental *CCR5* variations). Kulkarni *et al.* (2008) demonstrated that these genetic risk stratifications might be useful in improving the assessment of AIDS risk in HIV-1 infected individuals. They claimed that the combination of the laboratory markers (such as CD4 cell count and viral load) and *CCL3L1-CCR5* genetic markers will provide more prognostic information in HIV/AIDS than either marker alone.

Recently, two independent studies that used a modified real-time PCR assay, failed to replicate the results that *CCL3L1* copy number is associated with HIV-1 acquisition, viral load and AIDS progression (Bhattacharya *et al.* 2009; Urban *et al.* 2009). Furthermore, a study that compared two assays, PRT and real-time PCR, to measure *CCL3L1* gene copy number (Field *et al.* 2009). This study suspected that real-time PCR assays might provide an over-estimate of *CCL3L1* copy number. In reply, He *et al.* (2009) explained that using a different label in the real-time PCR assay might explain the different results and suggested that it is important to count the *CCL3L1* pseudogene, also known as *CCL3L2,* to give the correct copy number for *CCL3L1*. This issue of inconsistent copy number measurement is not limited to humans. Perry *et al.* (2008) used array CGH and found that chimpanzees have two copies of *CCL3L1* per diploid genome, whilst Degenhardt *et al.* (2009) used real-time PCR and found that *CCL3L1* is a copy number variable gene in chimpanzees with a mean copy number at 14.58.

Understanding how to measure the copy number of the *CCL3L1* gene accurately is essential to be able to investigate the relationship between this gene and HIV/AIDS.

1.5 The objectives of the study

The main objectives of this research are as follows:

- To understand the structure of the variable *CCL3L1/CCL4L1* region by investigating variants in this region. Also, to understand the evolutionary history of variation at the *CCL3L1/CCL4L1* cluster.
- To develop an assay for the presence of the *CCL3L1* pseudogene, and to investigate European (ECACC and CEPH HapMap), Asian (CHB/JPT) and African populations (YRI) for the presence of this pseudogene.
- To investigate an association between presence/absence of the *CCL3L1* pseudogene and *CCL3L1/CCL4L1* copy number in European (ECACC and CEPH HapMap), Asian (CHB/JPT) and African populations (YRI).
- To develop assays for known variants within the repeat, and for variable sites flanking the region of *CCL3L1/CCL4L1*. Then, use these single SNP markers and *CCL3* microsatellite to examine for associations with *CCL3L1/CCL4L1* copy number in European (ECACC and CEPH HapMap), Asian (CHB/JPT) and African populations (YRI), and investigate again for SNPs that could potentially predict *CCL3L1/CCL4L1* copy number.

- To sequence *CCL3L1/CCL4L1* and define the haplotype background of each of *CCL3L1/CCL4L1* copy. After that, look for SNPs or haplotype blocks using LD analysis, and, again examine their potential in prediction of *CCL3L1/CCL4L1* copy number.
- To investigate association between *CCL3L1/CCL4L1* copy number and tuberculosis.

Chapter 2: Materials and Methods

2.1 Genomic DNA samples

2.1.1 ECACC Human Random Control (HRC) samples

In this study, ECACC HRC DNA Panel 1 and 2 were purchased from the Health Protection Agency Culture Collections (http://www.hpacultures.org.uk/). This DNA was extracted from randomly selected unrelated UK Caucasian blood donors.

2.1.2 HapMap samples

All HapMap samples which were used in this study (including CEPH HapMap, CHB/JPT and YRI samples) were obtained from the Coriell Institute for Medical Research (http://ccr.coriell.org/Sections/Collections/NHGRI/hapmap.aspx?PgId=266&coll=HG). The CEPH HapMap and YRI samples comprise 30 parent-offspring trios and CHB/JPT samples include 45 unrelated Japanese from Tokyo and 45 unrelated Han Chinese from Beijing.

2.1.3 CEPH samples

A number of CEPH samples from the Centre de'Etude du Polymorphisme Humain (CEPH) (http://www.cephb.fr/) collection were specifically selected to type their *CCL3L1/CCL4L1* copy number, and to deduce unambiguously their *CCL3L1/CCL4L1* copy number haplotype in which microsatellite segregation analysis was uninformative. The selected CEPH families and their number in the pedigree are shown in Table 2.1.

Table 2.1: The list of selected CEPH families and their individual number in the pedigree.

CEPH family number	Number in the pedigree
1341	3,4,5,6,8
1344	2,7,8,9
1345	1,4,5,9,10,11
1347	1,10,11,16
1350	5,6
1362	4,5,7
1408	3,8,4

2.1.4 Basque samples

The origin and DNA extraction method of Basque individual DNA samples were described in Alonso and Armour (1998).

2.1.5 Tuberculosis (TB) case-control analysis samples

The DNA from Tuberculosis (TB) case-control analysis samples were kindly supplied by our collaborators, Dr. Jon Goulding and Prof. Mike Levin (Imperial College, London). Briefly, These samples are TB paediatrics-matched controls collected from African population (*i.e.* !Xhosa); all samples were diagnosed for TB including medical history, a physical examination, a chest X-ray, microbiological examination and Mantoux test (personal communication).

2.2 Polymerase Chain Reaction (PCR)

2.2.1 Primer design

PCR primers were designed using the UCSC genome browser March 2006 (NCBI36/HG18) assembly (http://genome.ucsc.edu/index.html?org=Human&db=hg18&hgsid=184944801) using the Primer 3 programme (http://frodo.wi.mit.edu/primer3/). Suggested PCR primers were checked using *in-silico* PCR available on UCSC genome browser to assure that the PCR primers were specific and produced only the target products. The sequences of PCR primers were also checked to confirm that they did not include common substitution variants by searching BLAST Human Sequences from NCBI (http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=9606).

2.2.2 10x "LD" PCR buffer and 10x PCR buffer

10x "LD" PCR buffer OR 10x PCR buffer	2	µl
10 µM forward primer	1	µl
10 µM reverse primer	1	µl
5U/µl *Taq* DNA polymerase	0.2	µl
H_2O	14.8	µl
10 ng/µl DNA	1	µl
Total	20	µl

Unless stated otherwise, all PCR products were amplified using the above mixture with 10x "LD" PCR buffer with the exception of *CCL3L1* and *CCL4L1* products which utilised 10x PCR buffer.

The 10x "LD" PCR buffer consists of final concentrations of 50mM Tris-HCl pH8.8, 12.5 mM $(NH_4)_2SO_4$, 1.4 mM $MgCl_2$, 7.5 mM 2-mercaptoethanol, 125 µg/ml Bovine serum albumin (BSA) and 200 µM of each dNTP. The 10x PCR buffer consists of final concentrations of 50mM Tris-HCl pH8.8, 12 mM $(NH_4)_2SO_4$, 5 mM $MgCl_2$, 7.4 mM 2-mercaptoethanol, 125 µg/ml Bovine serum albumin (BSA) and 1.1 mM each dNTP.

2.2.3 Primers and PCR conditions for *CCL3L1* pseudogene assay

The primers that were used in this assay are shown in Table 2.2, and the amplification protocol for this reaction was 5 minutes of initial denaturation at 95°C, followed by 37 amplification cycles (1 minute of denaturation at 95°C, 1 minute of annealing at 60°C and 1 minute for extension at 70°C).

Table 2.2: *CCL3L1* pseudogene assay primers.

Primer name	Primer sequence (5′-3′)
CCL3F	TGGCTGCTCGTCTCAAAGTA
CCL3R	AATTCCCTGAAGAGAACTGAGA
CCL3PR	GTGTGCAAGGACAATGCAAG

2.2.4 Primers and PCR conditions for *CCL3L1* and *CCL4L1* sequencing analysis

The primers and conditions used in *CCL3L1* and *CCL4L1* sequencing analysis are shown in Tables 2.3 to Table 2.5.

Table 2.3: *CCL3L1* PCR amplification and sequencing primers.

Primer name	Sequence (5′-3′)
CCL3L1F1	TCCTAGAGCGTGCATATTACGA
CCL3L1F2	CACACGCATGTTCCCAAG
CCL3L1F3	GAGGTGAGCAGGAAGACTGG
CCL3L1F4	AGGGTGAGCTGGAGAGTGAA
CCL3L1R	AAAGAGGAGAGATGGCTTCAGA

Table 2.4: *CCL4L1* PCR amplification sequencing primers.

Primer name	Sequence (5′-3′)
CCL4L1F1	CCTCCTTTTTAAAGGCATTTTT
CCL4L1F2	GACAGGAACTGCGGAGAGG
CCL4L1F3	TCCATATCTCACGGGACCT
CCL4L1F4	CCATTCCCACCTAACATGAG
CCL4L1F5	CAGTCACGCAGAGCTTCAT
CCL4L1R	CTGGTTCCCGCTTTGTTCT

Table 2.5: The conditions of *CCL3L1* and *CCL4L1* PCR amplification.

PCR name	Denaturing		Annealing		Extension		Cycles
	Temp. (°C)	Time (s)	Temp. (°C)	Time (s)	Temp. (°C)	Time (s)	
CCL3L1	95	30	57	30	70	180	40
CCL4L1*	95	30	56	30	70	180	40

Note: *Pre-denaturing at 95°C for 5 minutes was used before the cycles of PCR.

2.2.5 Primers and PCR conditions for testing of somatic variation and mixed samples

The primers and conditions used in this assay were modified from Kimpton *et al.* (1996). The primers are shown in Table 2.6, and the primers used fluorescent dyes as indicated on their primer names.

A PCR was performed in 1x LD buffer using 5 ng genomic DNA and 0.5 U *Taq* DNA polymerase in a total volume of 10 μl. Products were separately amplified with 1 μM each of FAM-D18S51F and D18S51R, or D21S11F and FAM-D21S11R for 26 cycles of 93°C for 30 seconds, 58°C for 1 minute and 15 seconds, 72°C for 15 seconds followed by a final hold at 72°C for 10 minutes.

Table 2.6: Microsatellite analysis primers (testing for somatic variation and mixed samples).

Primer name	**Primer sequence (5′-3′)**
FAM-D18S51F	CAAACCCGACTACCAGCAAC
D18S51R	GAGCCATGTTCATGCCACTG
D21S11F	ATATGTGAGTCAATTCCCCAAG
FAM-D21S11R	TGTATTAGTCAATGTTCTCCA

Note: FAM is the abbreviated form of Fluorescein amidite and used as fluorescein-labeled oligonucleotide probe.

2.3 DNA Sequencing

PCR products were purified using AMPure® XP (Agencourt®) according to the manufacturer's protocol prior to sequencing with Big Dye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems™) as the following mixture:

Big Dye® Terminator v3.1 Cycle Sequencing	1	µl
5x Sequencing buffer*	2	µl
10 µM primer	0.5	µl
H_2O	4.5	µl
10-20 ng/µl DNA	2	µl
Total	10	µl

*The sequencing buffer is composed of 250 mM Tris-HCl pH 9 and 10 mM $MgCl_2$.

Sequencing reactions were carried out as follows: 25 cycles of 96°C for 30 seconds, 50°C for 15 seconds and 60°C for 4 minutes, after which sequenced products were cleaned to remove unincorporated dyes using CleanSEQ® (Agencourt®) following the manufacturer's instructions. Lastly, all PCR products were sent to DBS genomics, School of Biological and Biomedical Sciences, Durham University for capillary electrophoresis and data collection on ABI 3730 DNA Analyser.

2.4. SNP assays

The SNPs that were investigated can be divided into three categories: SNPs within the *CCL3* region, SNPs at *CCL3L1* and *CCL4L1*, and additional SNPs within chromosome 17q12. The list of SNPs in each group is shown in Tables 2.7-2.9 and Figures 2.1-2.3, respectively.

SNPs within the *CCL3* region were selected from the HapMap database (http://hapmap.ncbi.nlm.nih.gov/) using the following criteria:

- They are located within the *CCL3* region (Chr.17:31,435,952-31,444,611).
- They are not located in a repeat sequence such as long terminal repeat (LTR).
- They have an average heterozygosity greater than 0.35.
- They are not associated with another chosen SNP which were also used in the SNP assays.

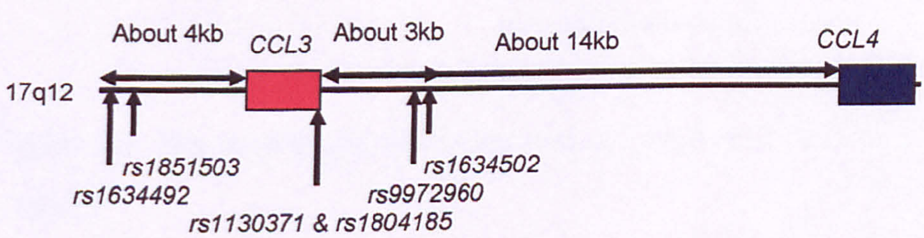


Figure 2.1: The positions of SNPs that were examined within the *CCL3* region.

SNPs at *CCL3L1* and *CCL4L1* were selected from UCSC Genome Browser March 2006 assembly (http://genome.ucsc.edu/) using the following criteria:

- They are located in the *CCL3L1* (Chr.17: 31,546,382-31,548,269 or 31,647,956-31,649,843) and *CCL4L1* (Chr.17: 31,562,581-31,564,387 or 31,664,147-31,665,959).
- They are not located in a constant region.

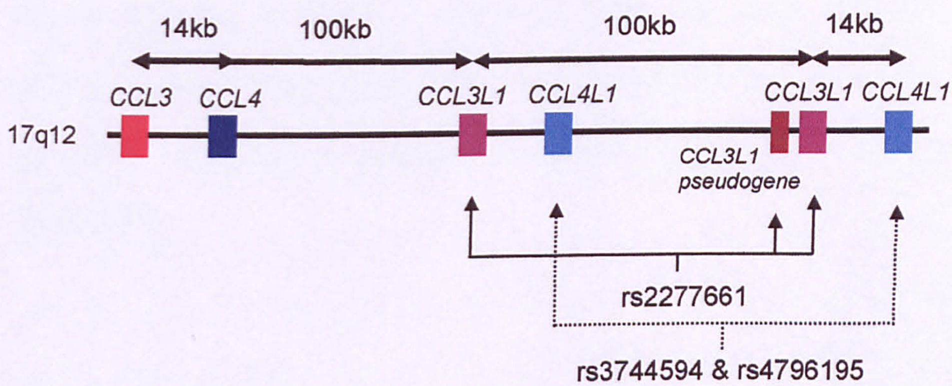


Figure 2.2: The positions of SNPs that were examined at *CCL3L1* and *CCL4L1*.

The additional SNP assays within chromosome 17q12 were selected from linkage disequilibrium analysis results that showed these SNPs might have an association with *CCL3L1/CCL4L1* copy number.

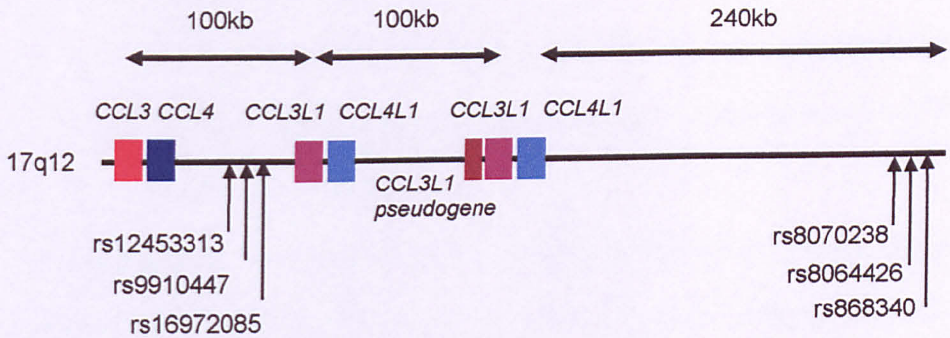


Figure 2.3: The positions of additional SNPs that were examined within the Chr.17q12 (These SNPs were suggested by Dr. Danielle Carpenter (personal communication), except SNP rs12453313 and rs9910447).

After all SNPs had been selected, the PCR primers were designed, to have the following characteristics using Primer3 software programme (http://frodo.wi.mit.edu/).

- The primers should not include a known SNP position themselves.
- The sequence of primers should be designed to discriminate against their paralogous sequence to avoid co-amplification.

For example, the SNP assay for rs1634492 within *CCL3* region has a similar sequence to its paralogue, *CCL3L1*. Therefore, the primers for this SNP should specifically amplify *CCL3* only and not *CCL3L1*. As a result, the primers were designed to have mismatches with the *CCL3L1* consensus sequence in both forward and reverse primers. An alignment of *CCL3* with *CCL3L1* is shown in Figure 2.4. Moreover, the PCR products were checked to confirm the products were *CCL3* and not from *CCL3L1* using the restriction enzymes, *Bsr*GI and *Tai*I. This PCR product is 365bp. The *CCL3* product will be digested into 29bp + 336bp fragments. If the products derived from *CCL3L1*, the product will be cut into 191bp + 174bp fragments.

All PCR products were pre-tested to confirm the identity of the product by digestion with diagnostic restriction enzymes as mentioned above. Then, all SNPs were genotyped by PCR-RFLP technique.

```
                                   BsrGI
CCL3     tccctgcagaaaatgaacaa.ggattcatgt.acactggcaggtactggc CCL3
>>>>>>>> ||| |||||||||||||||| |||||||||| |||||||||||||||||| >>>>>>>>
CCL3L1   tccttgcagaaaatgaacaaaggattcatgtaacactggcaggtactggc CCL3L1

CCL3     agccacccagggcctctcacaggaaagggagatcagaaagagaagcaaag CCL3
>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>
CCL3L1   agccacccagggcctctcacaggaaagggagatcagaaagagaagcaaag CCL3L1

CCL3     aggactcatgagataccacagggccgctgagtccagccttgcctggagct CCL3
>>>>>>>> |||||||||||||||||| ||||| |||| |||||||||||||||||||| >>>>>>>>
CCL3L1   aggactcatgagataccatagggctgctgcgtccagccttgcctggagct CCL3L1

CCL3     agggccacctcgatgccctatagtcttggagccacaaggtgcatttactc CCL3
>>>>>>>> ||||||||||||||||||||||||||||||||||||| |||||||||||| >>>>>>>>
CCL3L1   agggccacctcgatgccctatagtcttggagccacaacgtgcatttactc CCL3L1
                                             TaiI
CCL3     aaagcctctttgagtttggtttgcttgtttgctttctgcctggaaactgc CCL3
>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>
CCL3L1   aaagcctctttgagtttggtttgcttgtttgctttctgcctggaaactgc CCL3L1

CCL3     cagcatcctgagagatacgagatctgcatctgtgcagagacacagggttt CCL3
>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>
CCL3L1   cagcatcctgagagatacgagatctgcatctgtgcagagacacagggttt CCL3L1

CCL3     gttaaaagtcacaggccctgactgaagtgtggaactggctgaaatgagaa CCL3
>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>
CCL3L1   gttaaaagtcacaggccctgactgaagtgtggaactggctgaaatgagaa CCL3L1

CCL3     agtaggaggtaatttgg CCL3
>>>>>>>> |||    |||||||||| >>>>>>>>
CCL3L1   agt....ggtaatttgg CCL3L1
```

Figure 2.4: Alignment of *CCL3* with *CCL3L1* for designing the primers at SNP rs1634492.

The primers are shown in blue and the restriction sites for *Bsr*GI and *Tai*I are shown in red.

The assay for rs1634492 will be described in detail below as an example of a typical method for SNP typing. Most SNP assays were carried out in a similar way to SNP rs1634492 assay, except that the SNP assay for rs1130371 and rs1804185 used one restriction enzyme to assay two SNPs simultaneously.

All information for the SNP assays such as name of SNPs, the positions in the genome, primers that were used, PCR product sizes, and PCR conditions are shown in Tables 2.7-2.10.

2.4.1 SNP assay for rs1634492

PCR products were amplified with the primers and conditions as shown in Tables 2.7 and 2.10, respectively. The expected size of products is 365bp. This was verified by 1.5% agarose gel electrophoresis of PCR products. The PCR products were then digested with the restriction enzyme *Hin*fI. In a digestion, 8μl of PCR product was added to a 12μl digestion solution containing 9.5μl of purified water, 2μl of 10xNEBuffer 2 (NEB) and 5U *Hin*fI. After incubation at 37°C overnight, the sizes of the PCR products were re-checked again by 2.5% agarose gel electrophoresis, and showed one of these four possible types:

- Incomplete digestion = 365bp
- Homozygous (C/C; no variable site) = 263bp fragment
- Homozygous (A/A; has variable site) = 237bp fragment
- Heterozygous (C/A; one allele has variable site and the other one does not) = 237bp and 263bp fragments

The sequence of the PCR products (M = A or C) is:

TCCCTGCAGAAAATGAACAAg*gattc*atgtacactggcaggtactggcagccaccc

agggcctctcacaggaaagggagatcagaaagagaagcaaagag*gactc*atgagataccac

agggccgct

gMgtccagccttgcctggagctagggccacctcgatgccctatagtcttggagccacaaggtgc

atttactcaaagcctctttgagtttggtttgcttgtttgctttctgcctggaaactgccagcatcctgaga

gatacgagatctgcatctgtgcagagacacagggtttgttaaaagtcacaggccctgactgaagt

gtggaactggctgaaaTGAGAAAGTAGGAGGTAATTTGG

The PCR product generated includes two non-polymorphic *Hin*fI sites (see above, GAnTC in blue), so that if digestion is complete, all 365bp products should be cut into 23bp + 342bp fragments, and the 342bp fragment will further cut into 79bp + 263bp fragments. If the variable *Hin*fI site is present (*i.e.* if M = A in GMgTC in red above), all 263bp fragments should be cut into 26bp + 237bp fragments.

2.4.2 SNP assay for rs1130371 and rs1804185

This assay combines two SNP assays. The SNP rs1130371 was selected from the HapMap database and the SNP rs1804185 was selected from the UCSC genome browser. The SNPs are just 60bp apart so it is beneficial for both SNPs to be genotyped simultaneously using the same restriction enzyme. Fortunately, *Nci*I can be used to investigate both SNPs. However, the assay had to construct a *Nci*I constant site for use as a control for complete digestion by designing a mismatch into the forward primer.

The sequence of the PCR products (Mismatched primer base in green; Y = C or T, R = A or G and S = G or C) is:

CCCTTTCCTCTGGG*CCGGg*gcagcccttcctgactctgtaacacatgcctcactccag
ctccaagtcaggtcacacctcggagccctgcgtcctgtatccccgataggctcctgaaggctggg
cctttccaggatggccttctggcctgtctctgccccaaccctgaccctccctacctccatagaggtga
gcaggaagactggcacttacatgacaccYggcttggagcactggctgctcgtctcaaagtagtc
agctatgaaattctgtggaatctgccggRaggtgtagctgaagcagcaggcggtcggcgtgtca
gcagcaactgtggagaaaggaagagaataagcccgagtcacagctcagaagaaaaggcca
ggcagcttctgatccccgagcagttgaggaaggcaggcttgctcagaccaagtgactggaagg
CATTTGGGCATTTTTGCTG

The PCR product generated includes a non-polymorphic *Nci*I site (see above, ccsgg in blue), so that if digestion is complete, all 461bp products should be cut into 17bp + 444bp fragments. If the first variable *Nci*I site is present (see above, ccYgg in red), all 444bp fragments should be cut into 203bp + 241bp fragments. If the second variable *Nci*I site is present (see above ccggR in red), all products should further cut 241bp fragment into 62bp + 179bp fragments.

Table 2.7: SNP assays within the *CCL3* region.

SNP	Position [UCSC:Mar.2006 (NCBI36/HG18)]	Variant bases	Primer sequence (5′-3′)	Size of product (bp)	Restriction enzyme	Variant sizes after digestion(bp)
rs1634492	chr17:31,435,952	A/C	F: TCCCTGCAGAAAATGAACAA R: CCAAATTACCTCCTACTTTCTCA	365	*Hinf*I	A = 237 C = 263
rs1851503	chr17:31,436,315	G/T	F: ATGCTGCAGCCTTGGTATTT R: TGGCTGAAATGAGAAAGTAGGA	312	*Nsi*I	G = 241 T = 159
rs1130371	chr17:31,440,650	C/T	F: CCCTTTCCTCTGGGCCGG R: CAGCAAAAATGCCCAAATG	461	*Nci*I	C = 203 T = 265
rs1804185	chr17:31,440,710 chr17:31,547,378 chr17:31,635,421 chr17:31,648,952	A/G				A = 241 G = 179
rs9972960	chr17:31,444,192	A/G	F: GCTCTTCAAAAATCTAGACCATGC R: TCCCTATTATGGAAATCAGCGCA	205	*Hha*I	A = 187 G = 116+71
rs1634502	chr17:31,444,611	A/T	F: TCCAGACCATGTCTTGTTGG R: TGGCAAAGGGATAGTTCCTG	207	*Mse*I	A = 180 T = 152

Table 2.8: SNP assays at *CCL3L1* and *CCL4L1*.

SNP	Position [UCSC:Mar.2006 (NCBI36/HG18)]	Variant bases	Primer sequence (5'-3')	Size of product (bp)	Restriction enzyme	Variant sizes after digestion(bp)
rs2277661	chr17:31,547,099 chr17:31,635,142 chr17:31,648,673	A/G	F: CTCTCCTTTCTCCGCAGTTG R: TGCTGCCTCCTTCTTCCGGT	363	*Sau*96I	A = 143 G = 123
rs3744594	chr17:31,563,494 chr17:31,665,060	A/G	F: TAACTTCCTCCCCAGGAACC R: TATCAACCCCTGGCTGCGCT	215	*Hha*I	A = 200 G = 135
rs4796195	chr17:31,563,995 chr17:31,665,561	A/G	F: ACAGCTAAATCCAGTGAGTG R: GCCTCTTTTGGTTTGGGATC	212	*Bam*HI	A = 150 G = 132

Table 2.9: Additional SNP assays within chromosome 17q12.

SNP	Position [UCSC:Mar.2006 (NCBI36/HG18)]	Variant bases	Primer sequence (5′-3′)	Size of product (bp)	Restriction enzyme	Variant sizes after digestion(bp)
rs12453313	chr17:31,470,419	G/T	F: TGGTGCAGCTGAAAAGGAAT R: GTGGGTACCAGTGAGAATAGGAATTA	645	*Mse*I+*Ban*II	G = 322 T = 297
rs9910447	chr17:31,496,575	C/T	F: TGGCAGTTTCTTGCTGTGTC R: TTTTTACTGGAGGCCTGTCG	547	*Psp*GI	C = 399 T = 332
rs16972085	chr17:31,500,509	A/G	F: CCTTTAAGTCACTGGGCTTATATTGCAG R: TGTGAAATTTACAGAGATTTCCATCTCATC	240	*Bsa*BI	A = 188 G = 220
rs8070238	chr17:31,890,299	C/T	F: CTGGAAGTTTCCTGGCAGAG R: GGTAGGGGAAGGAGAATTGG	505	*Nde*I	C = 118+227 T = 345
rs8064426	chr17:31,893,863	A/G	F: GGCATCCAGAAATAAGGGAGA R: AGAAAGGGCCAATGCTTGT	300	*Nci*I	A = 278 G = 230
rs868340	chr17:31,906,733	A/T	F: CAACCGATGCCTTAAACACA R: GCCATGTTCAACTGTAATCTGC	279	*Alu*I	A = 126 T = 91

Table 2.10: PCR conditions for SNP assays.

SNP	Denaturing		Annealing		Extension		Cycles
	Temp. (°C)	Time (s)	Temp. (°C)	Time (s)	Temp. (°C)	Time (s)	
rs1634502	95	30	60	30	70	30	40
rs1634492	95	30	54	30	70	30	37
rs1130371 and rs1804185	95	30	57	30	70	30	37
rs4796195	95	30	60	30	70	30	37
rs8064426	95	30	60	30	70	30	40
rs1851503	95	30	60	30	70	30	37
rs9972960	95	30	57	30	70	30	40
rs16972085	95	30	60	30	70	30	37
rs3744594	95	30	65	30	72	30	37
rs2277661	95	30	60	30	72	30	37
rs868340	95	30	60	30	70	30	37
rs8070238	95	30	60	30	70	30	37
rs9910447	95	30	60	30	70	30	37
rs12453313	95	30	54	30	70	60	37

2.5 Allele-specific PCR assays

All allele-specific primers and conditions used at *CCL3L1* and *CCL4L1* are shown in Tables 2.11-2.12. The denaturing and extension temperature for all primers are 95°C and 70°C, respectively, and all 3 steps of the PCR cycle used 30 seconds in each cycle of PCR. Due to the length of allele-specific PCR products, sequencing using an additional primer was used to obtain full length sequence of the product. More details of the assays are explained in the section 5.4.

Table 2.11: Allele-specific PCR assays in *CCL3L1*.

Allele-specific name	rs number/Position [UCSC:Mar.2006 (NCBI36/HG18)]	Forward primer with allele-specific bases (shown in bold)	Annealing	cycles	Reverse primer
ASP87-532	Chr17: 31546263	TGTGAAGTGGATAACTCTGTCG**C**	68.1	37	ACCTGGAGCTGAGTGCCTGA
	Chr17: 31647837	TGTGAAGTGGATAACTCTGTCG**T**	67.0	40	
ASP532-1202	rs2944	CCAGTCCATAGAAGAGGTAGCTGT**A**	65.5	37	CCCAAATGCCTTGCAGTCAC
	Chr17: 31546708	CAGTCCATAGAAGAGGTAGCTGT**G**	65.5	37	
	Chr17: 31648282	*Additional primer: GCCTTCTGGCCTGTTTCTG	65.5	25	
ASP532-1863	rs2944	CCAGTCCATAGAAGAGGTAGCTGT**A**	65.0	37	ACACTCGAGCCCACATTCCA
	Chr17: 31546708	CAGTCCATAGAAGAGGTAGCTGT**G**	65.0	40	
	Chr17: 31648282	*Additional primer: TACCCCAGCCCAAGAGAAGC	67.0	20	
ASP87-632	Chr17: 31546263	TGTGAAGTGGATAACTCTGTCG**C**	69.2	37	CCTTCTCCACAGCTTCCTAACCA
	Chr17: 31647837	TGTGAAGTGGATAACTCTGTCG**T**	68.1	37	
ASP632-1863	rs17850251	GGACCCCTCAGGCACTC**A**	67.0	37	ACACTCGAGCCCACATTCCA
	Chr17: 31546808	GGACCCCTCAGGCACTC**G**	68.1	37	
	Chr17: 31648382	*Additional primer: TACCCCAGCCCAAGAGAAGC	67.0	20	

Note: *1:10 dilution of allele-specific PCR product was re-amplified using the additional primer.

Table 2.12: Allele-specific PCR assays in *CCL4L1.*

Allele-specific name	rs number/Position [UCSC:Mar.2006 (NCBI36/HG18)]	Forward primer with allele-specific bases (shown in bold)	Annealing	cycles	Reverse primer
ASP712-1467	rs4796195	GCCTCTTTTGGTTTGGAATC**C**	64.1	40	GGGAATGGATACAAGGGACCATA
	chr17: 31563995	GCCTCTTTTGGTTTGGAATC**T**	64.1	40	
	chr17: 31665561				
ASP1467-2086	rs3744597	TGCCCATGGGAGCTGTTTA**G**	65.5	37	GGAGCCATGACATCATCTCTACCA
	chr17: 31563240	TGCCCATGGGAGCTGTTTA**T**	65.5	40	
	chr17: 31664806				
ASP2086-2511	rs2277660	AAGAGGTTTTCTCAGAGGTGAG**G**	65.5	40	CTGGTTCCCGCTTTGTTCTTG
	chr17: 31562621	CAAAGAGGTTTTCTCAGAGGTGAG**T**	65.5	37	
	chr17: 31664187				
ASP850-1467	rs4796193	TGGATGGGATCCTTCCT**C**	64.1	40	GGGAATGGATACAAGGGACCATA
	chr17: 31563857	TGGATGGGATCCTTCCT**G**	64.1	40	
	chr17: 31665423				
ASP1213-2511	rs3744594	TGAGAGCTGCCTCCCCA**A**	68.1	37	CTGGTTCCCGCTTTGTTCTTG
	chr17: 31563494	TGAGAGCTGCCTCCCCA**G**	68.1	37	
	chr17: 31665060	*Additional primer CATGGCAATCTTCCTAGCTGC	65.0	15	

Note: *1:10 dilution of allele-specific PCR product was re-amplified using the additional primer.

2.6 Emulsion haplotype fusion PCR

Emulsion haplotype fusion PCR is a technique, published by Turner and Hurles (2009), to determine haplotypes over long distances. In this study, emulsion haplotype fusion PCR was developed and optimized by Dr. Jess Tyson (personal communication) to determine the phase for haplotypes of *CCL3L1/CCL4L1*. The principle of emulsion haplotype fusion PCR is shown in Figure 2.5.

To achieve *CCL3L1/CCL4L1* phased haplotypes by using emulsion haplotype fusion PCR, PCR amplification was done in two stages. The first PCR amplification was performed in emulsion droplets to produce fused products from two loci, *CCL3L1* and *CCL4L1*, and the second PCR amplification was performed in a similar manner to allele-specific PCR to construct a PCR product that once sequenced would reveal the variant bases on the same haplotype. In some samples, a third PCR amplification was needed using an additional primer to obtain the full length sequence of the product.

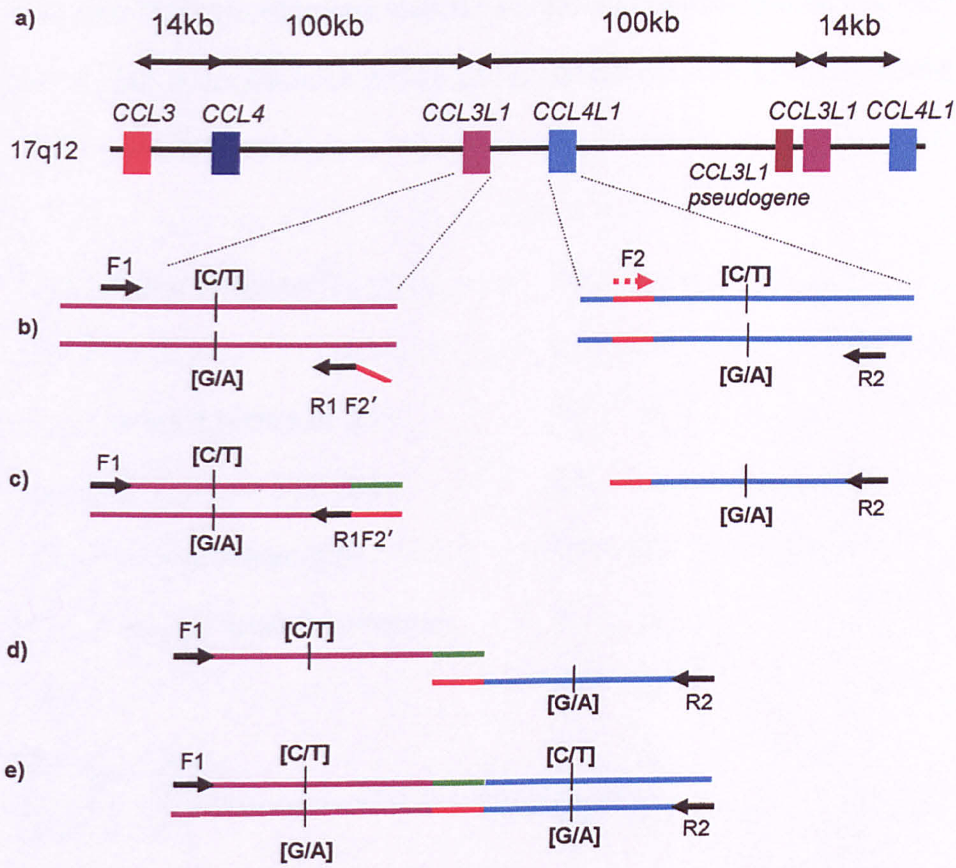


Figure 2.5: Emulsion haplotype fusion PCR. a) Diagram of *CCL3L1* and *CCL4L1* at Chr.17q12; they are about 14kb apart. b) The primer F1 is the forward primer which was designed to be specific for the *CCL3L1* locus, and R2 is the reverse primer which was designed to the locus on *CCL4L1*. F2, which in typical PCR is the forward primer for *CCL4L1* locus, was designed to be a reverse complementary sequence to append to the 5′ end of primer R1, which is then called the fusion primer R1F2′. The variant bases are indicated symbolically in both loci. c) PCR products from both loci were produced. d) The *CCL3L1* PCR product can prime on the PCR product from *CCL4L1* since the tail on the fusion primer matches the reverse complement of F2 from *CCL4L1* locus. e) The fused PCR product was amplified by the primers F1 and R2. Then, the variant bases from both loci of *CCL3L1* and *CCL4L1* can be defined as a *CCL3L1/CCL4L1* haplotype using allele-specic PCR and sequencing.

The first PCR amplification, which was done in emulsion droplets, was composed of an aqueous phase and an oil phase. The aqueous phase contained the following

5x GC buffer (Phusion®)	20	µl
10mM dNTPs	2	µl
10 µm forward primer (F1)	10	µl
1 µm fusion primer (R1-F2′)	2.5	µl
10 µm reverse primer (R2)	10	µl
2U/µl Phusion® DNA Polymerase	7	µl
H_2O	47.5	µl
200 ng/µl DNA	1	µl
Total	100	µl

To create emulsion droplets, 10 µl prepared aqueous phase was pipetted dropwise every 5 seconds to 200 µl oil phase (including 9 µl Span 80 (Sigma, cat.no. S6760), 0.8 µl Tween 80 (Sigma, cat.no. P8074) and 1 µl Triton X-100 (Sigma, cat.no. T9284) in 200 µl of light mineral oil (Sigma, cat.no. M5904)) while stirring at 1000 rpm with a magnetic stirrer. 100 µl mixture was aliquoted into 0.5 ml tubes and amplified at 98°C for 30 seconds followed by 40 cycles of 98°C for 10 seconds and 72°C for 45 seconds, followed by 72°C for 5 minutes, and a 4°C hold.

The fusion PCR products were extracted from emulsion droplets by mixing with an equal volume of hexane, vortexing and centrifuging at 13,000g for 3 minutes. The oil phase was removed and the hexane extraction was repeated for an additional 2 times or until the layer between oil phase and aqueous phase was extremely thin. Finally, the fusion PCR products in aqueous phase were diluted to 1:10 with water for the second round of PCR.

The second PCR was carried out using the mixture described below:

10x NH_4 buffer (Bioline®)	1.9	µl
50 mM $MgCl_2$	0.8	µl
25mM dNTPs	0.16	µl
10 µM forward primer (F1)	1	µl
10 µM reverse primer (R2)	1	µl
5U/µl Taq DNA Polymerase	0.2	µl
H_2O	13.94	µl
1:10 dilution of 1st round PCR product	1	µl
Total	20	µl

The primers and conditions used in both rounds of PCR are shown in Table 2.13 and Table 2.14.

Table 2.13: Primers used in emulsion haplotype fusion PCR (1st PCR amplification).

Emulsion-fusion PCR name	rs number/Position [UCSC:Mar.2006 (NCBI36/HG18)]	Primer name	Primer sequence (5′-3′)
EFP 1616-712	rs1828283 Chr17: 31547791, 31649366 to rs4796195 Chr17: 31563995, 31665561	F1 (sys1)	AGGGCTCCTGGGAGACCTAGG
		R1-F2′(sys1)	TCACTGGATTTAGCTGTGGGCCCCATGGTTAGACCACATCAGTCTTTTT
		R2 (sys4)	AGGACTCACTGGGGTCAGCG
EFP 532-712	rs2944 Chr17: 31546708, 31648282 to rs4796195 Chr17: 31563995, 31665561	F1 (sys5)	CACTCGGTTGTCACCAGACACAC
		R1-F2′ (sys5)	TCACTGGATTTAGCTGTGGGCCCTGCCCTCCTCAACCACTCA
		R2 (sys4)	AGGACTCACTGGGGTCAGCG
EFP 1616-2086	rs1828283 Chr17: 31547791, 31649366 to rs2277660 chr17: 31562621, 31664187	F1 (sys1)	AGGGCTCCTGGGAGACCTAGG
		R1-F2′(sys1-6)	TGTCATGGCTCCTGAAGCTAGCCCATGGTTAGACCACATCAGTCTTTTT
		R2 (sys6)	CCCAAAACAGGCCCCCTTTA

Table 2.14: Primers used in emulsion haplotype fusion PCR (2nd PCR amplification).

Emulsion-fusion PCR name	Primer name	Primer sequence (5′-3′) with allele-specific bases (shown in bold)	Denaturing Temp. (°C)	Denaturing Time (s)	Annealing Temp. (°C)	Annealing Time (s)	Extension Temp. (°C)	Extension Time (s)	Cycles
EFP 1616-712	F1 (1616C)	F: CTAGGGTGAGCTGGAGAGTGAA**C**	95	30	-	-	72	60	30
	F1 (1616G)	F: AGGGTGAGCTGGAGAGTGAA**G**							
	R2 (sys1)	R: GTCAGCGCAGACTTGCTTGCC							
EFP 532-712	F1 (532A)	F: CCAGTCCATAGAAGAGGTAGCTGT**A**	95	30	69.7	30	70	30	30
	F1 (532G)	F: CAGTCCATAGAAGAGGTAGCTGT**G**							
	R2 (sys1)	R: GTCAGCGCAGACTTGCTTGCC							
	ADD (532)	*Additional primer: F: GCCTTCTGGCCTGTTTCTG	95	30	65.5	30	70	30	20
EFP 1616-2086	F1 (2086G)	F: AAGAGGTTTTCTCAGAGGTGAG**G**	95	30	65.5	30	70	60	30
	F1 (2086T)	F: CAAAGAGGTTTTCTCAGAGGTGAG**T**							
	R2 (1616C)	R: CTAGGGTGAGCTGGAGAGTGAAC							

Note: *1:10 dilution of allele-specific PCR product was re-amplified using the additional primer.

2.7 Triplex PRT PCR assay

PRT is an abbreviation of Paralog Ratio Test, which is a technique to measure copy number. In brief, this technique uses a single pair of specifically designed primers to amplify two products from a copy number variable site of interest, called the test region, and a reference locus simultaneously. Then, the copy number of the test region is inferred from the ratio of test to reference products (Armour *et al.* 2007).

The triplex PRT PCR assay was developed based on the PRT technique designed by Susan Walker (Walker 2009; Walker *et al.* 2009). The assay combined 3 systems, including CCL3C, CCL4A and LTR61A as shown in Figure 2.6.

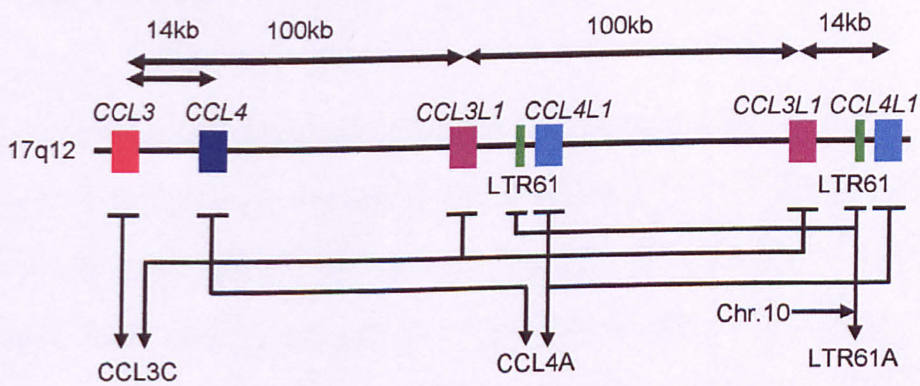


Figure 2.6: Schematic diagram of triplex PRT system.

The aims of CCL3C and CCL4A systems are to measure the copy number of *CCL3L1* and *CCL4L1* relative to their paralogues *CCL3* and *CCL4*. For LTR61A, the system is specially designed to measure a Long Terminal Repeat located between *CCL3L1* and *CCL4L1* relative to a homologous sequence on chromosome 10 at an unlinked reference locus known to be a single copy region. The advantage of including this system in triplex PRT PCR assay was to confirm the copy number of *CCL3L1* and *CCL4L1* determined by those two systems. Moreover, it would help to detect any copy number variation in *CCL3* and *CCL4* regions.

Here, the triplex PRT PCR assay was applied, mainly to determine *CCL3L1* and *CCL4L1* copy numbers in paediatric tuberculosis case-control analysis samples, with some modifications to the protocol as described below.

A PCR was performed in 1x LD buffer using 5 ng genomic DNA and 0.5 U *Taq* DNA polymerase in a total volume of 10 µl. Products were amplified with 0.5 µM each of FAM-CCL3CF, CCL3CR, FAM-CCL4AF and CCL4AR and 1.5 µM of FAM-LTR61AF and LTR61AR for 23 cycles of 95°C for 30 seconds, 55°C for 30 seconds, 70°C for 1 minute followed by a final hold at 70°C for 40 minutes. Sequences of all primers are shown in Table 2.15.

Table 2.15: Sequence of triplex PRT PCR primers.

Primer name	Primer sequence (5′-3′)
CCL3CF	GGCTAAGACCCCTTCTAGAG
CCL3CR	AATCATGCAGGTCTCCACT
CCL4AF	GAGTCTGCTTCCAGTGCT
CCL4AR	GAGGAGTCCTGAGTATGGAG
LTR61AF	AGTTTTCCTCTGCCTAGC
LTR61AR	TATTTATTTTAAGGTGTGCAC

For each sample, the PCR was carried out in triplicate as described above, substituting FAM-labelled forward primers with HEX and NED labelled primers. Fluorescently labelled PCR products were analyzed by capillary electrophoresis on an ABI3100 36 cm capillary using POP-4 polymer with an injection condition of 2kV for 45 seconds. 1 µl of each fluorescently labelled PCR product was mixed with 10µl Hi-Di^{TM} Formamide (Applied Biosystems) and ROX-500 marker (Applied Biosystems) and denatured at 95°C for three minutes prior to electrophoresis.

Finally, Genescan software (Applied Biosystems) was used for data analysis. The heights of test and reference peaks for each of the three systems were collected and used to calculate the ratio of test/reference signals. The average of the ratio of the three fluorescent dyes of each system was used to infer the copy number for each system in each sample using reference standards as calibration.

Chapter 3: *CCL3L1* pseudogene assay

3.1 Introduction

Pseudogenes are defective copies of regular genes. Generally, they are not functional because they are unable to be transcribed or translated. However, some pseudogenes can be transcribed and have a function (Hirotsune *et al.* 2003; Zheng *et al.* 2005). There are about 20,000 pseudogenes in the human genome (Harrison *et al.* 2002; Zhang *et al.* 2003).

The *CCL3L1* pseudogene is a 5′ truncated copy of *CCL3L1* located at chromosome 17q12 (Hirashima *et al.* 1992; Modi 2004). It is also known as G0S19-3; SCYA3L2; LD78gamma and *CCL3L2* (http://www.ncbi.nlm.nih.gov/gene/390788). This pseudogene has sequences similar to *CCL3L1* gene, but is lacking exon 1 of *CCL3L1* (Figure 3.1) (Hirashima *et al.* 1992; Modi 2004). As a result, it might cause problems with measurement of *CCL3L1* copy number by *CCL3L1* pseudogene acting as a template for the PCR for the *CCL3L1* gene (Field *et al.* 2009). Consequently, it might affect the interpretation in an association study between *CCL3L1* gene and diseases (Gonzalez *et al.* 2005; Field *et al.* 2009; Shostakovich-Koretskaya *et al.* 2009).

The aims of this study were to study the presence of the pseudogene in populations, to investigate the association between *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1*, and to observe the patterns of this association in different populations. To accomplish the objectives, reliable assays for detection of the *CCL3L1* pseudogene, which can distinguish between this pseudogene and *CCL3L1* gene, were constructed and applied in HapMap samples, including CEPH samples, Chinese-Japanese samples and Yoruba samples, and ECACC (UK) samples. The PCR assay for *CCL3L1* pseudogene is shown in Figure 3.1.

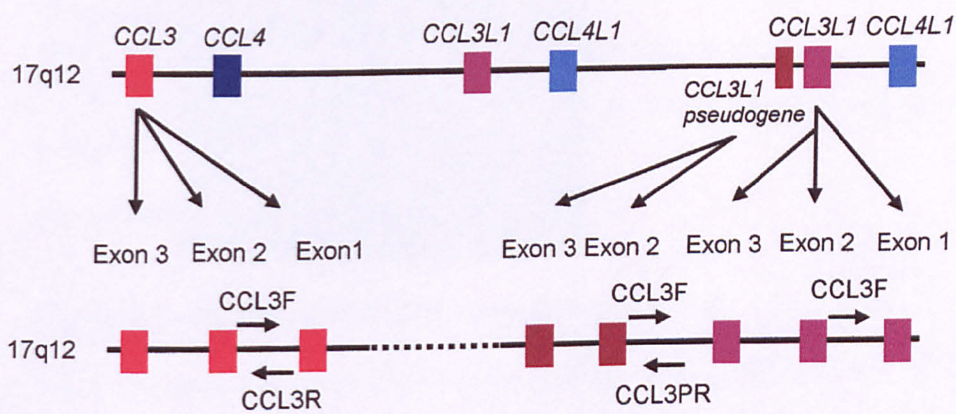


Figure 3.1: The structure of *CCL3L1* pseudogene and *CCL3L1* and *CCL3L1* pseudogene assay.

PCR products were amplified with the control primers, forward primer CCL3F and the reverse primer CCL3R, making a 358bp product specific to *CCL3* gene, and the alternative reverse primer CCL3PR, making a 233bp product specific to the DNA sequence around the novel junction created by the pseudogene duplication. The expected size of products from *CCL3* is 358bp, and products at 233bp indicated the presence of *CCL3L1* pseudogene. This was checked by 1.5% agarose gel electrophoresis as examples are shown in Figure 3.2. Then, the association between *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* was performed by using a Chi-square (χ^2) test.

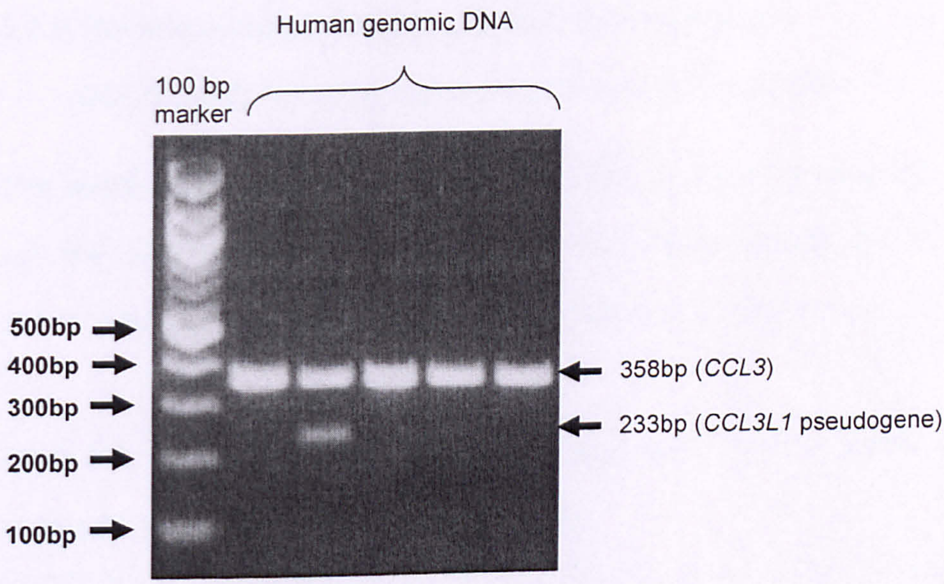


Figure 3.2: Examples of *CCL3L1* pseudogene assay.

The PCR products from *CCL3* gene appear at 358bp, and the PCR product at 233bp indicates the presence of *CCL3L1* pseudogene.

3.2 ECACC (UK) samples

3.2.1 *CCL3L1* pseudogene in ECACC (UK) samples

Using the above assay, The *CCL3L1* pseudogene was investigated in 192 ECACC (UK) samples, and it was found in 52 out of 192 samples (about 27%).

3.2.2 The relationship between *CCL3L1* pseudogene and *CCL3L1/CCL4L1* copy number in ECACC (UK) samples

The relationship between the presence of the *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* was further investigated in 192 ECACC (UK) samples. The results are shown in Tables 3.1-3.3.

Table 3.1: The presence of *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* in ECACC (UK) samples.

Copy number of *CCL3L1/CCL4L1*	0	1	2	3	4
CCL3L1 pseudogene present	0	4	14	29	5
CCL3L1 pseudogene absent	2	29	102	7	0
Total (samples)	2	33	116	36	5

Table 3.2: The presence of *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* in ECACC (UK) samples (grouped for χ^2 test).

Copy number of *CCL3L1/CCL4L1*	1-2	3	4
CCL3L1 pseudogene present	18	29	5
CCL3L1 pseudogene absent	131	7	0
Total (samples)	149	36	5

From the table, the χ^2 test shows that there is significant relationship between *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* (p = 0.0008). However, this tested relationship is based on all samples having an equal chance of possessing a *CCL3L1* pseudogene independent of copy number. Thus, taking into account the copy number of *CCL3L1/CCL4L1* is necessary to confirm the significant relationship between them.

Table 3.3: The *CCL3L1* pseudogene from observation, the *CCL3L1* pseudogene from prediction and the copy number of *CCL3L1/CCL4L1* in ECACC (UK) samples.

Copy number of *CCL3L1/CCL4L1*	1	2	3	4
Observed *CCL3L1* pseudogene	4	14	29	5
Predicted *CCL3L1* pseudogene	4	26.7	11.5	2
Total (samples)	33	116	36	5

NB. The numbers of predicted *CCL3L1* pseudogene are calculated from the probability of samples that will find at least one copy of *CCL3L1* pseudogene.

After taking the copy number of *CCL3L1/CCL4L1* into account, the χ^2 test still shows that there is significant relationship between *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* (p = 0.0006).

Lastly, the mean copy number of *CCL3L1/CCL4L1* between *CCL3L1* pseudogene present group and *CCL3L1* pseudogene absent group was calculated and tested for significant difference. The results are shown in Table 3.4.

Table 3.4: Mean copy numbers of *CCL3L1/CCL4L1* in *CCL3L1* pseudogene present group and *CCL3L1* pseudogene absent group in ECACC (UK) samples.

The presence of *CCL3L1* pseudogene	Pseudogene present (N=52)	Pseudogene absent (N=140)
Mean copy number of *CCL3L1/CCL4L1*	2.67	1.81

The t-test shows that the mean copy numbers of *CCL3L1/CCL4L1* between these two groups are significantly different ($p = 4.17 \times 10^{-16}$).

3.3 CEPH HapMap samples

3.3.1 *CCL3L1* pseudogene in CEPH HapMap samples

The *CCL3L1* pseudogene was also investigated in 90 CEPH HapMap samples. The *CCL3L1* pseudogene was found in 17 out of 90 samples, and in 13 out of 60 samples when children were excluded.

3.3.2 The relationship between *CCL3L1* pseudogene and *CCL3L1/CCL4L1* copy number in CEPH HapMap samples

The relationship between the presence of the *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* in CEPH HapMap samples was examined as for the ECACC (UK) samples. Moreover, CEPH HapMap samples allowed the investigation of the relationship between haplotype copy number of *CCL3L1/CCL4L1* and the presence of the *CCL3L1* pseudogene, since the haplotype copy number of *CCL3L1/CCL4L1* in these samples can be examined using segregation analysis. All results are shown in Tables 3.5-3.8.

Table 3.5: The presence of *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* in CEPH HapMap samples (excluding children).

Copy number of *CCL3L1/CCL4L1*	0	1	2	3
CCL3L1 pseudogene present	0	1	6	6
CCL3L1 pseudogene absent	2	13	29	3
Total (samples)	2	14	35	9

A χ^2 test shows that there is a significant relationship between *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* (p = 0.0144).

Similar to the analysis in ECACC (UK) samples, this tested relationship is based on all samples having an equal chance independent of copy number. Thus, taking into account the copy number of *CCL3L1/CCL4L1* is necessary for confirming the significant relationship between them.

Table 3.6: The *CCL3L1* pseudogene from observation, the *CCL3L1* pseudogene from prediction and the copy number of *CCL3L1/CCL4L1* in CEPH HapMap samples (excluding children).

Copy number of *CCL3L1/CCL4L1*	1	2	3
Observed *CCL3L1* pseudogene	1	6	6
Predicted *CCL3L1* pseudogene	1	4.9	1.8
Total (samples)	14	35	9

NB. The numbers of predicted *CCL3L1* pseudogene are calculated from the probability of samples that will find at least one copy of *CCL3L1* pseudogene if all repeat units have the same probability.

From the table, a t-test shows that there is no significant difference in *CCL3L1/CCL4L1* copy number between observation and prediction ($p = 0.218$).

Nevertheless, the association between the presence of *CCL3L1* pseudogene and the haplotype copy number of *CCL3L1/CCL4L1* was further analyzed, and data are shown in Table 3.7 and Table 3.8, respectively.

Table 3.7: The presence of *CCL3L1* pseudogene and the haplotype copy number of *CCL3L1/CCL4L1* in CEPH HapMap samples (excluding children).

Presence of *CCL3L1* pseudogene	Haplotype copy number of *CCL3L1/CCL4L1*		
	0	1	2
CCL3L1 pseudogene present	0	3	9
Unknown	0	8	0
CCL3L1 pseudogene absent	24	49	7
Total (haplotypes)	24	60	16

Table 3.8: The presence of *CCL3L1* pseudogene and the haplotype copy number of *CCL3L1/CCL4L1* in CEPH HapMap samples (excluding children, and grouped for Fisher's exact test).

Presence of *CCL3L1* pseudogene	Haplotype copy number of *CCL3L1/CCL4L1* (total)	
	1 (60)	2 (16)
CCL3L1 pseudogene present	3	9
CCL3L1 pseudogene absent	49	7

A Fisher's exact test shows that there is a significant relationship between *CCL3L1* pseudogene and the haplotype copy number of *CCL3L1/CCL4L1* ($p = 3.6 \times 10^{-5}$).

A t-test shows that the mean copy numbers of *CCL3L1/CCL4L1* between *CCL3L1* pseudogene present group and *CCL3L1* pseudogene absent group are significantly different ($p = 0.0008$). The mean copy numbers of *CCL3L1/CCL4L1* in these two groups are 2.38 and 1.70, respectively.

3.4 YRI HapMap samples

3.4.1 *CCL3L1* pseudogene in YRI HapMap samples

CCL3L1 pseudogene assay was carried out in YRI (African) samples. After investigation, the *CCL3L1* pseudogene was found in 58 out of 90 samples, and in 37 out of 60 samples when children were excluded.

3.4.2 The relationship between *CCL3L1* pseudogene and *CCL3L1/CCL4L1* copy number in YRI HapMap samples

The investigation of the relationship between the presence of the *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* in YRI samples was analysed as for CEPH HapMap samples. The data are shown in Table 3.9.

Table 3.9: The presence of *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* in YRI samples (excluding children).

Copy number of *CCL3L1/CCL4L1*	2	3	4	5	6	7	8	9	10
CCL3L1 pseudogene present	3	2	12	8	5	5	1	0	1
CCL3L1 pseudogene absent	4	8	3	1	4	1	0	2	0
Total (samples)	7	10	15	9	9	6	1	2	1

A χ^2 test grouping *CCL3L1/CCL4L1* copy numbers of 2-4, 5 and 6-10 shows that there is no significant relationship between *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* (p = 0.6623).

A Fisher's exact test showed that there is no significant relationship between *CCL3L1* pseudogene and the haplotype copy number of *CCL3L1/CCL4L1* (p = 0.5884).

A t-test showed that the mean copy numbers of *CCL3L1/CCL4L1* between *CCL3L1* pseudogene present group (4.95) and *CCL3L1* pseudogene absent group (4.26) are not significantly different (p = 0.087).

3.5 CHB/JPT HapMap samples

3.5.1 *CCL3L1* pseudogene in CHB/JPT HapMap samples

After *CCL3L1* pseudogene assay was carried out in CHB/JPT samples, the *CCL3L1* pseudogene was found in 48 out of 90 samples.

3.5.2 The relationship between *CCL3L1* pseudogene and *CCL3L1/CCL4L1* copy number in CHB/JPT HapMap samples

Like ECACC (UK) samples, the relationship between *CCL3L1* pseudogene and *CCL3L1/CCL4L1* copy number was investigated in CHB/JPT samples. The data are shown in Table 3.10.

Table 3.10: The presence of *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* in CHB/JPT samples.

Copy number of *CCL3L1/CCL4L1*	1	2	3	4	5	6	7	8
CCL3L1 pseudogene present	0	2	9	12	11	6	5	3
CCL3L1 pseudogene absent	3	11	8	11	4	5	0	0
Total samples	3	13	17	23	15	11	5	3

A χ^2 test grouping *CCL3L1/CCL4L1* copy numbers of 1-3, 4-5 and 6-8 shows that there is no significant relationship between *CCL3L1* pseudogene and the copy number of *CCL3L1/CCL4L1* (p = 0.2304).

The mean copy numbers of *CCL3L1/CCL4L1* in *CCL3L1* pseudogene present group and *CCL3L1* pseudogene absent group are 4.77 and 3.40, respectively. A t-test shows that the mean copy numbers of *CCL3L1/CCL4L1* between these two groups are significantly different ($p = 2.75 \times 10^{-5}$).

In summary, this *CCL3L1* pseudogene assay showed that the *CCL3L1* pseudogene is found to vary between individuals and in populations, as does the copy number of *CCL3L1/CCL4L1*. However, the *CCL3L1* pseudogene only has a significant association with *CCL3L1/CCL4L1* copy number and haplotype of *CCL3L1/CCL4L1* copy number in ECACC and CEPH HapMap samples.

To answer the next question in the objectives of this study, the SNP and microsatellite assays were carried out to investigate the association between single-copy markers, including SNPs flanking the region of *CCL3L1/CCL4L1* and *CCL3* microsatellite, and *CCL3L1/CCL4L1* copy number. In addition, the SNPs within the repeat unit of *CCL3L1/CCL4L1* also were tested for the association with *CCL3L1/CCL4L1* copy number. More details and results are provided in the next chapter.

Chapter 4: SNP and Microsatellite assays

4.1 Introduction

To search for a genetic marker that could be used to determine copy number of *CCL3L1/CCL4L1* and/or to tag the *CCL3L1* pseudogene, an easy method for investigation was a direct test of association between the *CCL3L1* pseudogene, *CCL3L1/CCL4L1* copy number and a selected single copy genetic marker.

In this study, the single copy genetic markers used to investigate were SNPs and microsatellites since they are widely used to discover genes for common and complex diseases, and to investigate association between markers and causal alleles.

4.2 SNP assays

4.2.1 Introduction

SNPs are a kind of human variation in which single base pair variation is found in > 1% of chromosomes in a given population (Feuk *et al.* 2006). They are abundant and can be found throughout the human genome both in coding and non-coding DNA. There are, currently, several techniques available for genotyping SNPs, from classical methods like restriction fragment length polymorphism (RFLP), to the high-throughput method such as SNP chips, in which thousands of SNPs can be analysed simultaneously in a single microarray. SNPs have been now well accepted and exploited (alone or combined with other markers) as a genetic tool in genome wide association studies (GWAS) to find variants underlying diseases. Examples include the discovery of the association between polymorphism of the complement factor H gene (*CHF*) and age-related macular degeneration (AMD) (Klein *et al.* 2005) and a SNP in a non-coding region creating a transcription factor binding site and altering the hepatic expression of *SORT1* gene effected plasma low-density lipoprotein cholesterol (LDL-C) and very low-density lipoprotein (VLDL) levels (Musunuru *et al.* 2010).

Consequently, SNPs were the markers selected to perform the investigation of association among SNPs, *CCL3L1/CCL4L1* copy number and the *CCL3L1* pseudogene. The investigated SNPs were selected from the *CCL3* region, inside the *CCL3L1/CCL4L1* variable region and SNPs suggested by linkage disequilibrium analysis. More details are described in the materials and methods, in section 2.4.

4.2.2 Association between SNPs and *CCL3L1/CCL4L1* copy number

All investigated SNPs were genotyped by PCR-RFLP techniques, and then their alleles were tested for the association with *CCL3L1/CCL4L1* copy number using a χ^2 test (*CCL3L1/CCL4L1* copy number in ECACC samples was determined by Dr. Susan Walker, and *CCL3L1/CCL4L1* copy number in Basque, CEPH HapMap, CHB/JPT and YRI samples was determined by Dr. Danielle Carpenter). The results are summarized in Tables 4.1 to Table 4.5.

Briefly, there were five SNPs that showed a strong association with *CCL3L1/CCL4L1* copy number in ECACC samples ($p = 0.0001$). These included rs3744594, rs2277661, which were located inside the *CCL3L1/CCL4L1* repeat unit, and rs8064426, rs16972085 and rs8070238 which were located in the flanking region and suggested by LD analysis. The association between SNPs located inside the *CCL3L1/CCL4L1* repeat unit and *CCL3L1/CCL4L1* copy number were also shown to vary significant in the HapMap populations. Lastly, there was no significant relationship between a SNP at *CCL4L1* region and *CCL3L1/CCL4L1* copy number.

Table 4.1: Relationship between SNPs at *CCL3* and *CCL3L1/CCL4L1* copy number in 192 ECACC samples.

SNP	Genotype (total)	*CCL3L1/CCL4L1* copy number (total)					p-value (χ^2)
		0 (2)	1 (33)	2 (116)	3 (36)	4 (5)	
rs1634502	AA (67)	1	18	35	12	1	0.1043
	AT (87)	1	10	58	14	4	
	TT (38)	0	5	23	10	0	
rs1634492	CC (59)	1	13	33	11	1	0.7378
	CA (88)	1	12	55	16	4	
	AA (45)	0	8	28	9	0	
rs1130371 and rs1804185*	CC (120)	2	25	73	18	2	0.0387
	CT (64)	0	7	37	17	3	
	TT (8)	0	1	6	1	0	
rs1851503	GG (25)	1	8	12	4	0	0.1738
	GT (77)	1	12	48	13	3	
	TT (90)	0	13	56	19	2	
rs9972960	AA (34)	1	12	16	5	0	0.0135
	AG (80)	1	13	49	14	3	
	GG (78)	0	8	51	17	2	

NB. *No polymorphism at rs1804185 in ECACC samples.

Table 4.2: Relationship between SNP rs3744594 at *CCL3L1* and *CCL3L1/CCL4L1* copy number in ECACC and HapMap samples.

Sample	Genotype	*CCL3L1/CCL4L1* copy number*											Total	p-value (χ^2)
		0**	1	2	3	4	5	6	7	8	9	10		
ECACC	AA	N/A	7	10	3	2	-	-	-	-	-	-	22	0.0001
	AG	N/A	0	36	28	3	-	-	-	-	-	-	67	
	GG	N/A	24	69	5	0	-	-	-	-	-	-	98	
	Total	N/A	31	115	36	5	-	-	-	-	-	-	187	
CEPH-HapMap	AA	N/A	2	2	2	-	-	-	-	-	-	-	6	0.3502
	AG	N/A	0	9	6	-	-	-	-	-	-	-	15	
	GG	N/A	12	24	1	-	-	-	-	-	-	-	37	
	Total	N/A	14	35	9	-	-	-	-	-	-	-	58	
CHB/JPT	AA	-	2	5	6	4	3	0	1	0	-	-	21	0.0281
	AG	-	1	7	11	19	12	11	3	3	-	-	67	
	GG	-	0	1	0	0	0	0	1	0	-	-	2	
	Total	-	3	13	17	23	15	11	5	3	-	-	90	
YRI	AA	-	-	3	1	0	0	2	0	0	0	0	6	0.0099
	AG	-	-	3	7	14	7	7	6	1	2	1	48	
	GG	-	-	1	2	1	2	0	0	0	0	0	6	
	Total	-	-	7	10	15	9	9	6	1	2	1	60	

NB. *The range of *CCL3L1/CCL4L1* copy number in ECACC, CEPH HapMap, CHB/JPT and YRI are 0-4, 0-3, 1-8 and 2-10, respectively. **N/A = since they are zero copy, the samples cannot be typed.

Table 4.3: Relationship between SNP rs2277661 at *CCL3L1* and *CCL3L1/CCL4L1* copy number in ECACC and HapMap samples.

Sample	Genotype	*CCL3L1/CCL4L1* copy number*											Total	*p*-value (χ^2)
		0**	1	2	3	4	5	6	7	8	9	10		
ECACC	AA	N/A	15	46	13	3	-	-	-	-	-	-	77	0.0001
	AG	N/A	7	61	20	2	-	-	-	-	-	-	90	
	GG	N/A	10	5	2	0	-	-	-	-	-	-	17	
	Total	N/A	32	112	35	5	-	-	-	-	-	-	184	
CEPH-HapMap	AA	N/A	13	26	4	-	-	-	-	-	-	-	43	0.0377***
	AG	N/A	0	9	5	-	-	-	-	-	-	-	14	
	GG	N/A	1	0	0	-	-	-	-	-	-	-	1	
	Total	N/A	14	35	9	-	-	-	-	-	-	-	58	
CHB/JPT	AA	-	3	6	1	5	2	5	1	1	-	-	24	0.0042
	AG	-	0	7	14	18	12	5	4	2	-	-	62	
	GG	-	0	0	1	0	0	1	0	0	-	-	2	
	Total	-	3	13	16	23	14	11	5	3	-	-	88	
YRI	AA	-	-	3	4	7	3	3	1	0	1	0	22	0.5260
	AG	-	-	4	3	7	6	6	3	1	1	1	32	
	GG	-	-	0	3	1	0	0	2	0	0	0	6	
	Total	-	-	7	10	15	9	9	6	1	2	1	60	

*The range of *CCL3L1/CCL4L1* copy number in ECACC, CEPH HapMap, CHB/JPT and YRI are 0-4, 0-3, 1-8 and 2-10, respectively. **N/A = since they are zero copy, the samples cannot be typed. ***Using Fisher's exact test.

Table 4.4: Relationship between SNPs suggested from LD analysis and *CCL3L1/CCL4L1* copy number in ECACC and Basque samples.

SNP	Sample	Genotype	*CCL3L1/CCL4L1* copy number					Total	*p*-value
			0	1	2	3	4		(χ^2)
rs8064426	ECACC	AA	1	0	3	0	0	4	0.0001
		AG	1	21	10	2	2	36	
		GG	0	12	103	34	3	152	
		Total	2	33	116	36	5	192	
rs8064426	Basque	AA	2	0	0	1	0	3	0.0001
		AG	1	23	2	3	0	29	
		GG	1	7	86	26	2	122	
		Total	4	30	88	30	2	154	
rs16972085	ECACC	AA	2	33	103	20	2	160	0.0001
		AG	0	0	10	16	1	27	
		GG	0	0	0	0	2	2	
		Total	2	33	113	36	5	189	
rs16972085	Basque	AA	4	29	83	25	1	142	0.0093
		AG	0	1	5	5	1	12	
		GG	0	0	0	0	0	0	
		Total	4	30	88	30	2	154	
rs868340	ECACC	AA	0	0	36	7	0	43	0.0002
		AT	1	19	54	17	1	92	
		TT	1	14	23	11	4	53	
		Total	2	33	113	35	5	188	
rs8070238	ECACC	CC	1	20	109	35	5	170	0.0001
		CT	1	13	4	1	0	19	
		TT	0	0	0	0	0	0	
		Total	2	33	113	36	5	189	
rs9910447	ECACC	CC	1	20	78	30	4	133	0.0515
		CT	1	10	33	5	1	50	
		TT	0	3	0	0	0	3	
		Total	2	33	111	35	5	186	
rs12453313	ECACC	GG	2	32	111	34	5	184	0.6489*
		GT	0	1	3	1	0	5	
		TT	0	0	0	0	0	0	
		Total	2	33	114	35	5	189	

NB. *using Fisher's exact test

Table 4.5: Relationship between SNP rs4796195 at *CCL4L1* and *CCL3L1/CCL4L1* copy number in 190 ECACC samples.

Genotype (Total)	***CCL3L1/CCL4L1* copy number (Total)**				***p*-value (χ^2)**
	1 (33)	**2 (116)**	**3 (36)**	**4 (5)**	
AA (128)	26	79	19	4	0.1133
AG (51)	0	33	17	1	
GG (11)	7	4	0	0	

4.2.3 Association between SNPs and *CCL3L1* pseudogene

Genotyped SNPs in the investigation of association between SNPs and *CCL3L1/CCL4L1* copy number were also used to examine a relationship with the *CCL3L1* pseudogene. The results are shown in Tables 4.6 to Table 4.10.

In brief, there were two SNPs, rs3744594 (allele A) and rs16972085 (allele G), which showed a strong relationship with *CCL3L1* pseudogene (p = 0.0001) in ECACC samples.

Table 4.6: Relationship between SNPs at *CCL3* and *CCL3L1* pseudogene in 192 ECACC samples.

SNP	Genotype (total)	The presence of *CCL3L1* pseudogene		p-value (χ^2)
		Present	Absent	
rs1634502	AA (67)	16	51	0.3131
	AT (87)	22	65	
	TT (38)	14	24	
rs1634492	CC (59)	13	46	0.5461
	CA (88)	25	63	
	AA (45)	14	31	
rs1130371 and rs1804185*	CC (120)	27	93	0.0650
	CT (64)	22	42	
	TT (8)	3	5	
rs1851503	GG (25)	5	20	0.0459
	GT (77)	15	62	
	TT (90)	32	58	
rs9972960	AA (34)	7	27	0.0752
	AG (80)	17	63	
	GG (78)	28	50	

NB. *No polymorphism at rs1804185 in ECACC samples.

Table 4.7: Relationship between SNP rs3744594 at *CCL3L1* and *CCL3L1* pseudogene in ECACC and HapMap samples.

Sample	Genotype	The presence of *CCL3L1* pseudogene		Total	*p*-value (χ^2)
		Present	Absent		
ECACC	AA	9	13	22	0.0001
	AG	35	32	67	
	GG	7	91	98	
	Total	51	136	187	
CEPH-HapMap	AA	4	2	6	0.0773*
	AG	6	9	15	
	GG	3	34	37	
	Total	13	45	58	
CHB/JPT	AA	13	8	21	0.4011
	AG	34	33	67	
	GG	1	1	2	
	Total	48	42	90	
YRI	AA	3	3	6	0.3527
	AG	31	17	48	
	GG	3	3	6	
	Total	37	23	60	

NB. *Using Fisher's exact test.

Table 4.8: Relationship between SNP rs2277661 at *CCL3L1* and *CCL3L1* pseudogene in ECACC and HapMap samples.

Sample	Genotype	The presence of *CCL3L1* pseudogene		Total	*p*-value (χ^2)
		Present	Absent		
ECACC	AA	21	56	77	0.2612
	AG	28	62	90	
	GG	2	15	17	
	Total	51	133	184	
CEPH-HapMap	AA	6	37	43	0.0088
	AG	7	7	14	
	GG	0	1	1	
	Total	13	45	58	
CHB/JPT	AA	8	16	24	0.0497
	AG	38	24	62	
	GG	2	0	2	
	Total	48	40	88	
YRI	AA	14	8	22	0.0535
	AG	22	10	32	
	GG	1	5	6	
	Total	37	23	60	

Table 4.9: Relationship between suggested SNPs from LD analysis and *CCL3L1* pseudogene in ECACC samples.

SNP	Genotype	*CCL3L1* pseudogene		Total	*p*-value (χ^2)
		Present	Absent		
rs8064426	AA	1	3	4	0.6408
	AG	11	25	36	
	GG	40	112	152	
	Total	52	140	192	
rs16972085	AA	31	129	160	0.0001
	AG	19	8	27	
	GG	2	0	2	
	Total	52	137	189	
rs868340	AA	7	36	43	0.1002
	AT	25	67	92	
	TT	19	34	53	
	Total	51	137	188	
rs8070238	CC	46	124	170	0.6756
	CT	6	13	19	
	TT	0	0	0	
	Total	52	137	189	
rs9910447	CC	42	91	133	0.0439
	CT	9	41	50	
	TT	0	3	3	
	Total	51	135	186	
rs12453313	GG	49	135	184	0.6129*
	GT	2	3	5	
	TT	0	0	0	
	Total	51	138	189	

NB. *Using Fisher's exact test

Table 4.10: Relationship between SNP rs4796195 at *CCL4L1* and *CCL3L1* pseudogene in 190 ECACC samples.

Genotype (Total)	The presence of *CCL3L1* pseudogene		*p*-value (χ^2)
	Present	Absent	
AA (128)	25	103	0.0004
AG (51)	26	25	
GG (11)	1	10	

4.2.4 Tag-SNPs for *CCL3L1/CCL4L1* copy number

To discover a SNP that could be used to determine copy number of *CCL3L1/CCL4L1*, linkage disequilibrium analysis was performed in CEPH HapMap samples using Haploview (Barrett *et al.* 2005). Two SNPs, rs16972085 and rs8064426, were found as shown in Table 4.11.

Table 4.11: The relationship between genotype and *CCL3L1/CCL4L1* copy number and their mean copy number in SNP rs16972085 and rs8064426.

SNP	Genotype	*CCL3L1/CCL4L1* copy number				Total	Mean copy number
		0	1	2	3		
rs16972085	AA	4	24	46	8	82	1.707
	AG	0	0	2	5	7	2.714
	GG	0	0	0	0	0	0.000
	Total	4	24	48	13	89	
rs8064426	AA	1	0	0	0	1	0.000
	AG	0	19	6	2	27	1.370
	GG	3	4	42	12	61	2.032
	Total	4	23	48	14	89	

From the table, based on the mean copy number and the relationship between genotype and *CCL3L1/CCL4L1* copy number, allele G of SNP rs16972085 could be used to tag 2 copy haplotypes, and allele A of SNP rs8064426 could be used to tag 0 copy haplotypes.

To evaluate the potential of these tag-SNPs as markers for the determination of *CCL3L1/CCL4L1* copy number, both SNPs were tested for association with *CCL3L1/CCL4L1* copy number in ECACC and Basque samples.

Apparently, the investigation of their relationship showed that both SNPs had strong association with the *CCL3L1/CCL4L1* copy number ($p<0.001$) in both ECACC and Basque samples. So, the combination of these two promising SNPs was used for prediction of *CCL3L1/CCL4L1* copy number in ECACC and Basque samples. The accuracy of prediction was investigated by comparison to the known *CCL3L1/CCL4L1* copy number using PRT (the data were kindly provided by Dr. Danielle Carpenter); the study in Basque samples was undertaken blind to *CCL3L1/CCL4L1* copy number.

The prediction of *CCL3L1/CCL4L1* copy number in ECACC and Basque samples was carried out using the hypothesis as shown in Figure 4.1 and Table 4.12.

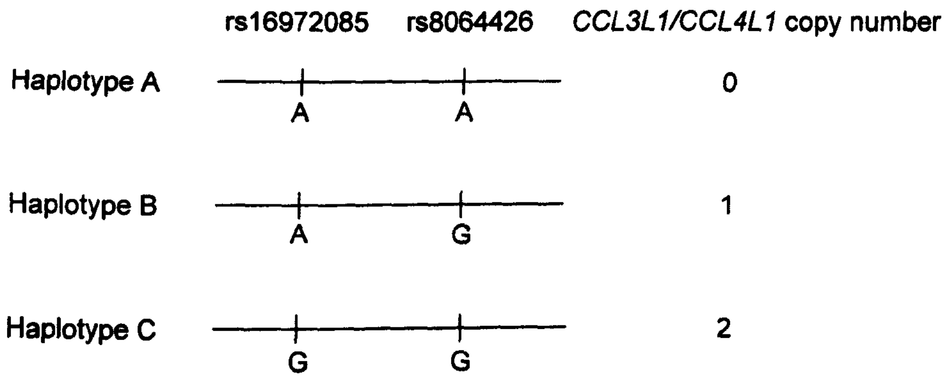


Figure 4.1: The prediction of *CCL3L1/CCL4L1* copy number haplotypes by using a combination SNPs of rs16972085 and rs8064426.

Table 4.12: genotypes of a combination of SNPs for prediction of *CCL3L1/CCL4L1* copy number.

rs16972085 (genotype)	**rs8064426 (genotype)**	***CCL3L1/CCL4L1* copy number**
AA	AA	0
AA	AG	1
AA	GG	2 (1/1)
AG	GG	3
GG	GG	4
AG	AG	2 (2/0)

4.2.5 Prediction of *CCL3L1/CCL4L1* copy number in ECACC and Basque samples

Using a combination of SNPs, rs16972085 and rs8064426, to predict the copy number of *CCL3L1/CCL4L1* in 192 ECACC and 157 Basque samples, they were approximately 70% correct (*i.e.* 72.34% and 71.43%, respectively). The data of *CCL3L1/CCL4L1* copy number prediction in ECACC and Basque samples are shown in Appendix.

4.3 *CCL3* microsatellite analysis

4.3.1 Introduction

Microsatellites, as genetic markers, are numerous and found throughout the human genome. Due to their polymorphic nature, microsatellites are generally informative, and as a consequence, their analysis has been applied in many fields, as mentioned in the introduction 1.3 (Goldstein and Schlötterer 1999).

Recently, microsatellites were found to be associated with copy number variants, and might have a role in the formation of copy number variants (Kim *et al.* 2008). Moreover, when analysis of a microsatellite is combined SNP genotyping, it will help to extend SNPs and show more detailed LD than the pair of SNPs or microsatellites alone (Payseur *et al.* 2008). Therefore, integration of these two genetic markers might help to improve LD of copy number variant in duplication-rich regions of the genome which may exist on multiple haplotypic backgrounds (Locke *et al.* 2006).

4.3.2 Association between *CCL3* microsatellite and the copy number of *CCL3L1/CCL4L1*

Here the compound microsatellite (CCL3CT) at *CCL3* region, which was designed by Susan Walker for investigation of the copy number variation at *CCL3* (Walker 2009), was applied to study the association with the copy number of *CCL3L1/CCL4L1* in 192 ECACC samples. The compound microsatellite (CCL3CT) is composed of CT-rich microsatellite with predominant repeat units of CTTT, CCTT, CCCT and CT. It is located about 99kb from the location of *CCL3L1* and its sequence as recorded in the human genome sequence assembly is shown in Figure 4.2.

ACCATGACAACTAGAAGATGGTCTtagcaactccctt$(ccct)_2$ccctt$(ccct)_4(cctt)_2$t$(ct)_3$
$(cttt)_3$tt$(cttt)_6$c$(cttt)_2(ct)_3$tt$(ct)_5$c$(cctt)_{16}(cttt)_{34}$tcttccaagacaaggtctcattctgtggcccagg
ctggagtgcaGTGACTCGATCTTGGCTCAC

Figure 4.2: The sequence of CCL3CT microsatellite.

At Chr.17:31446870-31447279 [UCSC: Mar.2006 (NCBI36/HG18)].

After the microsatellite amplification products were separated by capillary electrophoresis on an ABI3100, 180 out of 192 ECACC samples had two peaks with ratio 1:1, and 8 out of 192 had one peak. Although these eight samples were not investigated further as to whether they possess one or two copies of *CCL3*, this work implies that most UK samples have two copies of *CCL3*, and these samples do not show a copy number variation at the *CCL3* region.

Surprisingly, there were 4 samples that showed more than 2 peaks, which were C0896 and C0210 from the ECACC panel 1 and C0164 and C0990 from the ECACC panel 2. However, after they were re-analyzed again using the selected microsatellites, D18S51 and D21S11, which were used for individual identification (Kimpton *et al.* 1996), the results showed that C0896 and C0210 might have somatic mutation (Figure 4.3), and C0164 and C0990 were mixed samples (Figure 4.4).

Excluding these 4 samples, the assay showed there was extensive allelic diversity of CCL3CT microsatellite in this region. The microsatellite PCR products were found in two ranges, 270bp to 301bp and 341bp to 464bp, as shown in Figure 4.5.

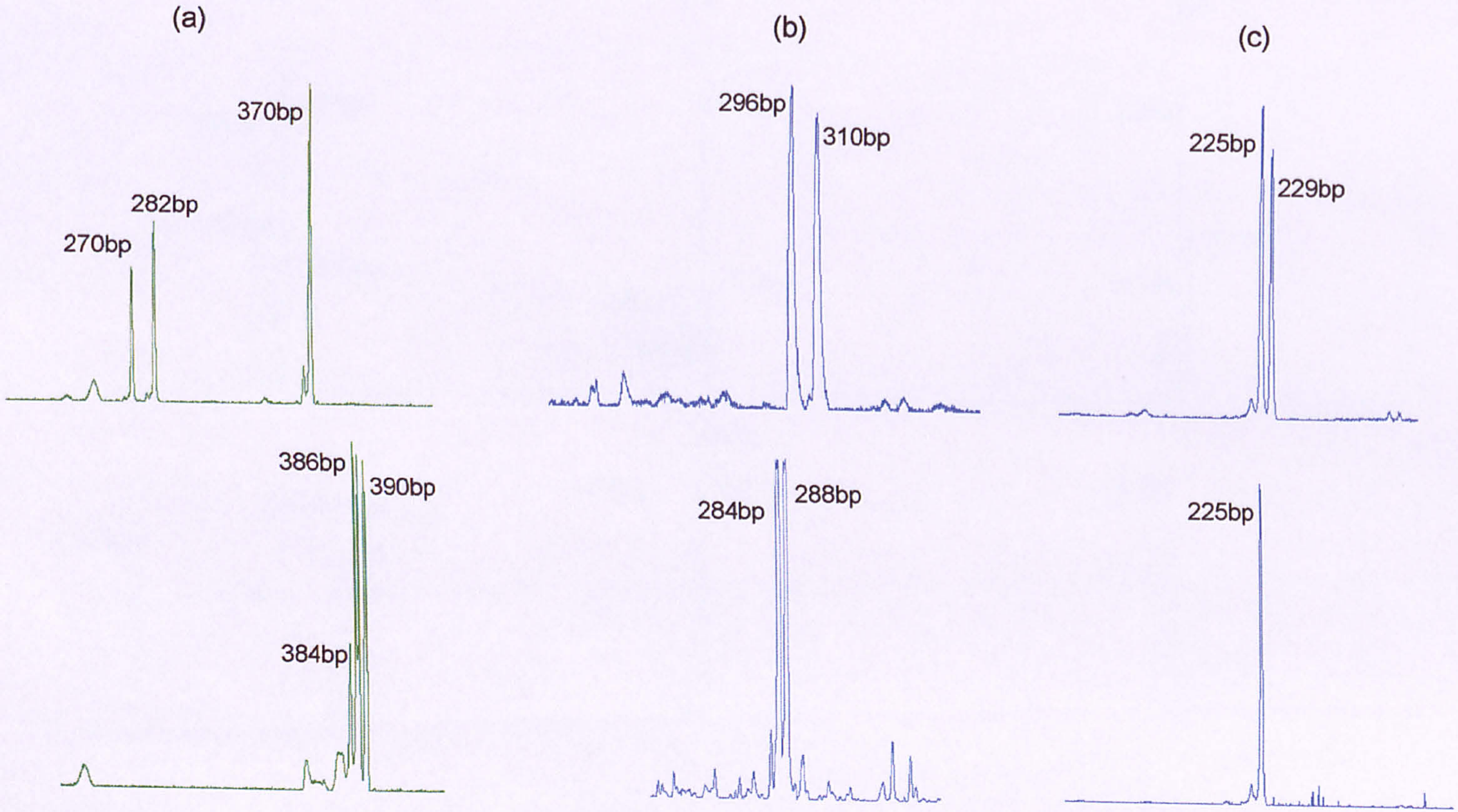


Figure 4.3: Microsatellite traces in C0896 (upper panel) and C0210 (bottom panel) which may have been the result of somatic mutation at the CCL3CT microsatellite. (a) CCL3CT (b) D18S51 and (c) D21S11

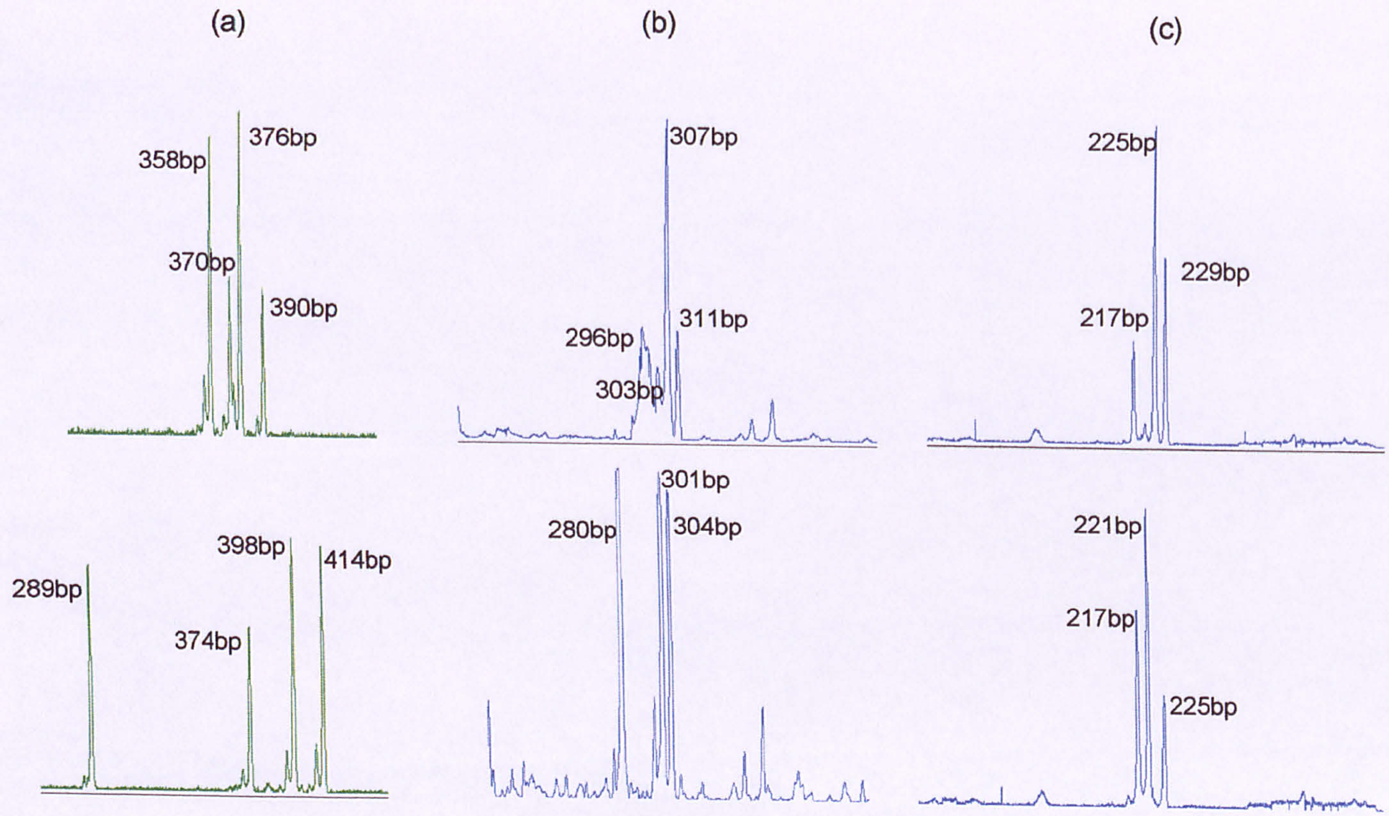


Figure 4.4: Microsatellite traces in C0164 (upper panel) and C0990 (bottom panel) which may be mixed samples.

(a) CCL3CT (b) D18S51 and (c) D21S11

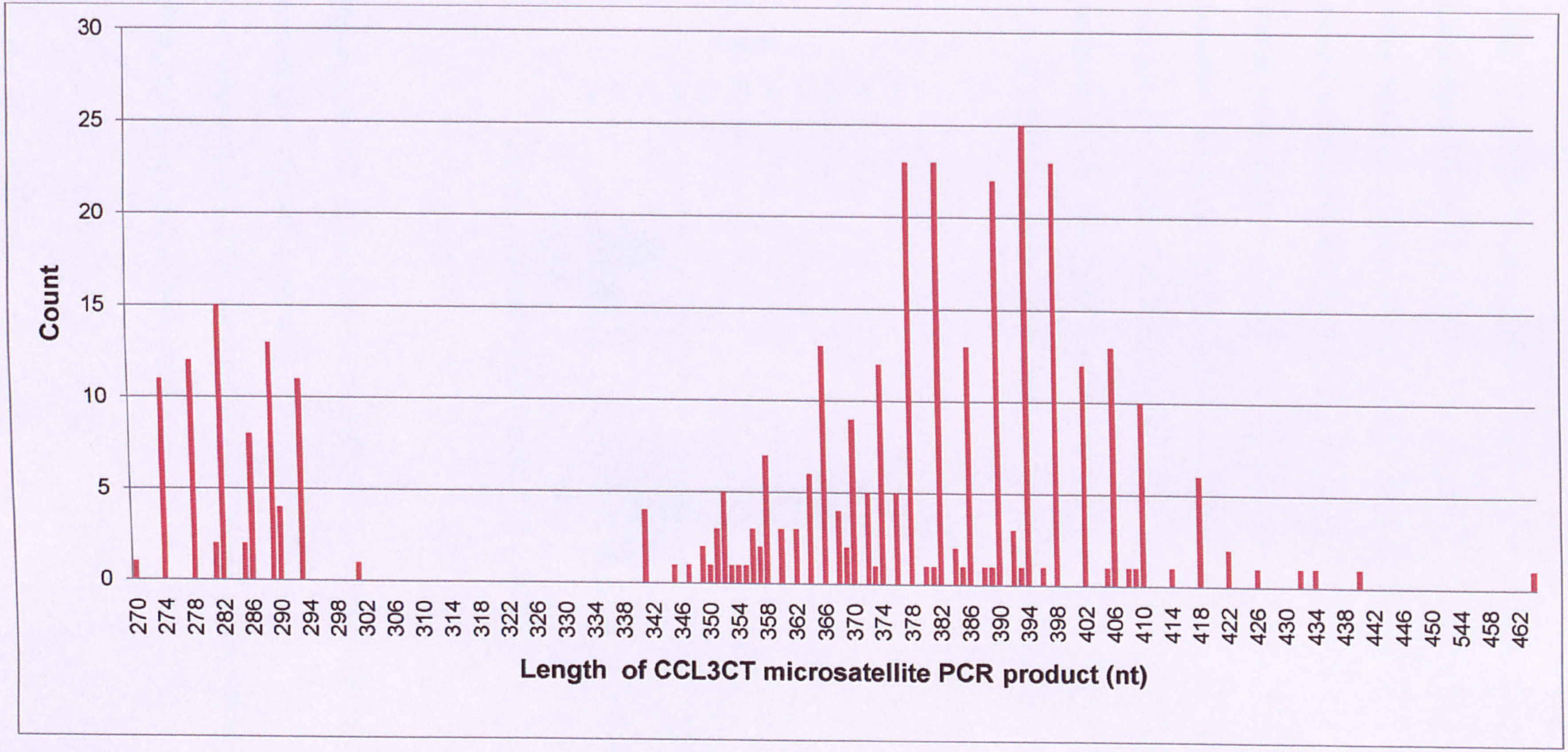


Figure 4.5: The distribution of CCL3CT microsatellite in ECACC samples (n=188).

After CCL3CT microsatellite alleles were categorized by their associated *CCL3L1/CCL4L1* copy number, their allele patterns do not differ between the groups. The alleles were still found in two ranges, except in samples of zero copy number of *CCL3L1/CCL4L1*, in which alleles were found only within the range of microsatellite PCR products between 341bp and 464bp. However, this observation for zero copy number of *CCL3L1/CCL4L1* samples still needs further investigation because there are only two samples of zero copy number of *CCL3L1/CCL4L1* in this study. The data are shown in Figure 4.6.

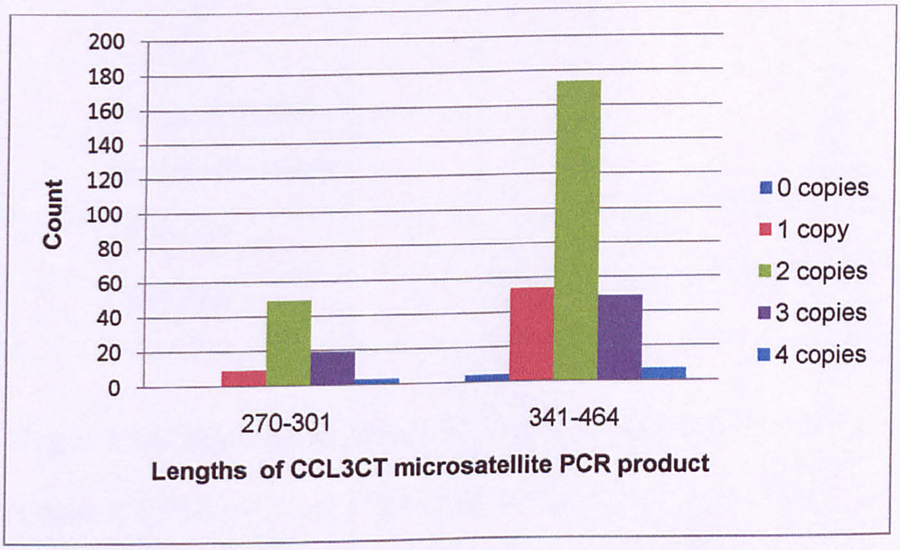


Figure 4.6: The comparison of CCL3CT microsatellite alleles grouped according to the copy number of *CCL3L1/CCL4L1*.

Two ranges of CCL3CT microsatellite PCR product lengths are seen, 270bp to 301bp and 341bp to 464bp.

To investigate the association between *CCL3* microsatellite and the copy number of *CCL3L1/CCL4L1*, CCL3CT microsatellite alleles were grouped by the ranges of CCL3CT microsatellite PCR product length and the copy number of *CCL3L1/CCL4L1*. The data are shown in Tables 4.13 and 4.14.

Table 4.13: The number of CCL3CT microsatellite alleles in each group of *CCL3L1/CCL4L1* copy number.

Copy number of *CCL3L1/CCL4L1* / Range of CL3CT microsatellite alleles	0	1	2	3	4
270-301	0	9	49	19	3
341-464	4	54	174	49	7

Table 4.14: The number of CCL3CT microsatellite alleles in each group of *CCL3L1/CCL4L1* copy number (grouped for χ^2 test).

Copy number of *CCL3L1/CCL4L1* / Range of CCL3CT microsatellite alleles	0-1	2	3-4
270-301	9	49	22
341-464	58	174	56

However, after the association was tested using Chi-square analysis, the results showed that there is no significant relationship between CCL3CT microsatellite alleles and *CCL3L1/CCL4L1* copy number (p = 0.0982) (n=188).

4.3.3 Association between *CCL3* microsatellite and *CCL3L1* pseudogene

In addition, the association between *CCL3* microsatellite genotype and *CCL3L1* pseudogene (presence/absence) was also examined in 188 ECACC samples. The CCL3CT microsatellite alleles were categorized into two groups, according to the presence of *CCL3L1* pseudogene group and the absence of *CCL3L1* pseudogene group. The results showed that the patterns of microsatellite alleles do not differ between the groups, and they were found in two ranges, 270bp to 301bp and 341bp to 464bp, the same as that was found in the study of the association between *CCL3* microsatellite and *CCL3L1/CCL4L1* copy number. The data are shown in Figure 4.7 and Table 4.15.

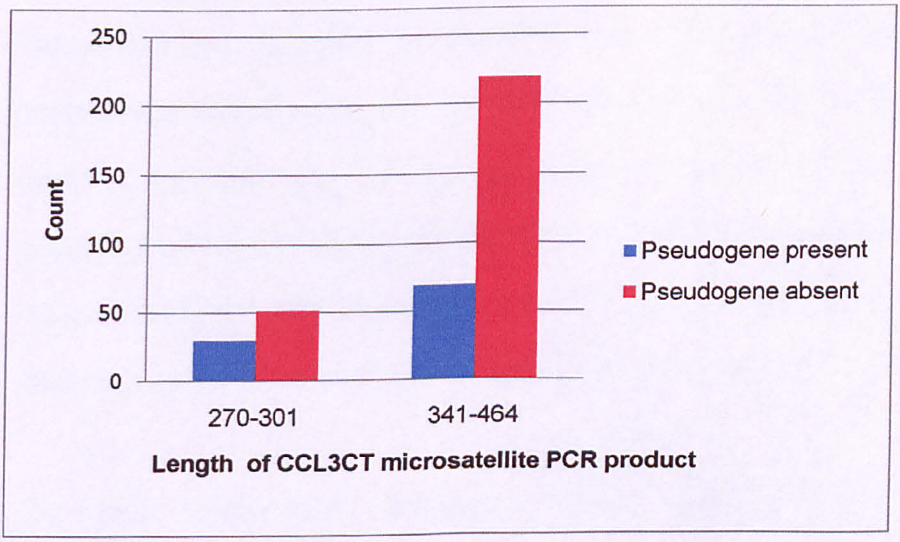


Figure 4.7: The comparison of CCL3CT microsatellite alleles grouped according to the presence of pseudogene.

Two ranges of CCL3CT microsatellite PCR product lengths are seen, 270bp to 301bp and 341bp to 464bp.

Table 4.15: The number of CCL3CT microsatellite alleles in each group of *CCL3L1* pseudogene.

Pseudogene / Range of CCL3CT microsatellite alleles	Presence	Absence
270-301	29	51
341-464	69	219

A χ^2 shows that there is a significant relationship between CCL3CT microsatellite alleles and *CCL3L1* pseudogene (p = 0.0277).

To summarize, CCL3CT microsatellite has not found to have a relationship with the copy number of *CCL3L1/CCL4L1*, but it shows a weak association with *CCL3L1* pseudogene. However, this might suggest that the CCL3CT microsatellite has a high rate of mutation, the *CCL3L1/CCL4L1* CNV has a high rate of mutation, and/or there has been high recombination at the *CCL3L1* and *CCL4L1* region.

As the association between SNPs, microsatellites and *CCL3L1/CCL4L1* copy number were not found, there is no one marker that can be used to predict the *CCL3L1/CCL4L1* copy number. One plausible hypothesis is that *CCL3L1/CCL4L1* copy numbers might have different haplotype backgrounds. As a result, the relationship between single-copy markers and *CCL3L1/CCL4L1* will be reduced. Therefore, to overcome this, the haplotype of *CCL3L1/CCL4L1* was identified and tested for association with *CCL3L1/CCL4L1* copy number again using LD analysis. Moreover, the defined haplotypes of *CCL3L1/CCL4L1* were also analysed for evolutionary history of variation in this region by comparison to chimpanzee. More details and results are shown in the next chapter.

Chapter 5: Linkage disequilibrium and haplotype block analysis at *CCL3L1* and *CCL4L1*

5.1 Introduction

Since LD was characterized in the human genome (Reich *et al.* 2001), and the International HapMap Project provided an invaluable human variant resource (The International HapMap Consortium 2005, 2007; The International HapMap 3 Consortium 2010), LD has been a powerful genetic tool to investigate direct and indirect gene-disease mapping, especially in genome wide association studies (GWAS) as mentioned in the introduction. As a result, LD was also exploited as a tool to search for a tag-SNP which can be used to determine *CCL3L1/CCL4L1* copy number.

In this first approach, studies of association between SNPs, microsatellites and *CCL3L1/CCL4L1* copy number, did not provide satisfactory results. So, LD was applied to find a tag-SNP for determination of *CCL3L1/CCL4L1* copy number using CEPH-HapMap samples.

Unfortunately, although suggested tag-SNP from LD analysis provided significant associations in ECACC and Basque samples, using them to determine *CCL3L1/CCL4L1* copy number in these samples was only 70% concordant. A new method, therefore, was needed to improve this association study.

Promisingly, haplotype-based tests of association were introduced that might be able to improve the association study, since it could provide greater power than using single genetic markers (Clark 2004; de Bakker *et al.* 2005). Furthermore, haplotype-based tests of association have been successful in gene mapping, such as at the APOE locus in Alzheimer's disease (Fallin *et al.* 2001), the *P-selectin* gene and myocardial infarction (Tregouet *et al.* 2002), and the Renin-angiotensin-aldosterone system genes and hypertension (Niu *et al.* 2009).

Therefore, sequence haplotype-based tests were applied to further investigate a SNP or a haplotype block, which showed a strong association with *CCL3L1/CCL4L1* copy number haplotypes, in the CEPH-HapMap samples using LD.

The application of phased haplotypes and thus LD in this association study should not only provide SNPs which could be used to determine the *CCL3L1/CCL4L1* copy number haplotype, but it also could provide evidence of the genomic evolution of *CCL3L1/CCL4L1* segment as illustrated in Figure 5.1.

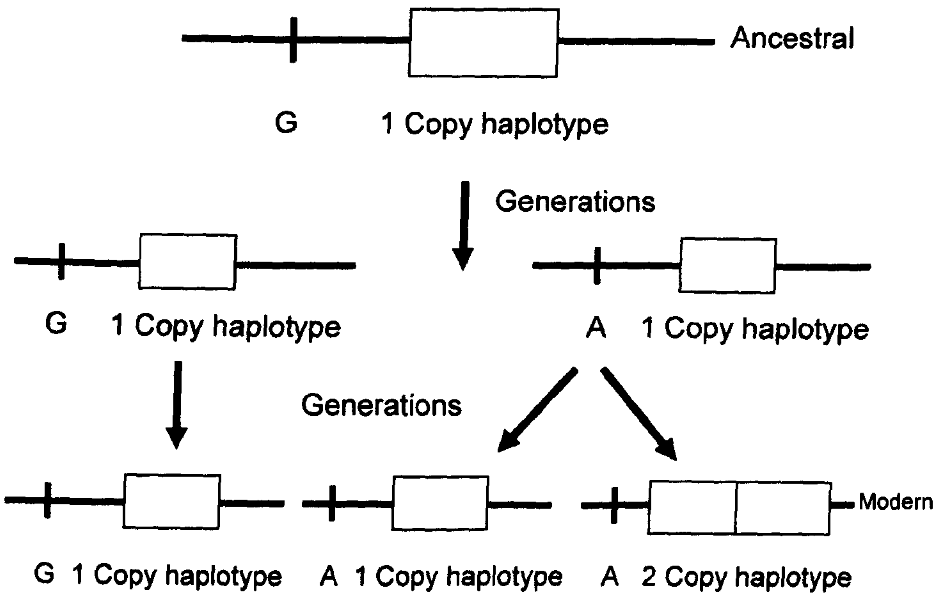


Figure 5.1: Testing the association between hypothetical SNPs and *CCL3L1/CCL4L1* haplotypes on the basis of linkage disequilibrium (LD). The boxes represent to copy number haplotype, and "A" and "G" are SNPs. If the association between SNPs and *CCL3L1/CCL4L1* haplotype was not confounded by recombination or recurrent mutation of *CCL3L1/CCL4L1* haplotype, and/or reverse mutation of *CCL3L1/CCL4L1* haplotype, the study would show the relationship between SNPs and *CCL3L1/CCL4L1* haplotypes when the SNPs were typed and tested for association with *CCL3L1/CCL4L1* haplotypes. As illustrated in Figure 5.1 suppose base G was the ancestral state and was associated with a 1 copy haplotype. After many generations, a SNP was created, and the base G in one lineage was changed to base A. Again, after further generations, the CNV occurred. SNP genotyping would show a relationship with CNV as in the figure; base G would correlate with 1 copy haplotype, and base A would be associated with 2 copy haplotype.

So, by identify tag-SNPs to determine *CCL3L1/CCL4L1* haplotypes, the evolution of *CCL3L1/CCL4L1* haplotypes could be identified as well.

To generate the *CCL3L1* and *CCL4L1* sequence haplotypes in CEPH HapMap samples, however, several techniques were necessary; sequencing, segregation analysis, allele-specific PCR and emulsion-fusion PCR because sequencing alone cannot identify the *CCL3L1/CCL4L1* sequence haplotype in some samples. For example, for samples which have two or more copies of *CCL3L1/CCL4L1*, sequencing cannot provide the information of which variant bases occur together as a haplotype. Therefore, the samples also needed segregation analysis and/or allele-specific PCR to establish the *CCL3L1/CCL4L1* haplotype. Lastly, some samples also required emulsion-fusion PCR to combine both *CCL3L1* and *CCL4L1* haplotypes together to generate complete haplotypes.

5.2 Sequence analysis at *CCL3L1* and *CCL4L1*

5.2.1 Introduction

CCL3L1 and *CCL4L1* sequencing was performed for the both genes in 90 CEPH HapMap samples, to identify the sequence variants in both loci. The primers and conditions for *CCL3L1* and *CCL4L1* PCR amplification and sequencing are shown in Tables 2.3-2.5 in section 2.2.4.

5.2.2 *CCL3L1* sequencing

CCL3L1 is a copy number variable gene which is paralogous to *CCL3*, and it has a partial repeat, usually called the *CCL3L1* pseudogene. Therefore, to read the sequence of *CCL3L1*, the *CCL3L1* PCR amplification primers were especially designed to be specific to *CCL3L1* to avoid co-amplification with *CCL3* or *CCL3L1* pseudogene, and they were also designed to avoid common variants themselves.

To complete the whole sequence of *CCL3L1* (about 1.9kb in size), a 2.1kb PCR product of *CCL3L1* was used for sequencing. However, the length of these *CCL3L1* PCR products was too long for a single sequencing reaction. Thus, nested primers were also designed (to use in sequencing PCR) to obtain full length sequence of *CCL3L1*. The *CCL3L1* sequencing assay and sequence variant bases are shown in Figure 5.2.

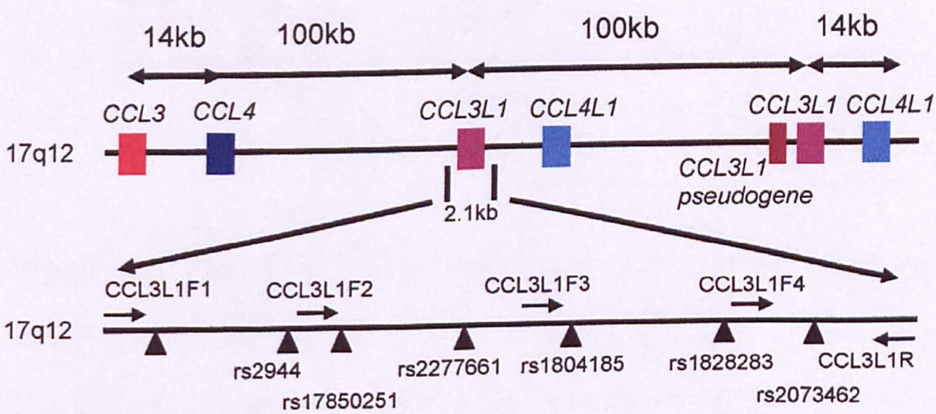


Figure 5.2: The *CCL3L1* sequencing assay and discovered sequence variant bases at *CCL3L1* (Chr.17:31546177-31548296 and 31647751-31649870). For simplicity, one repeat unit is shown.

Sequencing identified 7 common variant bases, and 6 out of 7 were validated in the SNP database (dbSNP129). The positions of the common sequence variant base are described in section 5.6.

5.2.3 *CCL4L1* sequencing

CCL4L1 sequencing was performed similarly to *CCL3L1* sequencing. The *CCL4L1* PCR amplification primers were specifically designed for *CCL4L1*, to avoid co-amplification with the paralogous *CCL4*, and *CCL4L1*. PCR products for sequencing were produced (about 2.5kb) to cover the whole gene sequence (about 1.8kb). The *CCL4L1* sequencing assay and sequence variant bases are shown in Figure 5.3.

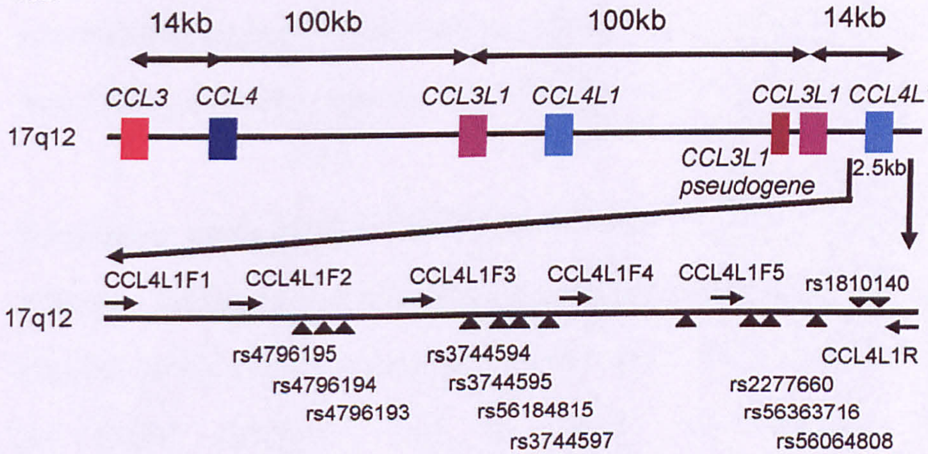


Figure 5.3: The *CCL4L1* sequencing assay and discovered sequence variant bases at *CCL4L1* (Chr.17: 31663731-31666272).

The *CCL4L1* sequencing identified 13 common variant bases, and 11 out of 13 were validated in the SNP database (dbSNP129). The positions of the common sequence variant bases are described in section 5.6.

5.3 Segregation analysis

Segregation analysis was performed to determine the copy number of *CCL3L1/CCL4L1* per chromosome, and it was also used to confirm the copy number of *CCL3L1/CCL4L1* typed by the PRT technique. The data on *CCL3L1/CCL4L1* copy number inheritance transmission in CEPH-HapMap used in this study were carried out and kindly provided by Dr. Danielle Carpenter. In brief, the *CCL3L1/CCL4L1* copy number inheritance transmission patterns were inferred by segregation of microsatellites within the copy variable region as described in Walker (2009).

However, there were some CEPH-HapMap families in which microsatellite segregation analysis could not provide the information of their *CCL3L1/CCL4L1* copy number inheritance transmission pattern.

Fortunately, some of these CEPH-HapMap families were the same as reference families from the Centre d'Etude du Polymorphisme Humain (CEPH), which can be investigated more closely using markers from the CEPH database. Thus, to deduce unambiguously the *CCL3L1/CCL4L1* copy number haplotypes in these families, more samples from these CEPH families were selected to type their *CCL3L1/CCL4L1* copy number and to do segregation analysis. This segregation analysis was carried out using three selected flanking markers from the CEPH database; they were AFM179xg11 / (AC)n, UT752 / pcr(n) and Mfd15-2 / (AC)25. These three markers were located flanking the region of *CCL3L1/CCL4L1* on Chr.17: 28313925-28314303, 30890413-30890839 and 34405617-34405867, respectively. *CCL3L1/CCL41* is located at Chr.17: 31546382-31665958. An example is CEPH-HapMap family 1341 in which the *CCL3L1/CCL4L1* copy number inheritance pattern in the maternal line cannot be determined by segregation since all members, including the grandfather, grandmother and mother, have 2 copies of *CCL3L1/CCL4L1*, and all have the same pattern of microsatellite length. Therefore, the chromosome copy number of *CCL3L1/CCL4L1* in these trios could be either 2/0 or 1/1. Consequently, more children were selected to type their *CCL3L1/CCL4L1* copy number as shown in Figure 5.4.

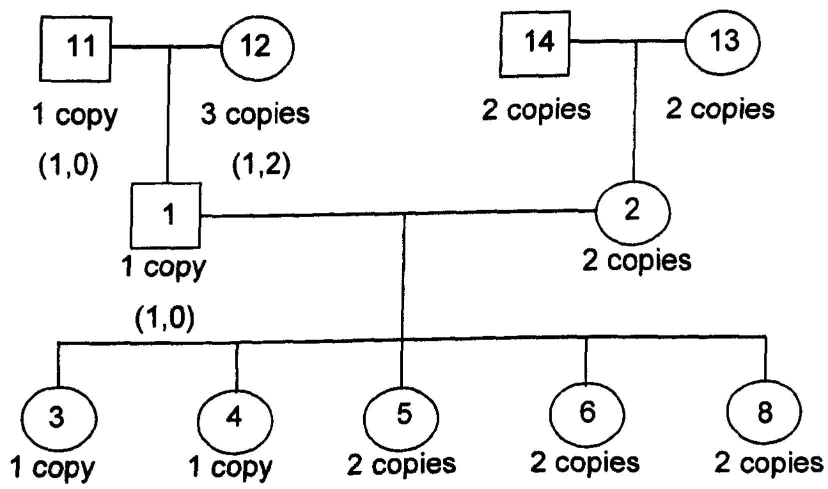


Figure 5.4: The pedigree of CEPH family 1341 and their *CCL3L1/CCL4L1* copy number.

The numbers 11, 12, 1, 13, 14 and 2 in this pedigree are NA07034, NA07055, NA07048, NA06993, NA06985 and NA06991 in CEPH-HapMap samples, respectively. The numbers in the bracket are chromosomal copy numbers of *CCL3L1/CCL4L1*.

The children selected to be typed for *CCL3L1/CCL4L1* copy number were children who showed different pattern of segregation, and their patterns of transmission inheritance do not change between the three flanking markers. The information provided by the three markers indicated that this DNA segment defined by these markers had been transmitted as a whole, with no evidence for recombination. The pedigree of this family and their genotypes for the three markers are shown in Figures 5.5 to Figure 5.7.

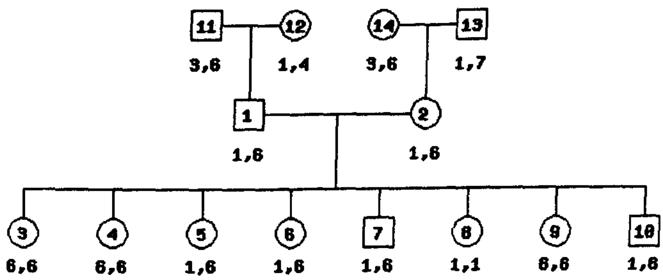


Figure 5.5: CEPH family 1341 and their AFM179xg11 / (AC)n genotypes.

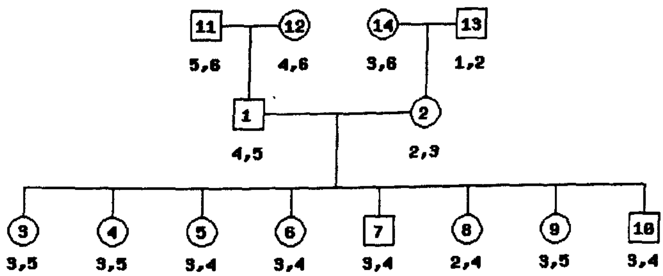


Figure 5.6: CEPH family 1341 and their UT752 / pcr(n) genotypes.

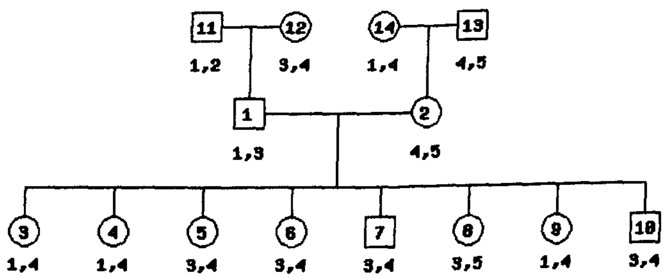


Figure 5.7: CEPH family 1341 and their Mfd15-2 / (AC)25 genotypes.

To summarize, after more children were typed for *CCL3L1/CCL4L1* copy number, the segregation patterns were examined in three selected markers and inferred that all grandfather, grandmother and mother have chromosome copy number as 1/1. This is because if mother (number 2 in the Figure 5.4) has the chromosome copy number as 2/0, her children should have the characteristics as follows,

- Based on segregation in these 3 markers, child number 8 should have 3 or 1 copies of *CCL3L1/CCL4L1* copy number and children numbers 3 and 4 should have 2 or 0 copy of *CCL3L1/CCL4L1* copy number.
- Based on segregation in markers UT752 / pcr(n) and Mfd15-2 / (AC)25, children numbers 5 and 6 should have 3 or 1 copies of *CCL3L1/CCL4L1* copy number.

However, the *CCL3L1/CCL4L1* copy number was found 1, 1, 2, 2 and 2 in children numbers 3, 4, 5, 6 and 8, respectively. So, apparently, this mother (number 2 in the Figure 5.4) should have the chromosome copy number 1/1, and this implies that her parents should have the chromosome copy number 1/1, as well. In addition, the *CCL3L1/CCL4L1* copy number and its inheritance pattern were confirmed by the repetition of the results in the pairs of children, numbers 3 and 4 and 5 and 6.

5.4 Allele-specific PCR assays

An allele-specific PCR technique was applied to help deduce *CCL3L1* and *CCL4L1* haplotypes in CEPH-HapMap samples whose haplotypes could not be inferred by segregation analysis.

An example is CEPH-HapMap family 1334, which comprises father (NA12146), child (NA10847) and mother (NA12239). Using PRT, these individuals possess 3, 2 and 1 copies of *CCL3L1* respectively, as shown in Table 5.1. Based on segregation analysis, the *CCL3L1* haplotypes of mother and child can be established. Since the mother possesses 1 copy, the transmission of this haplotype to the child allows the retrospective identification of which haplotype is transmitted from father. Further dissection of the father's 2 other copies of *CCL3L1* into a haplotype is difficult. From sequencing alone it is not known which variant bases at positions 7837, 8282, 8382 and 9613 occur together as *CCL3L1* haplotypes. Consequently, allele specific PCR assays were applied to deduce the *CCL3L1* haplotypes in this sample. After amplification with allele specific primer at position 7837, the allele specific PCR products were sequenced to reveal the variant bases on the same haplotype as the discriminatory base used in the allele specific primers. The sequencing results of allele specific PCR products of this sample are shown in Figures 5.8 and 5.9, and the *CCL3L1* haplotypes of this sample are shown in Table 5.2.

Table 5.1: Sequence variant bases of CEPH-HapMap family 1334 before using allele specific PCR.

Sample	***CCL3L1* CN**	**Sequence Variant (SNP129) UCSC:Mar.2006 (NCBI36/HG18); Chr.17:31647837-31649613* [for simplicity 1 browser position is shown]**						
			rs2944	rs17850251	rs2277661	rs1804185	rs1828283	rs2073462
		7837**	8282	8382	8673	8952	9366	9613
NA12146	1 copy Chr.	T	A	A	C	C	G	G
	2 copy Chr.	CT	AG	AG	TT	CC	CC	AG
NA10847	1 copy Chr.	T	A	A	C	C	G	G
	1 copy Chr.	T	A	A	T	T	C	G
NA12239	0 copy Chr.	-	-	-	-	-	-	-
	1 copy Chr.	T	A	A	T	T	C	G

Note:
1. NA12146 is father, NA10847 is a child, and NA12239 is mother, respectively.
2. The red and blue shading represent the chromosome transmitted from father and mother, respectively.
3. The yellow shading represent to the variant bases which cannot be determined to reveal *CCL3L1* haplotypes using segregation analysis
4. *The locations of variant bases are provided as the last 4 digit-number under the rs number. For example, the variant base of rs2944 in this table is located at Chr17: 31648282.
5. **This variant base position is not included in dbSNP129.

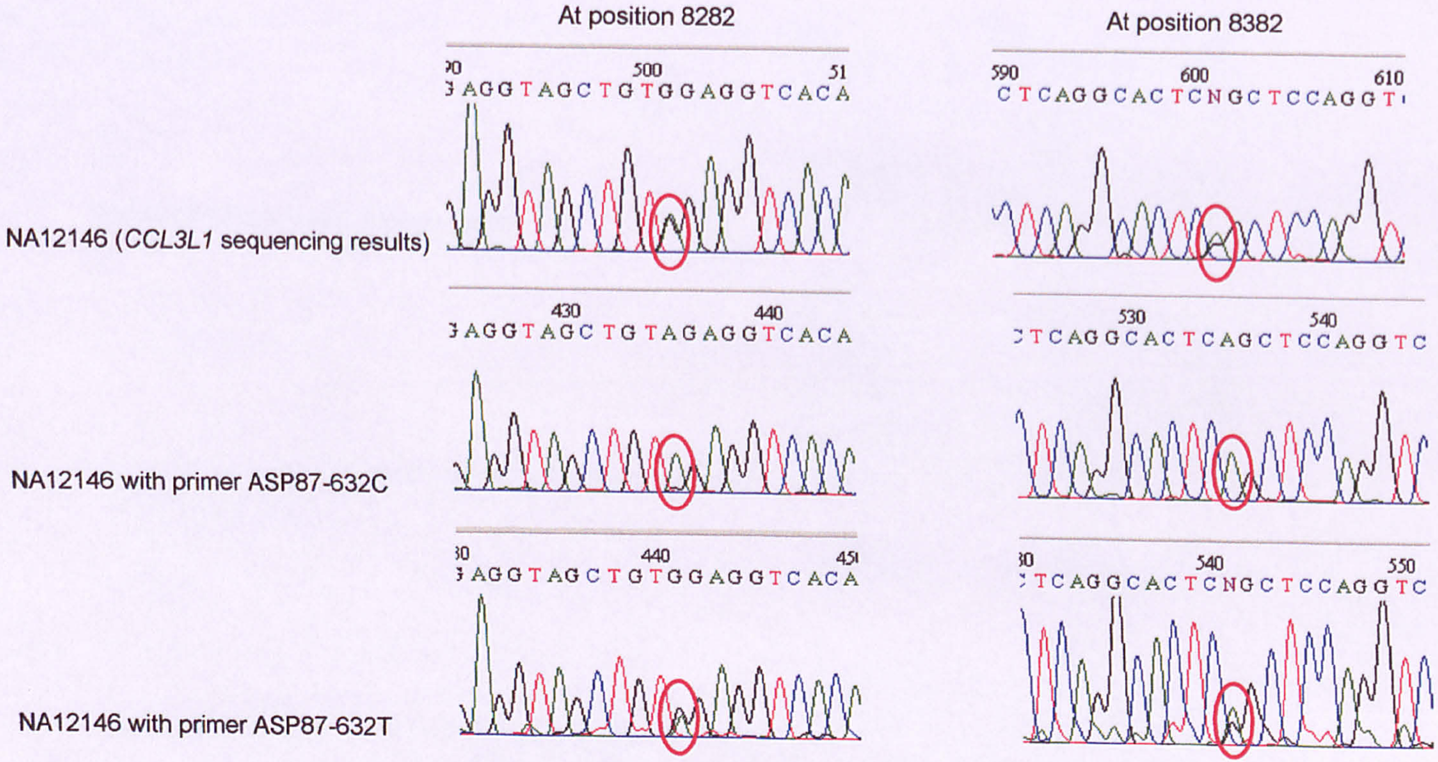
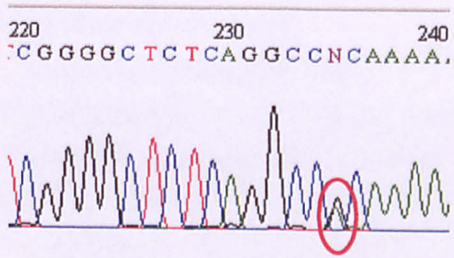


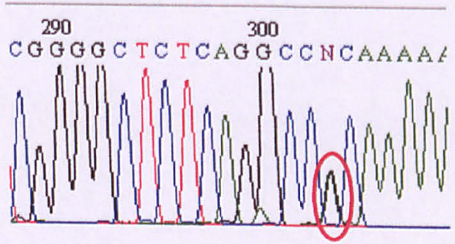
Figure 5.8: Comparison of sequencing results at position 8282 and 8382 of *CCL3L1* in sample NA12146.

NA12146 has 3 copies of *CCL3L1*. Sequencing shows the presence of variant bases, AAG and AAG at position 8282 and 8382, respectively (top panels). Allele specific primers designed to amplify a C and T at position 7837 demonstrate this C occurs with both A at positions 8282 and 8382, and T at position 7837 occurs with A and G at both positions 8282 and 8382.

NA12146 (*CCL3L1* sequencing results)



NA12146 with primer ASP632-1863A (with additional primer)



NA12146 with primer ASP632-1863G (with additional primer)

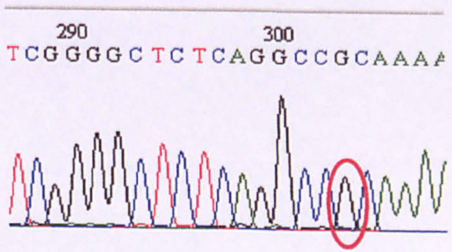


Figure 5.9: Comparison of sequencing results at position 9613 of *CCL3L1* in sample NA12146.

NA12146 has 3 copies of *CCL3L1*. Sequencing shows the presence of variant bases AGG at position 9613 (top panel). Allele specific primers designed to amplify an A and a G at position 8382 demonstrate this A occurs with A and G at positions 9613, and G at position 8382 occurs with G at position 9613.

Table 5.2: Sequence variant bases of CEPH-HapMap family 1334 after using allele specific PCR.

Sample	***CCL3L1* CN**	**Sequence Variant (SNP129) UCSC:Mar.2006 (NCBI36/HG18); Chr.17:31647837-31649613* [for simplicity 1 browser position is shown]**						
			rs2944	rs17850251	rs2277661	rs1804185	rs1828283	rs2073462
		7837**	8282	8382	8673	8952	9366	9613
NA12146	1 copy Chr.	T	A	A	C	C	G	G
	2 copy Chr.	C	A	A	T	C	C	A
		T	G	G	T	C	C	G
NA10847	1 copy Chr.	T	A	A	C	C	G	G
	1 copy Chr.	T	A	A	T	T	C	G
NA12239	0 copy Chr.	-	-	-	-	-	-	-
	1 copy Chr.	T	A	A	T	T	C	G

Note:
1. NA12146 is father, NA10847 is a child, and NA12239 is mother, respectively.
2. The red and blue shading represent the chromosome transmitted from father and mother, respectively.
3. The yellow shading represent to the variant bases which are deduced *CCL3L1* haplotypes using allele-specific PCR.
4. *The locations of variant bases are provided as the last 4 digit-number under the rs number. For example, the variant base of rs2944 in this table is located at Chr17: 31648282.
5. **This variant base position is not included in dbSNP129.

As is the case for NA12146, additional CEPH HapMap samples whose *CCL3L1/CCL4L1* haplotypes could not be identified by segregation analysis were examined using their allele specific PCR assays. Primers are shown in section 2.5. A combination of allele-specific PCRs was used to provide complete coverage of *CCL3L1* and *CCL4L1*.

5.5 Emulsion haplotype fusion PCR

Emulsion-fusion PCR was used to make a condensed haplotype of sequence variants in *CCL3L1* with sequence variants in *CCL4L1* from the same repeat unit in CEPH-HapMap samples whose phase haplotypes of *CCL3L1/CCL4L1* could not be inferred by segregation analysis alone. This included examples that possess two-copy haplotypes of *CCL3L1/CCL4L1*.

An example is CEPH-HapMap family 1408, which comprises father (NA12155), child (NA10831) and mother (NA12156). Using PRT, these individuals possess 2, 3 and 1 copies of *CCL3L1/CCL4L1* respectively. Based on segregation analysis, the transmission pattern of *CCL3L1/CCL4L1* copy number in this family is shown in Table 5.3. In this family, transmission of the 2-copy *CCL3L1/CCL4L1* haplotype from father to child is ambiguous. The combination of sequence variants in *CCL3L1* that go together with sequence variants in the neighbouring copy of *CCL4L1* in each of the two repeats cannot be determined by segregation analysis alone. As a result, emulsion-fusion PCR was used to combine the haplotype of *CCL3L1* and *CCL4L1*, and then the *CCL3L1/CCL4L1* variant combinations in each repeat unit were revealed using allele-specific PCR as shown in Figure 5.10 and Table 5.4.

Table 5.3: Phase haplotype of *CCL3L1/CCL4L1* of CEPH-HapMap family 1408 before using emulsion-fusion PCR.

Sample	*CCL3L1/CCL4L1* copy number	Sequence Variant (SNP129) UCSC:Mar.2006 (NCBI36/HG18) [for simplicity 1 browser position is shown]				
		*CCL3L1**		*CCL4L1****		
		rs2944	rs1804185	rs4796195	rs4796194	rs4796193
		Chr.17:31648282	Chr.17:31648952	Chr.17:31665561	Chr.17:31665424	Chr.17:31665423
NA12155	2 copy	G	C	C	C	G
		A	T	T	T	G
	0 copy	-	-	-	-	-
NA10831	2 copy	G	C	C	C	G
		A	T	T	T	G
	1 copy	A	T	T	T	C
NA12156	0 copy	-	-	-	-	-
	1 copy	A	T	T	T	C

Note: 1. NA12155 is father, NA10831 is a child, and NA12156 is mother, respectively.

2. The red and blue shading represent the chromosome transmitted from father and mother, respectively.

3. The yellow shading represents to the ambiguous phase haplotypes of *CCL3L1/CCL4L1*.

4. *These are not a full haplotype.

5. ***CCL4L1* sequencing results are shown in negative strand.

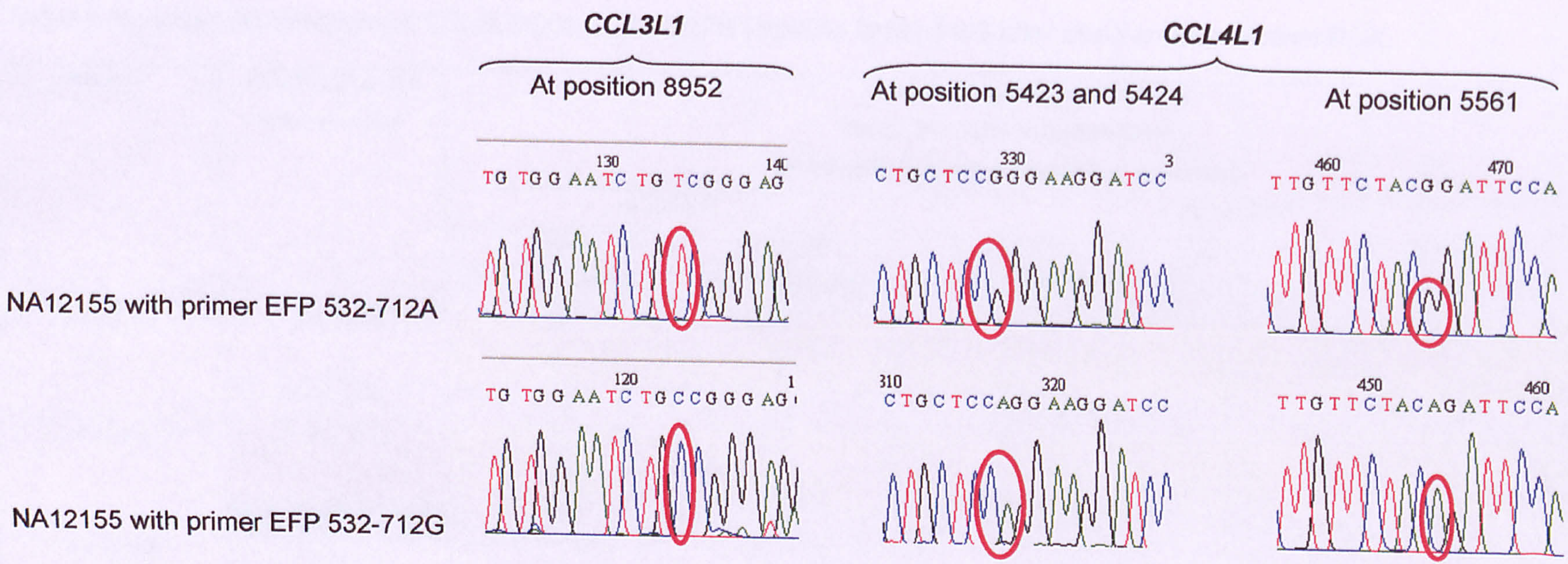


Figure 5.10: Sequencing results reveal the variant combinations of *CCL3L1/CCL4L1* in sample NA12155. NA12155 has 2 copies of *CCL3L1/CCL4L1*. After performing emulsion-fusion PCR, allele-specific primers designed to amplify either an A or a G at position 8282 demonstrate this A occurs with T at positions 8952 of *CCL3L1* and with C, G and G at positions 5423, 5424 and 5561 of *CCL4L1*, respectively. The G at position 8282 occurs with C at positions 8952 of *CCL3L1* and C, A and A at positions 5423, 5424 and 5561 of *CCL4L1*, respectively (Note: *CCL4L1* sequencing results are shown in positive strand).

Table 5.4: Variant combinations of *CCL3L1/CCL4L1* of CEPH-HapMap family 1408 after using emulsion-fusion PCR.

Sample	***CCL3L1/CCL4L1* copy number**	**Sequence Variant (SNP129) UCSC:Mar.2006 (NCBI36/HG18) [for simplicity 1 browser position is shown]**				
		CCL3L1*		***CCL4L1*****		
		rs2944	rs1804185	rs4796195	rs4796194	rs4796193
		Chr.17:31648282	Chr.17:31648952	Chr.17:31665561	Chr.17:31665424	Chr.17:31665423
NA12155	2 copy Chr.	G	C	T	T	G
		A	T	C	C	G
	0 copy Chr.	-	-	-	-	-
NA10831	2 copy Chr.	G	C	T	T	G
		A	T	C	C	G
	1 copy Chr.	A	T	T	T	C
NA12156	0 copy Chr.	-	-	-	-	-
	1 copy Chr.	A	T	T	T	C

Note: 1. NA12155 is father, NA10831 is a child, and NA12156 is mother, respectively.

2. The red and blue shading represent the chromosome transmitted from father and mother, respectively.

3. The yellow shading represent to the phase haplotype of *CCL3L1/CCL4L1* after using emulsion-fusion PCR.

4. *These are not a full haplotype.

5. ***CCL4L1* sequencing results are shown in negative strand.

5.6 CEPH HapMap haplotypes at *CCL3L1 and CCL4L1*

After using several techniques to identify the *CCL3L1/CCL4L1* haplotypes in CEPH-HapMap samples, their copy number haplotypes and frequency can be deduced as in Table 5.5 and 5.6.

Overall, there were found 11 and 10 different sequences in 1- and 2-copy number *CCL3L1/CCL4L1* haplotypes, respectively. However, there were only 3 different major *CCL3L1/CCL4L1* haplotypes (as indicated in colours in the tables), and it seems that 2-copy number *CCL3L1/CCL4L1* haplotypes were a combination and duplication of 1-copy number *CCL3L1/CCL4L1* haplotypes. Moreover, evidence of recombination between *CCL3L1/CCL4L1* was also found in 2-copy number haplotypes (*i.e.* 2B6Hybrid, 2C2Hybrid, 2CM1Hybrid1 and 2CM2Hybrid2 haplotypes). These 2-copy number haplotypes consisted of *CCL3L1* haplotype type 1B and *CCL4L1* haplotype type 1A on the same chromosome as shown in Figure 5.11 and Table 5.6.

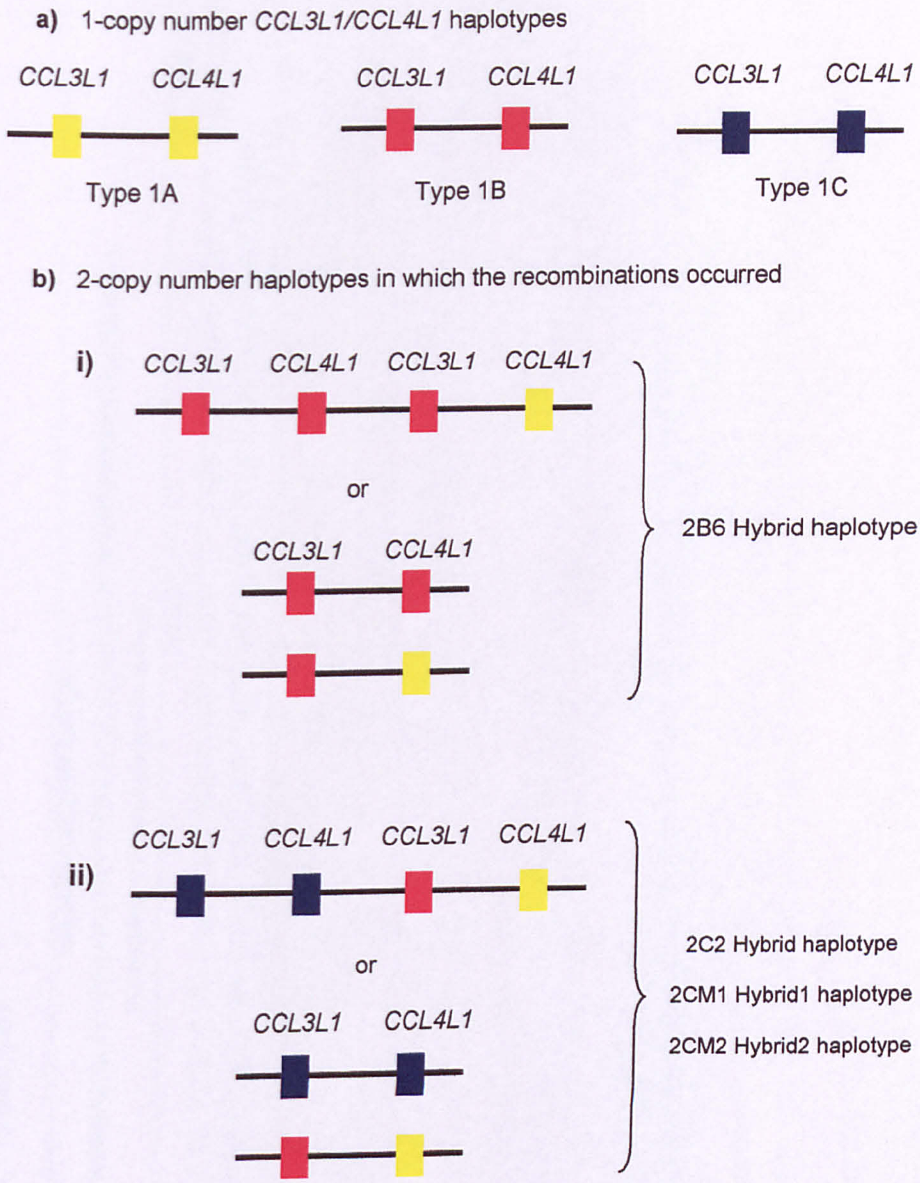


Figure 5.11: The evidence of recombination between *CCL3L1/CCL4L1* in CEPH-HapMap samples.

a) shows the 3 major haplotypes in 1-copy *CCL3L1/CCL4L1* haplotypes (type 1A, 1B and 1C). b) shows evidence of recombination which were found in 2-copy *CCL3L1/CCL4L1* haplotypes. i) 2B6 Hybrid haplotype consisted of *CCL4L1* of type 1A and *CCL3L1* of type 1B on the same chromosome. ii) The same as 2B6 Hybrid haplotype, *CCL4L1* of type 1A was also found with *CCL3L1* of type 1B on the same chromosome in 2C2Hybrid, 2CM1Hybrid1 and 2CM2Hybrid2 haplotypes.

Table 5.5: 1-copy number *CCL3L1/CCL4L1* haplotypes.

Haplotype	Sequence Variant (dbSNP129) UCSC:Mar.2006 (NCBI36/HG18); Chr.17:31647837-31649613 for *CCL3L1* and Chr.17:31663762-31665561 for *CCL4L1* [for simplicity 1 browser position is shown]																			% frequency***
	*CCL3L1**							*CCL4L1**												
		rs2	rs3	rs4	rs5	rs6	rs7	rs8	rs9	rs10	rs11	rs12	rs13	rs14		rs16	rs17	rs18	rs19		
	7837**	8282	8382	8673	8952	9366	9613	5561	5424	5423	5060	4914	4897	4806	4359**	4187	4135	3921	3792	3762**	
1A_1	C	G	A	C	T	C	A	C	C	G	A	C	G	T	A	G	G	A	G	A	1.09%
1A_2	C	G	A	C	T	G	A	C	C	G	A	C	G	T	A	G	G	A	G	A	8.70%
1A_3	C	G	A	C	T	G	A	C	C	G	A	C	G	T	C	G	G	A	G	A	3.26%
1B_1	T	A	A	C	C	G	G	T	T	C	G	C	G	T	C	G	G	A	G	A	2.17%
1B_2	T	A	A	C	T	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	2.17%
1B_3	T	A	A	T	T	C	G	T	T	C	G	C	G	T	A	G	C	A	G	T	2.17%
1B_4	T	A	A	T	T	C	G	T	T	C	G	C	G	T	A	G	G	G	G	T	6.52%
1B_5	T	A	A	T	T	C	G	C	C	G	G	C	G	T	A	G	G	A	G	A	1.09%
1B_6	T	A	A	T	T	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	67.39%
1C_1	T	G	G	T	C	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	1.09%
1C_2	T	G	G	T	C	C	G	T	T	G	A	T	G	G	C	T	G	A	T	A	4.35%

Note: 1. rs2= rs2944, rs3 = rs17850251, rs4 = rs2277661, rs5 = rs1804185, rs6 = rs1828283, rs7 = rs2073462, rs8 = rs4796195, rs9 = rs4796194, rs10 = rs4796193, rs11 = rs3744594, rs12 = rs3744595, rs13 = rs56184815, rs14 = rs3744597, rs16 = rs2277660, rs17 = rs56363716, rs18 = rs56064808, rs19 = rs1810104

2. *The locations of variant bases are provided as the last 4 digit-number under the rs number. For example, the variant base of rs2 in this table is located at Chr17: 31648282.

3. **These variant base positions are not included in dbSNP129.

4. ***Total is 92 haplotypes, and there are 32 haplotypes cannot be determined whether they are 1 copy or 2 copy number *CCL3L1/CCL4L1* haplotypes.

Table 5.6: 2-copy number *CCL3L1/CCL4L1* haplotypes.

Haplotype (show in 2 repeat units)		Sequence Variant (dbSNP129) UCSC:Mar.2006 (NCBI36/HG18); Chr.17:31647837-31649613 for *CCL3L1* and Chr.17:31663762-31665561 for *CCL4L1* [for simplicity 1 browser position is shown]																				***% frequency
		*CCL3L1**							*CCL4L1**													
			rs2	rs3	rs4	rs5	rs6	rs7	rs8	rs9	rs10	rs11	rs12	rs13	rs14		rs16	rs17	rs18	rs19		
		7837**	8282	8382	8673	8952	9366	9613	5561	5424	5423	5060	4914	4897	4806	4359**	4187	4135	3921	3792	3762**	
2B6B6	1B_6	T	A	A	T	T	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	15.79%
	1B_6	T	A	A	T	T	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	
2B6C2	1B_6	T	A	A	T	T	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	10.53%
	1C_2	T	G	G	T	C	C	G	T	T	G	A	T	G	G	C	T	G	A	T	A	
2B6 Hybrid	1B_6	T	A	A	T	T	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	10.53%
	Hybrid	T	A	A	T	T	C	G	C	C	G	A	C	G	T	A	G	G	A	G	T	
2C2 Hybrid	1C_2	T	G	G	T	C	C	G	T	T	G	A	T	G	G	C	T	G	A	T	A	26.32%
	Hybrid	T	A	A	T	T	C	G	C	C	G	A	C	G	T	A	G	G	A	G	T	
2C1M1	1C_1	T	G	G	T	C	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	10.53%
	1BM1	T	A	A	C	C	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	

Table 5.6: 2-copy number *CCL3L1/CCL4L1* haplotypes (continued).

Haplotype (show in 2 repeat units)		Sequence Variant (dbSNP129) UCSC:Mar.2006 (NCBI36/HG18); Chr.17:31647837-31649613 for *CCL3L1* and Chr.17:31663762-31665561 for *CCL4L1* [for simplicity 1 browser position is shown]																				***% frequency
		*CCL3L1**							*CCL4L1**													
			rs2	rs3	rs4	rs5	rs6	rs7	rs8	rs9	rs10	rs11	rs12	rs13	rs14		rs16	rs17	rs18	rs19		
		7837**	8282	8382	8673	8952	9366	9613	5561	5424	5423	5060	4914	4897	4806	4359**	4187	4135	3921	3792	3762**	
2BM2	1BM2	T	A	A	T	T	C	G	T	T	C	G	C	G	T	C	G	G	A	T	A	
	1BM2	T	A	A	T	T	C	G	T	T	C	G	C	G	T	C	G	G	A	T	A	5.26%
2CM1 Hybrid1	Hybrid1	T	A	A	T	T	G	G	C	C	G	A	C	G	T	A	G	G	A	G	T	
	1CM1	T	G	G	T	T	C	G	T	T	G	A	T	G	G	C	T	G	A	T	A	5.26%
2CM2 Hybrid2	Hybrid2	T	A	A	T	T	G	G	C	C	G	A	C	G	T	C	G	G	A	G	A	
	1CM2	T	G	A	T	C	G	A	T	T	C	G	C	A	T	A	G	G	A	G	T	5.26%
2A1C2	1AM1	C	A	A	T	C	C	A	T	T	G	A	T	G	T	C	G	G	A	G	A	
	1C_2	T	G	G	T	C	C	G	T	T	G	A	T	G	G	C	T	G	A	T	A	5.26%
2A1C1	1AM2	C	G	A	C	T	C	A	C	C	G	A	C	G	T	C	G	G	A	G	A	
	1CM3	T	G	G	T	C	C	G	C	C	G	A	C	G	T	C	G	G	A	G	A	5.26%

Note: 1. rs2= rs2944, rs3 = rs17850251, rs4 = rs2277661, rs5 = rs1804185, rs6 = rs1828283, rs7 = rs2073462, rs8 = rs4796195, rs9 = rs4796194, rs10 = rs4796193, rs11 = rs3744594, rs12 = rs3744595, rs13 = rs56184815, rs14 = rs3744597, rs16 = rs2277660, rs17 = rs56363716, rs18 = rs56064808, rs19 = rs1810104

2. *The locations of variant bases are provided as the last 4 digit-number under the rs number. For example, the variant base of rs2 in this table is located at Chr17: 3164<u>8282</u>.

3. **These variant base positions are not included in dbSNP129

4. ***Total is 38 haplotypes, and there are 32 haplotypes cannot be determined whether they are 1 or 2 copy number *CCL3L1/CCL4L1* haplotypes.

5.7 Tag-SNP and haplotype block analysis at *CCL3L1/CCL4L1*

To search for a SNP or a haplotype block which could show strong association with *CCL3L1/CCL4L1* haplotypes as defined by full sequencing, 1,250 SNPs flanking the *CCL3L1/CCL4L1* region (nearly 3 Mb) from the HapMap database were identified by using Haploview.

After the association between SNPs and *CCL3L1/CCL4L1* sequence haplotypes was tested, both in terms of copy number haplotypes and sub-group of copy number haplotypes divided by their similarity (*i.e.* as indicated in Table 5.5 of section 5.6), the results showed that there were no SNPs with a strong association with *CCL3L1/CCL4L1* sequence haplotypes. The results are shown in Tables 5.7 and 5.8.

Table 5.7: The best SNPs which were associated with *CCL3L1/CCL4L1* haplotypes.

Copy number haplotype	*D′*	LOD	r^2	NCBI B36 assembly (dbSNP b126)	
				SNP	Position
0	85.8	2.70	0.100	rs2287352	32679407
1	55.9	2.50	0.110	rs1266180	32521381
2	48.1	2.11	0.141	rs8067329	30864441

Note: *D′* and r^2 are the two most common measures for LD. Both of them are in the ranges between 0 and 1 which *D′* and r^2 = 1 will be perfect LD, and indicates that there is no recombination between the two loci of SNPs. LOD stands for log odds which displays to the statistical significance of LD between a pair of SNPs. The SNP which shows the highest score of D', LOD and r^2 with *CCL3L1/CCL4L1* copy number haplotype are represented in the table, but the r^2 is a more useful measure than D'.

Table 5.8: The best SNPs which were associated with *CCL3L1/CCL4L1* haplotype (analysis in terms of sub-group of copy number haplotypes).

Copy number haplotype	**D'**	**LOD**	**r^2**	**NCBI B36 assembly (dbSNP b126)**	
				SNP	**Position**
1A	64.5	1.77	0.201	rs2291189	30616734
	100	2.39	0.091	rs321597	30911907
1B	45.4	3.13	0.142	rs916841	31501593
1C	100	3.04	0.490	rs17138323	32043413
	100	3.04	0.490	rs8081787	31937223
2A	100	1.97	0.165	rs854632	31398892
	49.0	1.22	0.240	rs7218876	32096773
2B	100	1.45	0.242	rs8071957	32608804
2C	100	1.78	0.060	rs854658	31376779

Note: D' and r^2 are the two most common measures for LD. Both of them are in the ranges between 0 and 1 which D' and $r^2 = 1$ will be perfect LD, and indicates that there is no recombination between the two loci of SNPs. LOD stands for log odds which displays to the statistical significance of LD between a pair of SNPs. The SNP which shows the highest score of D', LOD and r^2 with *CCL3L1/CCL4L1* copy number haplotype are represented in the table, but the r^2 is a more useful measure than D'.

5.8 Evolution of haplotypes at *CCL3L1/CCL4L1*

In order to study evolution of human variation, comparison of human patterns to closely related species is required because the comparison will provide evolutionary information to assess uniqueness in humans (Ruvolo 1997; Tung *et al.* 2010).

Based on the phylogeny of the hominoid primates it has been suggested that chimpanzee is the closest relative organism to human, and orang-utan is the most diverged of the great apes (Sibley and Ahlquist 1984, 1987; Chen and Li 2001). Therefore, to infer an ancestral state of *CCL3L1/CCL4L1* haplotypes in human, the *CCL3L1/CCL4L1* haplotypes in human were compared to *CCL3L1/CCL4L1* sequence haplotypes in chimpanzee and orang-utan.

To generate the *CCL3L1/CCL4L1* haplotype of chimpanzee and orang-utan, the whole gene sequence of *CCL3L1/CCL4L1* in these primates was searched using two browsers, including NCBI browser (http://www.ncbi.nlm.nih.gov/) and Ensembl browser (http://www.ensembl.org/), and chimpanzee *CCL3L1/CCL4L1* sequence haplotype was also derived by using the ancestral state of SNPs (the dbSNP129) from UCSC genome browser (http://genome.ucsc.edu/). The list of sequence assembly and location are shown in the Table 5.9.

Table 5.9: Assembly and location of *CCL3L1/CCL4L1* sequences in orang-utan and chimpanzee.

Browser	Specie	Gene	Assembly	Location	Remark
NCBI	*Pongo abelii*	*CCL3L1*	P_pygmaeus_2.0.2	Chr.17_random: 22011-24122	GenBank accession number: 002963241
NCBI	*Pongo abelii*	*CCL4L1*	P_pygmaeus_2.0.2	Chr.17_random: 37419-39946	GenBank accession number: 002963241
NCBI	*Pongo abelii*	*CCL3L1*	P_pygmaeus_2.0.2	Chr.17: 736132-738247	GenBank accession number: 002888650
NCBI	*Pongo abelii*	*CCL4L1*	P_pygmaeus_2.0.2	Chr.17: 747483-749484	GenBank accession number: 002888650
NCBI	*Pongo abelii*	*CCL4L1*	P_pygmaeus_2.0.2	Chr.17: 749484-749928	GenBank accession number: 002888650
NCBI	*Pan troglodytes*	*CCL3L1*	Pan_troglodytes-2.1	Chr.17: 321836-323931	GenBank accession number: 001226927
NCBI	*Pan troglodytes*	*CCL4L1*	Pan_troglodytes-2.1	Chr.17: 337784-340325	GenBank accession number: 001226927
NCBI	*Pan troglodytes*	*CCL4L1*	Pan_troglodytes-2.1	Chr.17: 1960335-1962862	GenBank accession number: 001226927
NCBI	*Pan troglodytes*	*CCL3L1*	Pan_troglodytes-2.1	Chr.17: 1280-3396	GenBank accession number: 001227489
NCBI	*Pan troglodytes*	*CCL3L1*	Pan_troglodytes-2.1	Chr.17: 3350-5467	GenBank accession number:001227474
Ensembl	*Pongo abelii*	*CCL3L1*	PPYG2, Sep 2007	Chr.17_random 17278607-17280718	
Ensembl	*Pongo abelii*	*CCL4L1*	PPYG2, Sep 2007	Chr.17_random 17294038-17296542	
Ensembl	*Pan troglodytes*	*CCL3L1*	CHIMP2.1, Mar 2006	Chr.17: 19411281-19413223	
Ensembl	*Pan troglodytes*	*CCL4L1*	CHIMP2.1, Mar 2006	Chr.17: 19427229-19429770	

Unfortunately, the UCSC genome browser did not provide the *CCL3L1/CCL4L1* in orang-utan, and the haplotype information given for chimpanzee and orang-utan obtained from NCBI and Ensembl browsers were identified as *CCL3* and *CCL4* when the sequences were aligned and inspected for characteristic positions. Therefore, finally, *CCL3L1/CCL4L1* haplotypes in human were compared to *CCL3L1/CCL4L1* haplotypes in chimpanzee by using the data from UCSC database only, and the *CCL3L1/CCL4L1* haplotypes of chimpanzee are shown in Table 5.10.

After the *CCL3L1/CCL4L1* haplotypes in human were compared to chimpanzee, the results showed that the majority of *CCL3L1/CCL4L1* haplotypes (about 70%) which were found in human were the same as found in chimpanzee, *i.e.* the ancestral state (Table 5.10).

Regarding the human *CCL3L1/CCL4L1* haplotypes, the haplotypes might be divided by the location according to the UCSC browser of *CCL3L1/CCL4L1* repeat unit into two types. The first type was the major group including 1B and 1C haplotype groups; these haplotypes can be found at Chr17: 31647837-31649613 for *CCL3L1* and Chr.17:31663762-31665561 for *CCL4L1* (Table 5.10). The second type was 1A haplotype group which *CCL3L1* can be found at Chr17: 31546263-31548039, and *CCL4L1* at Chr.17:31562196-31563995 (Table 5.11).

Table 5.10: Comparison of *CCL3L1/CCL4L1* haplotypes between human and chimpanzee.

Haplotype	Sequence Variant (dbSNP129) UCSC:Mar.2006 (NCBI36/HG18); Chr.17:31647837-31649613 for *CCL3L1* and Chr.17:31663762-31665561 for *CCL4L1*																				% frequency***
	*CCL3L1**							*CCL4L1**													
		rs2	rs3	rs4	rs5	rs6	rs7	rs8	rs9	rs10	rs11	rs12	rs13	rs14		rs16	rs17	rs18	rs19		
	7837**	8282	8382	8673	8952	9366	9613	5561	5424	5423	5060	4914	4897	4806	4359**	4187	4135	3921	3792	3762**	
1A_1	C	G	A	C	T	C	A	C	C	G	A	C	G	T	A	G	G	A	G	A	1.09%
1A_2	C	G	A	C	T	G	A	C	C	G	A	C	G	T	A	G	G	A	G	A	8.70%
1A_3	C	G	A	C	T	G	A	C	C	G	A	C	G	T	C	G	G	A	G	A	3.26%
1B_1	T	A	A	C	C	G	G	T	T	C	G	C	G	T	C	G	G	A	G	A	2.17%
1B_2	T	A	A	C	T	C	G	T	T	C	G	C	G	T	A	G	G	A	G	A	2.17%
1B_3	T	A	A	T	T	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	2.17%
1B_4	T	A	A	T	T	C	G	T	T	C	G	C	G	T	A	G	C	A	G	T	2.17%
1B_5	T	A	A	T	T	C	G	T	T	C	G	C	G	T	A	G	G	G	G	T	6.52%
1B_6	T	A	A	T	T	C	G	C	C	G	G	C	G	T	A	G	G	A	G	A	1.09%
1C_1	T	G	G	T	C	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	67.39%
1C_2	T	G	G	T	C	C	G	T	T	C	G	C	G	T	A	G	G	A	G	T	1.09%
Chimpanzee	T	A	A	T	T	C	G	T	T	G	A	T	G	G	C	T	G	A	T	A	4.35%
								T	T	C	G	C	G	T	A	G	G	A	G	T	-

Note: 1. rs2= rs2944, rs3 = rs17850251, rs4 = rs2277661, rs5 = rs1804185, rs6 = rs1828283, rs7 = rs2073462, rs8 = rs4796195, rs9 = rs4796194, rs10 = rs4796193, rs11 = rs3744594, rs12 = rs3744595, rs13 = rs56184815, rs14 = rs3744597, rs16 = rs2277660, rs17 = rs56363716, rs18 = rs56064808, rs19 = rs1810104

2. 1A_1 to 1C_2 haplotypes are *CCL3L1/CCL4L1* haplotypes in human. The SNPs which are shown variant base identical to chimpanzee are shaded in red.

3. *The locations of variant bases are provided as the last 4 digit-number under the rs number. For example, the variant base of rs2 in this table is located at Chr17: 3164<u>8282</u>.

4. **These variant base positions are not included in dbSNP129.

5. ***Total is 92 haplotypes, and there are 32 haplotypes cannot be determined whether they are 1 copy or 2 copy number *CCL3L1/CCL4L1* haplotypes.

Table 5.11: Comparison of *CCL3L1/CCL4L1* haplotype at Chr.17: 31546263-31548039 and 31562196-31563995 between 1A haplotype group and reference haplotype from UCSC browser.

Haplotype	Sequence Variant (dbSNP129) UCSC:Mar.2006 (NCBI36/HG18); Chr.17:31546263-31548039 for *CCL3L1* and Chr.17:31562196-31563995 for *CCL4L1*																		
	*CCL3L1**							*CCL4L1**											
		rs2	rs3	rs4	rs5	rs6	rs7	rs8	rs9	rs10	rs11	rs12	rs13	rs14		rs16	rs17	rs18	rs19	
	6263**	6708	6808	7099	7378	7792	8039	3995	3858	3857	3494	3348	3331	3240	2793**	2621	2569	2355	2226	2196**
Ref_UCSC	C	G	A	C	T	G	A	C	C	G	A	C	G	T	A	G	G	A	G	A
1A_1	C	G	A	C	T	C	A	C	C	G	A	C	G	T	A	G	G	A	G	A
1A_2	C	G	A	C	T	G	A	C	C	G	A	C	G	T	A	G	G	A	G	A
1A_3	C	G	A	C	T	G	A	C	C	G	A	C	G	T	C	G	G	A	G	A

Note: 1. rs2= rs2944, rs3 = rs17850251, rs4 = rs2277661, rs5 = rs1804185, rs6 = rs1828283, rs7 = rs2073462, rs8 = rs4796195, rs9 = rs4796194, rs10 = rs4796193, rs11 = rs3744594, rs12 = rs3744595, rs13 = rs56184815, rs14 = rs3744597, rs16 = rs2277660, rs17 = rs56363716, rs18 = rs56064808, rs19 = rs1810104

2. The SNPs which are shown variant base identical to reference haplotype from UCSC (Ref_UCSC) are shaded in red.

3. *The locations of variant bases are provided as the last 4 digit-number under the rs number. For example, the variant base of rs2 in this table is located at Chr17: 31546708.

4. **These variant base positions are not included in dbSNP129.

Collectively, *CCL3L1/CCL4L1* haplotypes in human might be represented as an evolutionary tree as Figure 5.12.

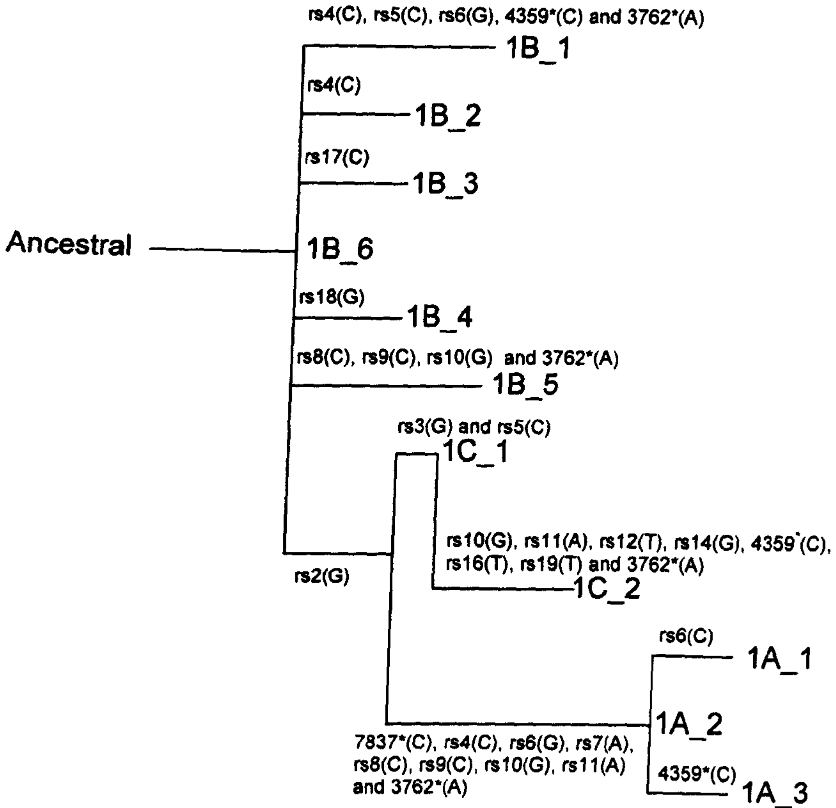


Figure 5.12: The hypothetical evolutionary tree of *CCL3L1/CCL4L1* haplotypes.

Note: rs2= rs2944, rs3 = rs17850251, rs4 = rs2277661, rs5 = rs1804185, rs6 = rs1828283, rs7 = rs2073462, rs8 = rs4796195, rs9 = rs4796194, rs10 = rs4796193, rs11 = rs3744594, rs12 = rs3744595, rs14 = rs3744597, rs16 = rs2277660, rs17 = rs56363716, rs18 = rs56064808 and rs19 = rs1810104. (all SNPs are in dbSNP129)

4359* and 3762* are variant bases which are not included in dbSNP129.

Although the association between *CCL3L1/CCL4L1* haplotype and their copy number was not found, this study has shown that the *CCL3L1/CCL4L1* cluster is a complex region, with variation thought to have occurred by recombination. There have been reports showing that the variation in copy number of *CCL3L1* might be involved in the role of macrophages, especially with respect to susceptibility to infection, such as HIV and HCV infection (Gonzalez *et al.* 2005; Grünhage *et al.* 2010). In addition, CCL5, which is a ligand that binds and activates the same CCR5 receptor as CCL3L1, has been shown to be involved and associated (its functional polymorphisms) with tuberculosis (Saukkonen *et al.* 2002; Skwor *et al.* 2006; Chu *et al.* 2007). Therefore, the association study between *CCL3L1/CCL4L1* copy number and tuberculosis should be investigated. More details are provided in the next chapter.

Chapter 6: Association study

(*CCL3L1/CCL4L1* copy number and Tuberculosis)

6.1 Introduction

Tuberculosis (TB) is one of the most important infectious diseases and is caused by *Mycobacterium tuberculosis*. Globally, in 2009 there were an estimated 9.4 million new TB cases, and 1.7 million people died of TB (about 0.4 million of these people were HIV positive) (World Health Organization 2010).

Interestingly, only about 10% of infected persons develop TB disease (Bloom and Murray 1992; Vynnycky and Fine 2000). Therefore human genetic variation might have a contribution to this difference in outcome. Consequently, human genetic risk factors for TB susceptibility have been interesting to study, to provide more understanding of aetiology of the disease which might be useful in personalized therapy and vaccination strategies (Möller *et al.* 2010).

One of the most interesting candidate genes is *CCL5*, also known as RANTES (regulated on activation, normal T cell expressed and secreted). It can inhibit intracellular growth of *M. tuberculosis* and plays an important role as a protective immune response against *M. tuberculosis* infection (Saukkonen *et al.* 2002; Skwor *et al.* 2006). Moreover, RANTES functional polymorphisms also have an association with TB susceptibility (Chu *et al.* 2007).

CCR5 is one of the receptors of *CCL5*, and also can be activated by other ligands, one of which is CCL3L1 (more details are described in 1.4.4.2).

As CCL3L1 shares the receptor with CCL5 and is the most potent ligand for CCR5 to suppress HIV entry in HIV infection, *CCL3L1* might also have a role in TB susceptibility. To test this hypothesis, therefore, an association study between *CCL3L1* copy number and TB was carried out by typing copy number and using the criteria to sieve the samples for association analysis as described in Walker (2009), with some modifications.

The investigation of association between *CCL3L1/CCL4L1* copy numbers and TB was performed in 219 TB patients and 158 controls in African population (*i.e.* !Xhosa) using a triplex PRT assay to type for *CCL3L1/CCL4L1* copy number. However, due to the quality of DNA and the limited DNA quantity, only 170 cases and 129 controls were successfully typed using 2/3 systems in majority of samples.

Based on the data from the triplex PRT assay, the numbers of cases and controls could be divided into two groups according to the agreement of the triplex PRT assay components as shown in Table 6.1.

Table 6.1: Numbers of cases and controls divided by agreement of triplex PRT assay.

Category	**Description**	**Number of cases**	**Number of controls**
A	Any sample where the integer copy number is in unanimous agreement from all 3 measurements	112	79
B	Any sample with 2 measurements in agreement and 1 failed measurement	42	29
	Any sample where 3 measures are in agreement with each other and the overall average, and the third value is within 0.75 of that integer	16	21
	Total of useful results	170	129
	Total of unusable results	49	29
	Total	219	158

6.2 Results

According to the data from Table 6.1, to reach a reliable result of the investigation of association between *CCL3L1/CCL4L1* copy number and TB, the association was examined using two data sets. In the first, association was investigated using total numbers of cases and controls typed, which were 170 and 129, respectively. In the other, the association was analysed using only those cases and controls which showed complete agreement of triplex PRT assay, totally 112 cases and 79 controls.

All results for both analyses, such as mean and standard deviation of normalised copy number in each system of triplex PRT of cases and controls, distribution of copy number value in cases and distribution of integer copy numbers were compared between cases and controls, and are shown in Tables 6.2 to Table 6.5 and Figures 6.1 to Figure 6.4.

Table 6.2: Mean and standard deviation of normalised copy number values for 170 Tuberculosis samples (all TB cases).

	CCL3C	CCL4A	LTR61A	Average
Mean	3.15	3.18	3.64	3.33
Standard deviation	1.41	1.32	1.55	1.35

Table 6.3: Mean and standard deviation of normalised copy number values for 129 control samples (all control samples).

	CCL3C	CCL4A	LTR61A	Average
Mean	3.67	3.71	4.12	3.83
Standard deviation	1.40	1.32	1.69	1.41

Table 6.4: Mean and standard deviation of normalised copy number values for 112 selected Tuberculosis samples (complete agreement of triplex PRT assay TB cases).

	CCL3C	CCL4A	LTR61A	Average
Mean	2.97	3.03	3.24	3.08
Standard deviation	1.28	1.26	1.29	1.25

Table 6.5: Mean and standard deviation of normalised copy number values for 79 control samples (complete agreement of triplex PRT assay control samples).

	CCL3C	CCL4A	LTR61A	Average
Mean	3.46	3.49	3.71	3.55
Standard deviation	1.35	1.29	1.31	1.30

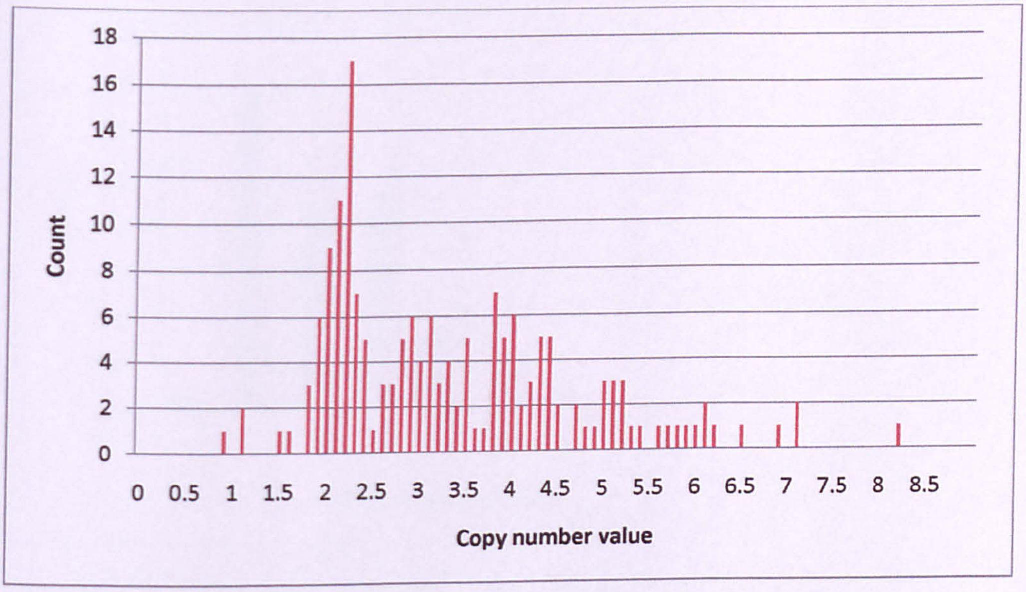


Figure 6.1: Distribution of copy number values for 170 Tuberculosis samples (all TB cases).

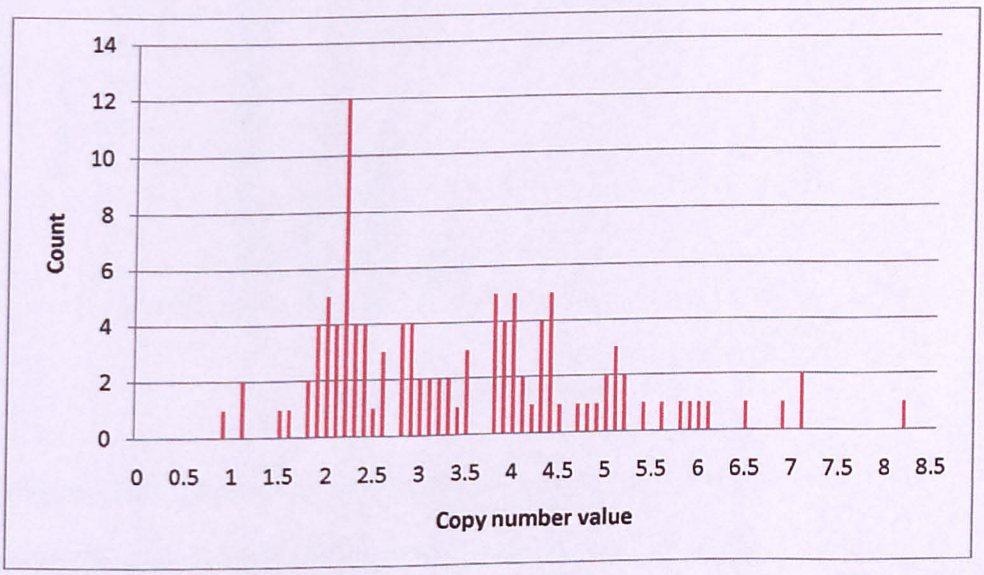


Figure 6.2: Distribution of copy number values for 112 selected Tuberculosis samples (complete agreement of triplex PRT assay TB cases).

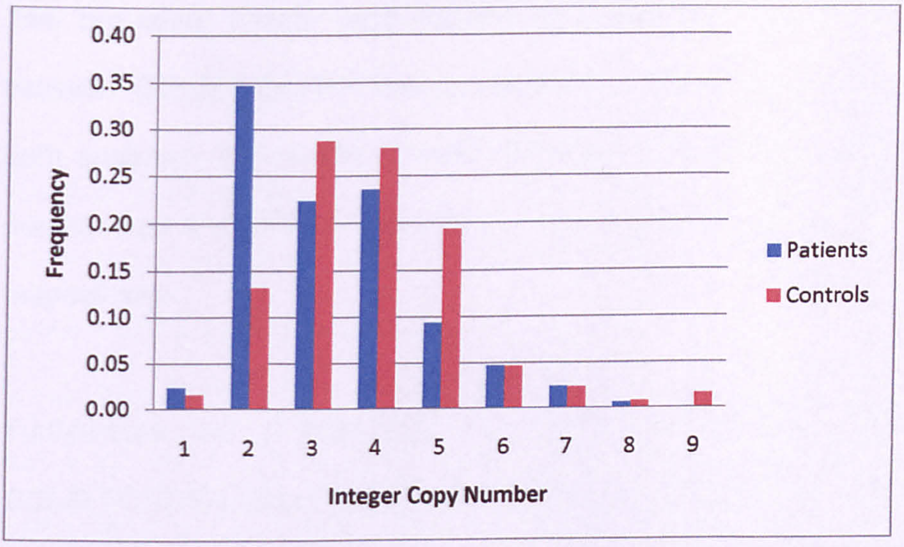


Figure 6.3: Distribution of integer copy numbers for 170 Tuberculosis patients and 129 controls (all samples).

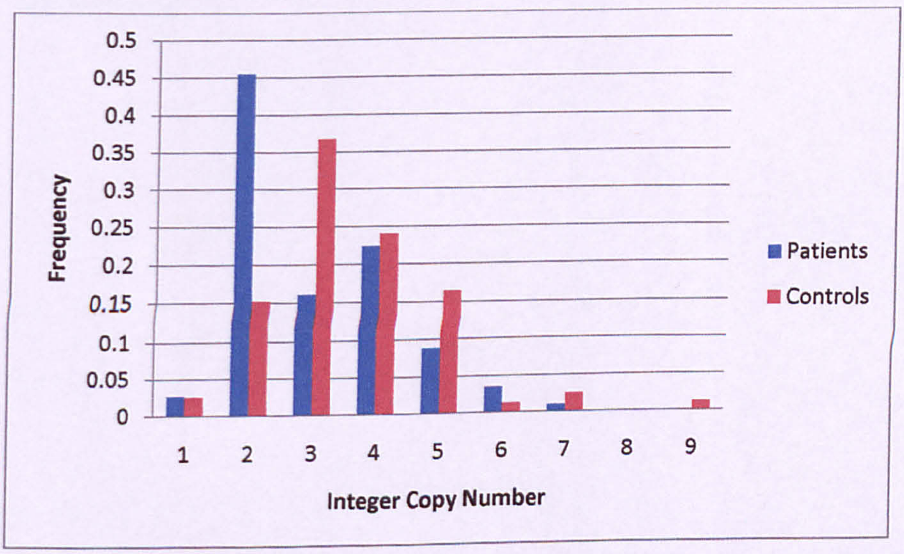


Figure 6.4: Distribution of integer copy numbers for 112 Tuberculosis patients and 79 controls (complete agreement of triplex PRT assay samples).

Two two-tailed T-tests were applied to investigate the association between *CCL3L1/CCL4L1* copy number and tuberculosis disease in both analyses. The results showed that lower *CCL3L1/CCL4L1* copy number had a significant association with TB (p = 0.001 and 0.013, respectively).

Furthermore, two χ^2 tests were used to examine for failure rate of *CCL3L1/CCL4L1* copy number typing between cases and controls, and for frequency of complete agreement in triplex PRT assay among accepted samples. Both tests showed that there are no significant difference between cases and controls (p = 0.4395 and 0.6970, respectively).

Chapter 7: Discussion

Studying the evolutionary history of variation at the *CCL3L1/ CCL4L1* cluster might have an advantage for understanding the functions of the genes and the genetic basis of susceptibility or resistance to disease in which these genes play a role. The main objective of this study was to define the relationship between copy number variation of *CCL3L1* and SNPs flanking the *CCL3L1/CCL4L1* cluster and provided information on the evolution of haplotypes at *CCL3L1/CCL4L1*.

7.1 *CCL3L1* pseudogene assay

The *CCL3L1* pseudogene or *CCL3L2* is a 5′ truncated copy of *CCL3L1*; it has sequences similar to the *CCL3L1* gene, but is lacking exon 1 of *CCL3L1*. Interestingly, the *CCL3L1* pseudogene is found in humans but not in chimpanzee (Gornalusse *et al.* 2009), implying that the *CCL3L1* pseudogene has just occurred in a single evolutionary lineage.

Similar to *CCL3L1/CCL4L1* copy number, this study demonstrated that the *CCL3L1* pseudogene varies between individuals and in populations as the results showed that it is not present in every person and its frequency is found to significantly differ between populations (p = 0.001). Extrapolating from this result, the copy number measurement of *CCL3L1* should be carried out carefully because the presence of this pseudogene might contribute to the measurement error of some assays due to its sequence similarity to *CCL3L1*. For example, the *CCL3L1* pseudogene may account for the differences in copy number between *CCL3L1* and *CCL4L1* in the study of Townson *et al.* (2002).

The association between copy number variants (CNVs) and pseudogenes has been characterised and it has been reported that there was an association between them (Kim *et al.* 2008). Although in this study *CCL3L1* pseudogene and *CCL3L1/CCL4L1* copy number and haplotype of *CCL3L1/CCL4L1* copy number did not show a significant association in most populations, the results in this study showed a trend to support Kim's study for the following reasons; the mean copy number in *CCL3L1* pseudogene containing group is higher than in the group for which the *CCL3L1* pseudogene is absent in all populations of study. Moreover, a higher frequency of presence of the *CCL3L1* pseudogene is found in the population with the highest copy number of *CCL3L1/CCL4L1* (*i.e.* the rank of decreasing of *CCL3L1* pseudogene is found in Yoruba, CHB/JPT and UK and CEPH-HapMap, respectively).

Recently, Shostakovich-Koretskaya *et al.* (2009) claimed that they found novel 5′ exons for this pseudogene using bioinformatics and mRNA profiling, which could mean it has two alternatively spliced transcripts, and is predicted to be a functional gene rather than a pseudogene. This has led to a controversial issue of how to count the

CCL3L1 copy number because the difference in method for quantifying *CCL3L1* copy number could affect an investigation of association between *CCL3L1* and diseases in which this gene might be involved (Shrestha *et al.* 2009). This is not a unique situation. Two variable pseudogenes (NCF1B and NCF1C) of neutrophil cytosolic factor-1 (*NCF1*), which is a component of NADPH oxidase, also have been found to have alternative spliced expression and the variation of these pseudogenes may be biologically relevant (Brunson *et al.* 2010). Moreover, it has been estimated that about 20% of pseudogenes could be transcribed (Zheng *et al.* 2007).

7.2 SNPs and Microsatellites

7.2.1 SNPs

This study used 15 SNPs, which lie in both the flanking region and the copy-variable region of *CCL3L1/CCL4L1*, to test for an association of a SNP with *CCL3L1/CCL4L1* copy number. Whilst there is no perfect SNP which can be used to determine *CCL3L1/CCL4L1* copy number, the results suggested a combination of SNP rs8064426 and SNP rs16972085 can be used to predict the copy number of *CCL3L1/CCL4L1* in UK and Basque populations with about 70% accuracy. Neither of these SNPs, rs8064426 and 16972085, have been reported to be associated with a disease in the NHGRI GWAS catalogue (www.genome.gov/gwastudies). However, it is interesting that SNP rs16972085 is a rare variant in European and Asian populations, but not in African populations. This is also found in SNP rs1804185, which was selected from UCSC genome browser. The SNP did not show any variation in all ECACC samples tested.

SNP rs4796195, a SNP which was analysed in this study, has been reported to have an association with the copy number of *CCL4L1* (Colobran *et al.* 2008). Surprisingly, the investigation of relationship between this SNP and the copy number of *CCL3L1/CCL4L1* in this study did not find any correlation.

Regarding the other studied SNPs, it seems that SNPs at the *CCL3L1/CCL4L1* variable region, rs3744594 and rs2277661, have a tentative relationship with *CCL3L1/CCL4L1* copy number as the results showed that these SNPs were associated with *CCL3L1/CCL4L1* copy number in most populations studied (*i.e.* ECACC, CHB/JPT and YRI for SNP rs3744594 and ECACC, CEPH HapMap and CHB/JPT for SNP rs2277661). Although these two SNPs were not useful for prediction of *CCL3L1/CCL4L1* copy number, they still can be exploited for confirmation that a sample possesses a *CCL3L1/CCL4L1* copy number of zero.

In addition, the suggested SNPs from LD analysis which were used in this study also show potential and support the concept of LD application. Most suggested SNPs, 6 out of 8, based on LD analysis showed significant association with *CCL3L1/CCL4L1* copy number except SNP rs9910447 and SNP rs12453313.

For the study of the relationship between SNPs and the *CCL3L1* pseudogene, the study showed that there were some SNPs associated with the presence of the *CCL3L1* pseudogene, but there was no one SNP that had an absolute association, even SNP rs16972085 which was suggested from LD analysis to be associated with the pseudogene. Therefore, in order to find a tag-SNP to detect the presence of the *CCL3L1* pseudogene, further investigation is still needed.

7.2.2 Microsatellites

The compound microsatellite (CCL3CT) which is composed of CT-rich microsatellite was applied in a similar manner to SNPs to investigate association with *CCL3L1/CCL4L1* copy number and the *CCL3L1* pseudogene in ECACC samples. This microsatellite is located within the *CCL3* constant region (Chr.17:31447038-31447215) which is approximately 99kb from the *CCL3L1/CCL4L1* variable region. It was investigated to see whether *CCL3* varied in copy number (Walker 2009). In summary, Walker (2009) analysed 87 Yoruba samples and found that most samples produced two products (including 2 samples which produced unequal intensity of peaks). The exception was 8 samples which produced a single product and 4 samples produced three distinct products. Based on this microsatellite analysis, Walker (2009) suspected that *CCL3* in African population could vary in copy number not just *CCL3L1/CCL4L1*. In addition, Walker (2009) suspected that a child in the study might have a somatic mutation.

Similar to the study of Walker (2009), the study in ECACC samples also found that most samples (180 out of 192 samples) produced two products with equal height peaks, and a homozygous sample for the same allele length can be found as well (8 out of 192 samples). Interestingly, there were 4 samples 2 of which produced three distinct products of different size and 2 samples produced four different sized products. It was confirmed that 2 samples were mixed samples by using other microsatellites (*i.e.* D18S51 and D21S11, which usually are used for individual identification (Kimpton *et al.* 1996)), and 2 samples remained that have been proposed to have a somatic mutation at CCL3CT. Based on the prevalence of samples with somatic mutation in this study and on the study of Walker (2009), it may be suggested that the *CCL3* microsatellite is a sensitive region for somatic mutation, but further investigation is required. However, it could be summarized that there is no variation of *CCL3* copy number in European population.

CCL3CT microsatellite was not found to have a relationship with the copy number of *CCL3L1/CCL4L1*, and the association with *CCL3L1* pseudogene was weak (p = 0.0277). This might suggest that the CCL3CT microsatellite has a high rate of mutation, the *CCL3L1/CCL4L1* CNV has a high rate of mutation, and/or there has been frequent recombination at the *CCL3L1* and *CCL4L1* region.

Lastly, although the study showed that this microsatellite does not have a relationship with *CCL3L1/CCL4L1* copy number, again this microsatellite might be used as a marker for detection and/or confirmation of zero copy number of *CCL3L1/CCL4L1*. However, more investigation is still needed to confirm this benefit.

7.3 Tag-SNPs and Haplotype block analysis

Haplotype-based association studies were continued after SNP and microsatellite association studies were not successful to find a tag-SNP or a haplotype block to use for determination of copy number at *CCL3L1/CCL4L1*. LD was also investigated in this study by analysing 1,250 HapMap SNPs flanking the *CCL3L1/CCL4L1* region (about 3 Mb). However, for haplotype-based association studies, it is not easy to obtain valid haplotype phase information and errors of inferred phase information could be frequent (Clark 2004). In this study, several methods such as sequencing, segregation analysis, allele-specific PCR and emulsion haplotype fusion PCR were applied to identify *CCL3L1/CCL4L1* sequence haplotypes, but the problem of phase uncertainty was also still found; there were some haplotypes that cannot be defined for their *CCL3L1/CCL4L1* copy number (*i.e.* 32 out of 124 haplotypes).

Surprisingly, although haplotype-based association studies were used to identify a SNP which had an association with *CCL3L1/CCL4L1* copy number, the results showed that there were no SNPs with a strong association with *CCL3L1/CCL4L1* haplotypes or even with a sub-group of copy number haplotypes (see Tables 5.7-5.8). For example, for "1C" *CCL3L1/CCL4L1* copy number haplotype group, which was the lowest frequency haplotype group (about 5% of observed 1-copy number haplotypes; see Table 5.5), an absolute LD tag-SNP was still not found in this group. In order to identify this "1C" *CCL3L1/CCL4L1* copy number haplotype, a combination of two tag-SNPs was needed.

Moreover, a combination of SNPs (rs8064426 and rs16972085) which was used to predict the *CCL3L1/CCL4L1* copy number in the UK and Basque populations is not suggested as a candidate tag-SNP in this study. In fact, in this study the SNP rs8064426 was the best SNP associated with 1 copy number of *CCL3L1/CCL4L1* (D' = 45.1, LOD = 2.82 and r^2 = 0.126), but this association was reduced in testing of association with haplotypes (1-copy number haplotype; D' = 46.4, LOD = 1.66 and r^2 = 0.08) and sub-group haplotypes (type 1A haplotype: D' = 9, LOD = 0.00 and r^2 = 0.000; type 1B haplotype: D' = 48.1, LOD = 0.89 and r^2 = 0.039, and type 1C haplotype: D' = 100, LOD = 0.18 and r^2 = 0.005).

Usually haplotype-based association studies are assumed to provide greater power than using single genetic markers, but in this study it seemed that haplotyped-based approaches were not more advantageous than using single genetic makers. There were some suggestions that this might be because of uncertainty in the haplotype inference, or, as well as increasing the degrees of freedom in the haplotype-based association study (Cordell and Clayton 2005). An example of this evidence has also been found in the study of Pinnaduwage and Briollais (2005) which showed that haplotype-based analysis had no advantage over genotype-based (single locus) analysis in the association study of alcohol dependence.

Taken together, and based on several approaches, including SNP, microsatellite and haplotype, this work did not show a strong association between single-copy marker and *CCL3L1/CCL4L1* copy number; this may suggest that *CCL3L1/CCL4L1* is a complex region and one plausible hypothesis is that there is a high rate of recombination in this region (more discussion in section 7.4).

7.4 Evolution of haplotypes at *CCL3L1/CCL4L1*

To generate *CCL3L1/CCL4L1* haplotypes, three different major single-copy *CCL3L1/CCL4L1* haplotypes were found. There were classified as 1A, 1B and 1C types according to their sequence similarity (see Table 5.5). Comparison of these sequence haplotypes with reference sequences on UCSC browser showed that, type 1A and type 1B haplotypes can be identified as type 1A located at chr17:31,546,382-31,548,269 and chr17:31,562,581-31,564,393 for *CCL3L1* and *CCL4L1*, respectively, and type 1B annotated as *CCL3L3/CCL4L2* located at chr17:31,647,956-31,649,843 and chr17:31,664,147-31,665,959, respectively. However, both "*CCL3L3*" and "*CCL4L2*" are still designated in this study as *CCL3L1* and *CCL4L1*, respectively because in each pair, *CCL3L* and *CCL4L* had identical sequences in coding regions (Colobran *et al.* 2010; Shrestha *et al.* 2010).

Although *CCL4L1* shared 100% nucleotide identity with *CCL4L2* in the coding region, a variant intron 2 acceptor splice site was found in *CCL4L1* and creates a variety of mRNA isoforms (Colobran *et al.* 2005). Thus, the variation in the copy number of *CCL4L* might have functional effects. To date, the functional difference between *CCL4L1* and *CCL4L2* is unknown, but it was suggested that the new splicing in *CCL4L1* might reduce its activity (Colobran *et al.* 2010). In fact, Colobran *et al.* (2005) found that in individuals with only *CCL4L1*, it was expressed at a lower level than if only *CCL4L2* is present (12% vs 52% compared to *CCL4* transcripts); they also found that the *CCL4L1* allele was at higher frequency in HIV infected individuals than controls. However, further investigation is still needed to identify the importance of variation in the forms of *CCL4L*.

Interestingly, it seemed that the type 1C haplotype, which had the lowest observed frequency (5%), was mutated from the type 1B haplotype. Two missense mutations were found in this group. One was found at *CCL3L1* where SNP rs17850251 changed leucine to proline, and the other one was found in *CCL4L1* in which SNP rs3744595 altered arginine to histidine. Although the effects of these changes are still unknown, it is possible that the substitution of these amino acids might influence the folding of protein, resulting in an altered functional effect of *CCL3L1/CCL4L1* (Paximadis *et al.* 2009).

Moreover, 2-copy *CCL3L1/CCL4L1* haplotypes were found to be duplicated from 1-copy *CCL3L1/CCL4L1* haplotypes, supporting the study of Modi (2004) that found internal duplication in these regions. Evidence of recombination between *CCL3L1* and *CCL4L1* was also found in 2-copy haplotypes. These recombinations might explain the difficulty in finding a tag-SNP or a haplotype block in this region.

Similarly, Paximadis *et al.* (2009) characterized *CCL3* and *CCL3L1* genes in 43 South African African (SAA) mother-infant pairs from a mother-to-infant HIV-1 transmission study, and in 28 South African Caucasian (SAC) adult volunteers. They also found only 3 classes of

CCL3L1 haplotypes in their study, which were the same as shown in this study. However, it should be noted that there were two variant bases at the 3′ end of *CCL3L1* in the type 1A haplotype found only in the study of Paximadis *et al.* (2009) and not in this study, and a few bases at 3′ and 5′ end of *CCL3L1* overlapped between both studies. In summary, they found type 1B in SAA population and type 1C and type 1A in SAC population. Surprisingly, type 1B, which was most common in the European population in this study, was found only in the SAA population and did not appear in the SAC population in Paximadis *et al.* (2009). In addition, they also reported that type 1C was found at higher frequency than type 1A in the SAC population, while in this study the frequency of type 1A was found at twice the frequency compared to type 1C.

To investigate the evolution of *CCL3L1/CCL4L1* haplotypes, the human haplotypes of *CCL3L1/CCL4L1* were compared to *CCL3L1/CCL4L1* haplotypes in chimpanzee. Interestingly, the *CCL3L1/CCL4L1* haplotype in chimpanzee was found to be identical to type 1B (*i.e.* 1B_6 haplotype) which was found at a frequency of nearly 70% in this study. Taken together with the result of Paximadis *et al.* (2009), which reported type 1B haplotype as the only haplotype found in the African population, it could be surmised that type 1B was the ancestral state haplotype of *CCL3L1/CCL4L1*. Also type 1A represents an internal duplication of type 1B, and type 1C was a (mutated) derivative of type 1B. Moreover, it could be implied that the type 1B haplotype had a significant functional role, in that it was kept through the evolution of *CCL3L1/CCL4L1* haplotypes without major changes.

It is also important to note that a variant at the 3′ end of *CCL3* can be found both in sequences of chimpanzee and orang-utan, in which they were similar to *CCL3L1*. However, when these sequences were aligned to *CCL3L1*, they showed a lot of base substitutions similar to *CCL3*. For example in the study of Paximadis *et al.* (2009), they also compared *CCL3L1* sequences from their population study with the sequence of chimpanzee from NCBI browser (GenBank accession number :NW_001226927). This sequence should not be annotated *CCL3L1* because there were a lot of base substitutions characteristic of *CCL3*, although this sequence of chimpanzee had a deletion (the same as *CCL3L1*) at the 3′end.

7.5 Association study

TB is an important infectious disease. Many studies have suggested that genetic factors play a role in TB susceptibility (reviewed in Möller, *et al.* 2009). *CCL5* is one of the candidate genes shown to have an association with TB susceptibility (Chu *et al.* 2007). CCL3L1 is one of chemokines that shares its receptor with CCL5 (*i.e.* CCR5), and it has associations with HIV infection (Gonzalez *et al.* 2005; Nakajima *et al.* 2007), Kawasaki disease (Burns *et al.* 2005; Mamtani *et al.* 2010) and chronic hepatitis C (Grünhage *et al.* 2010). As a result, *CCL3L1* should be also an interesting candidate gene. Surprisingly, until now, there is no report on a study of the association between *CCL3L1* and TB.

Thus, this might be the first study to show lower *CCL3L1/CCL4L1* copy number has a significant association with TB (p = 0.013). Moreover, this study is one of many to show that *CCL3L1* is a genetic risk factor for infection susceptibility. It is not the only association study that shows possession of *CCL3L1* copy number lower than their population average to be associated with TB. The other studies also show that the possession of *CCL3L1* copy number lower than their population average is correlated with susceptibility to HIV (Gonzalez *et al.* 2005; Nakajima *et al.* 2007) and HCV infection (Grünhage *et al.* 2010).

However, the numbers of cases and controls that were used in this study is still small. Ideally this association should be investigated and replicated in a large number of cases and population matched controls to confirm the significant association between *CCL3L1/CCL4L1* copy number and TB.

In addition to *CCL3L1/CCL4L1*, their non-copy number variable paralogues *CCL3* and *CCL4,* which bind to receptor CCR5, have also been tested for association with TB. Although, studies showed that variation at *CCL3* is not associated with TB (Flores-Villanueva *et al.* 2005), variation at *CCL4* is (p = 0.002) (Jamieson *et al.* 2004). Collectively, several ligands of CCR5 show an association with TB. Therefore, the study of functional polymorphisms of *CCR5* should also be investigated further for its association with TB.

Although this study provides a good preliminary result, it still has a point that could be improved to make it more accurate and reliable in a further study; that is the measurement accuracy for *CCL3L1/CCL4L1* copy number.

The triplex PRT assay which was applied to measure *CCL3L1/CCL4L1* copy number in this study still remains a problem in agreement of integer copy number between systems in some case and control samples (Table 6.1). It seems that the triplex PRT can accurately measure low copy number of *CCL3L1/CCL4L1* but still struggles to measure a high copy number of *CCL3L1/CCL4L1*. This problem was also found in the study of measurement of *CCL3L1/CCL4L1* copy number in Pygmy samples (with relatively high copy number, 3-10 copies (Walker 2009)), although the measurement of *CCL3L1/CCL4L1* in UK population (with relatively low copy number, 0-4 copies) provides a highly robust result (Walker 2009). Therefore, the development of a triplex PRT assay for typing *CCL3L1/CCL4L1* in high copy number samples such as more than 3 or 4 copies is required. Moreover, the issue of the quality and amount of DNA samples also should not be avoided because it could affect the estimation of copy number. Urban *et al.* (2009) have reported that difference of DNA amounts can cause the differential bias in copy number using the real-time PCR method, and DNA shearing can lead to systematic copy number overestimation using the PRT method.

In summary, to understand the evolutionary history of variation at the *CCL3L1/CCL4L1* region, CEPH-HapMap samples have been characterised for the *CCL3L1/CCL4L1* haplotypes and compared to the chimpanzee *CCL3L1/CCL4L1* sequence. This study shows that there are 3 major haplotypes of *CCL3L1/CCL4L1* (type 1A, 1B and 1C) in the European population (*i.e.* CEPH-HapMap samples). The major 1-copy *CCL3L1/CCL4L1* haplotype (type 1B, found at about 70% frequency) is the ancestral state, as inferred by comparison with chimpanzee. The 2-copy *CCL3L1/CCL4L1* haplotypes are duplicated from 1-copy *CCL3L1/CCL4L1* haplotypes. Interestingly, evidence of recombination is also found in 2-copy *CCL3L1/CCL4L1* haplotypes. As a result, this may suggest that *CCL3L1/CCL4L1* is a complex region and one plausible hypothesis is that there is a high rate of recombination in this region. With this assumption and the evidence of recombination which was identified in this study, this might be an explanation why there is no strong association between single-copy markers (SNPs and microsatellites), *CCL3L1/CCL4L1* haplotypes and *CCL3L1/CCL4L1* copy number. The study also shows the variation of the *CCL3L1* pseudogene in individuals and in populations (UK, CEPH-HapMap, CHB/JPT and YRI). Although this *CCL3L1* pseudogene is not significantly associated with *CCL3L1/CCL4L1* copy number, there is a positive correlation between its presence and copy number. Work is still needed on the function and possible expression of this pseudogene. This is because the *CCL3L1* pseudogene can affect counting the *CCL3L1* copy number and subsequently affect interpretation of association studies of *CCL3L1* copy number and disease (Shostakovich-Koretskaya *et al.* 2009). Moreover, a robust standard technique to determine copy number variation particularly in the high copy number range is still urgently needed, in particular, for *CCL3L1/CCL4L1* copy number.

References

Alkan, Can, Bradley P. Coe and Evan E. Eichler (2011). "Genome structural variation discovery and genotyping." Nat Rev Genet **advance online publication**.

Allen, Samantha J., Susan E. Crown and Tracy M. Handel (2007). "Chemokine:Receptor Structure, Interactions, and Antagonism." Annual Review of Immunology **25**(1): 787-820.

Alonso, S. and J. A. L. Armour (1998). "MS205 minisatellite diversity in Basques: Evidence for a pre-neolithic component." Genome Research **8**(12): 1289-1298.

Ardlie, Kristin G., Leonid Kruglyak and Mark Seielstad (2002). "Patterns of linkage disequilibrium in the human genome." Nat Rev Genet **3**(4): 299-309.

Arenzana-Seisdedos, Fernando and Marc Parmentier (2006). "Genetics of resistance to HIV infection: Role of co-receptors and co-receptor ligands." Seminars in Immunology **18**(6): 387-403.

Armour, John A. L., Raquel Palla, Patrick L. J. M. Zeeuwen, Martin den Heijer, Joost Schalkwijk and Edward J. Hollox (2007). "Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats." Nucl. Acids Res. **35**(3): e19.

Bailer, Ursula, Friedrich Leisch, Kurt Meszaros, Elisabeth Lenzinger, Ulrike Willinger, Rainer Strobl, Angela Heiden, Christian Gebhardt, Elisabeth Döge, Karoline Fuchs, Werner Sieghart, Siegfried Kasper, Kurt Hornik and Harald N. Aschauer (2002). "Genome scan for susceptibility loci for schizophrenia and bipolar disorder." Biological Psychiatry **52**(1): 40-52.

Bailey, Justin R., Karen O'Connell, Hung-Chih Yang, Yefei Han, Jie Xu, Benjamin Jilek, Thomas M. Williams, Stuart C. Ray, Robert F. Siliciano and Joel N. Blankson (2008). "Transmission of Human Immunodeficiency Virus Type 1 from a Patient Who Developed AIDS to an Elite Suppressor." J. Virol. **82**(15): 7395-7410.

Barrett, J. C., B. Fry, J. Maller and M. J. Daly (2005). "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics **21**(2): 263-265.

Beckmann, Jacques S., Xavier Estivill and Stylianos E. Antonarakis (2007). "Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability." Nat Rev Genet **8**(8): 639-646.

Berger, Edward A., Philip M. Murphy and Joshua M. Farber (1999). "CHEMOKINE RECEPTORS AS HIV-1 CORECEPTORS: Roles in Viral Entry, Tropism, and Disease." Annual Review of Immunology **17**(1): 657-700.

Bhattacharya, Tanmoy, Jennifer Stanton, Eun-Young Kim, Kevin J. Kunstman, John P. Phair, Lisa P. Jacobson and Steven M. Wolinsky (2009). "CCL3L1 and HIV/AIDS susceptibility." Nat Med **15**(10): 1112-1115.

Bloom, Barry R. and Christopher J. L. Murray (1992). "Tuberculosis: Commentary on a Reemergent Killer." Science **257**(5073): 1055-1064.

Brookes, Anthony J. (1999). "The essence of SNPs." Gene **234**(2): 177-186.

Brunson, Tiffany, Qingwei Wang, Isfahan Chambers and Qing Song (2010). "A copy number variation in human NCF1 and its pseudogenes." BMC Genetics **11**(1): 13.

Burns, Jane C., Chisato Shimizu, Enrique Gonzalez, Hemant Kulkarni, Sukeshi Patel, Hiroko Shike, Robert S. Sundel, Jane W. Newburger and Sunil K. Ahuja (2005). "Genetic Variations in the Receptor-Ligand Pair CCR5 and CCL3L1 Are Important Determinants of Susceptibility to Kawasaki Disease." J Infect Dis **192**(2): 344-349.

Carrington, Mary, Michael Dean, Maureen P. Martin and StephenJ. O'Brien (1999). "Genetics of HIV-1infection: chemokine receptor CCR5 polymorphism and its consequences." Hum. Mol. Genet. **8**(10): 1939-1945.

Chen, Feng-Chi and Wen-Hsiung Li (2001). "Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees." American journal of human genetics **68**(2): 444-456.

Chu, S. F., C. M. Tam, H. S. Wong, K. M. Kam, Y. L. Lau and A. K. S. Chiang (2007). "Association between RANTES functional polymorphisms and tuberculosis in Hong Kong Chinese." Genes Immun **8**(6): 475-479.

Clark, Andrew G. (2004). "The role of haplotypes in candidate gene studies." Genetic Epidemiology **27**(4): 321-333.

Collins, Francis S., Lisa D. Brooks and Aravinda Chakravarti (1998). "A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation." Genome Research **8**(12): 1229-1231.

Colobran, R, P Adreani, Y Ashhab, A Llano, JA Este, O Dominguez, R Pujol-Borrell and M Juan (2005). "Multiple products derived from two CCL4 loci: high incidence of a new polymorphism in HIV+ patients." J Immunol **174**(9): 5655 - 5664.

Colobran, R., N. Casamitjana, A. Roman, R. Faner, E. Pedrosa, J. I. Arostegui, R. Pujol-Borrell, M. Juan and E. Palou (2009). "Copy number variation in the CCL4L gene is associated with susceptibility to acute rejection in lung transplantation." Genes Immun **10**(3): 254-259.

Colobran, R., D. Comas, R. Faner, E. Pedrosa, R. Anglada, R. Pujol-Borrell, J. Bertranpetit and M. Juan (2008). "Population structure in copy number variation and SNPs in the CCL4L chemokine gene." Genes Immun **9**(4): 279-288.

Colobran, R., E. Pedrosa, L. Carretero-Iglesia and M. Juan (2010). "Copy number variation in chemokine superfamily: the complex scene of CCL3L–CCL4L genes in health and disease." Clinical & Experimental Immunology **162**(1): 41-52.

Cordell, Heather J. and David G. Clayton (2005). "Genetic association studies." The Lancet **366**(9491): 1121-1131.

Cummings, C. J. and H. Y. Zoghbi (2000). "TRINUCLEOTIDE REPEATS: Mechanisms and Pathophysiology." Annual Review of Genomics and Human Genetics **1**(1): 281-328.

Daly, Mark J., John D. Rioux, Stephen F. Schaffner, Thomas J. Hudson and Eric S. Lander (2001). "High-resolution haplotype structure in the human genome." Nat Genet **29**(2): 229-232.

de Bakker, Paul I. W., Roman Yelensky, Itsik Pe'er, Stacey B. Gabriel, Mark J. Daly and David Altshuler (2005). "Efficiency and power in genetic association studies." Nat Genet **37**(11): 1217-1223.

Deeks, Steven G. and Bruce D. Walker (2007). "Human Immunodeficiency Virus Controllers: Mechanisms of Durable Virus Control in the Absence of Antiretroviral Therapy." Immunity **27**(3): 406-416.

Degenhardt, JD, P de Candia, A Chabot, S Schwartz, L Henderson, B Ling, M Hunter, Z Jiang, RE Palermo, M Katze, EE Eichler, M Ventura, J Rogers, P Marx, Y Gilad and CD Bustamante (2009). "Copy number variation of CCL3-like genes affects rate of progression to simian-AIDS in Rhesus Macaques (Macaca mulatta)." PLoS Genet **5**(1): e1000346.

Dolan, Matthew J., Hemant Kulkarni, Jose F. Camargo, Weijing He, Alison Smith, Juan-Manuel Anaya, Toshiyuki Miura, Frederick M. Hecht, Manju Mamtani, Florencia Pereyra, Vincent Marconi, Andrea Mangano, Luisa Sen, Rosa Bologna, Robert A. Clark, Stephanie A. Anderson, Judith Delmar, Robert J. O'Connell, Andrew Lloyd, Jeffrey Martin, Seema S. Ahuja, Brian K. Agan, Bruce D. Walker, Steven G. Deeks and Sunil K. Ahuja (2007). "CCL3L1 and CCR5 influence cell-mediated immunity and affect HIV-AIDS pathogenesis via viral entry-independent mechanisms." Nat Immunol **8**(12): 1324-1336.

Ellegren, Hans (2004). "Microsatellites: simple sequences with complex evolution." Nat Rev Genet **5**(6): 435-445.

Estivill, Xavier and Lluís Armengol (2007). "Copy Number Variants and Common Disorders: Filling the Gaps and Exploring Complexity in Genome-Wide Association Studies." PLoS Genet **3**(10): e190.

Fallin, Daniele, Annick Cohen, Laurent Essioux, Ilya Chumakov, Marta Blumenfeld, Daniel Cohen and Nicholas J. Schork (2001). "Genetic Analysis of Case/Control Data Using Estimated Haplotype Frequencies: Application to APOE Locus Variation and Alzheimer's Disease." Genome Research **11**(1): 143-151.

Fellay, Jacques, Dongliang Ge, Kevin V. Shianna, Sara Colombo, Bruno Ledergerber, Elizabeth T. Cirulli, Thomas J. Urban, Kunlin Zhang, Curtis E. Gumbs, Jason P. Smith, Antonella Castagna, Alessandro Cozzi-Lepri, Andrea De Luca, Philippa Easterbrook, Huldrych F. Günthard, Simon Mallal, Cristina Mussini, Judith Dalmau, Javier Martinez-Picado, José M. Miro, Niels Obel, Steven M. Wolinsky, Jeremy J. Martinson, Roger Detels, Joseph B. Margolick, Lisa P. Jacobson, Patrick Descombes, Stylianos E. Antonarakis, Jacques S. Beckmann, Stephen J. O'Brien, Norman L. Letvin, Andrew J. McMichael, Barton F. Haynes, Mary Carrington, Sheng Feng, Amalio Telenti, David B. Goldstein and Niaid Center for HIV/AIDS Vaccine Immunology (2009). "Common Genetic Variation and the Control of HIV-1 in Humans." PLoS Genet **5**(12): e1000791.

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

Feuk, Lars, Christian R. Marshall, Richard F. Wintle and Stephen W. Scherer (2006). "Structural variants: changing the landscape of chromosomes and design of disease studies." Human Molecular Genetics **15**(suppl 1): R57-R66.

Field, Sarah F., Joanna M. M. Howson, Lisa M. Maier, Susan Walker, Neil M. Walker, Deborah J. Smyth, John A. L. Armour, David G. Clayton and John A. Todd (2009). "Experimental aspects of copy number variant assays at CCL3L1." Nat Med **15**(10): 1115-1117.

Flores-Villanueva, Pedro O., Jorge A. Ruiz-Morales, Chang-Hwa Song, Ludmila M. Flores, Eun-Kyeong Jo, Marta Montaño, Peter F. Barnes, Moises Selman and Julio Granados (2005). "A functional promoter polymorphism in monocyte chemoattractant protein-1 is associated with increased susceptibility to pulmonary tuberculosis." The Journal of Experimental Medicine **202**(12): 1649-1658.

Gabriel, Stacey B., Stephen F. Schaffner, Huy Nguyen, Jamie M. Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, Shau Neen Liu-Cordero, Charles Rotimi, Adebowale Adeyemo, Richard Cooper, Ryk Ward, Eric S. Lander, Mark J. Daly and David Altshuler (2002). "The Structure of Haplotype Blocks in the Human Genome." Science **296**(5576): 2225-2229.

Ginzinger, David G., Tony E. Godfrey, Janice Nigro, Dan H. Moore, Seiji Suzuki, Maria G. Pallavicini, Joe W. Gray and Ronald H. Jensen (2000). "Measurement of DNA Copy Number at Microsatellite Loci Using Quantitative PCR Analysis." Cancer Research **60**(19): 5405-5409.

Goldstein, David B. and Christian Schlötterer, Eds. (1999). Microsatellites: *Evolution and Applications*. New York, Oxford University Press.

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

Gornalusse, German, Srinivas Mummidi, Weijing He, Guido Silvestri, Mike Bamshad and Sunil K. Ahuja (2009). "CCL3L Copy Number Variation and the Co-Evolution of Primate and Viral Genomes." PLoS Genet **5**(1): e1000359.

Goulder, Philip J. R., Rodney E. Phillips, Robert A. Colbert, Stephen McAdam, Graham Ogg, Martin A. Nowak, Paul Giangrande, Graz Luzzi, Barbara Morgana, Anne Edwards, Andrew J. McMichael and Sarah Rowland-Jones (1997). "Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS." Nat Med **3**(2): 212-217.

Grünhage, Frank, Jacob Nattermann, Olav A. Gressner, Hermann E. Wasmuth, Claus Hellerbrand, Tilman Sauerbruch, Ulrich Spengler and Frank Lammert (2010). "Lower copy numbers of the chemokine CCL3L1 gene in patients with chronic hepatitis C." Journal of Hepatology **52**(2): 153-159.

Harrison, Paul M., Hedi Hegyi, Suganthi Balasubramanian, Nicholas M. Luscombe, Paul Bertone, Nathaniel Echols, Ted Johnson and Mark Gerstein (2002). "Molecular Fossils in the Human Genome: Identification and Analysis of the Pseudogenes in Chromosomes 21 and 22." Genome Research **12**(2): 272-280.

Hästbacka, Johanna, Albert de la Chapelle, Ilkka Kaitila, Pertti Sistonen, Alix Weaver and Eric Lander (1992). "Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland." Nat Genet **2**(3): 204-211.

He, W, H Kulkarni, J Castiblanco, C Shimizu, U Aluyen, R Maldonado, A Carrillo, M Griffin, A Lipsitt, L Beachy, L Shostakovich-Koretskaya, A Mangano, L Sen, RJ Nibbs, CT Tiemessen, H Bolivar, MJ Bamshad, RA Clark, JC Burns, MJ Dolan and SK Ahuja (2009). "Reply to: "Experimental aspects of copy number variant assays at CCL3L1"." Nat Med **15**(10): 1117 - 1120.

Hirashima, M., T. Ono, M. Nakao, H. Nishi, A. Kimura, H. Nomiyama, F. Hamada, M. C. Yoshida and K. Shimada (1992). "Nucleotide sequence of the third cytokine LD78 gene and mapping of all three LD78 gene loci to human chromosome 17." Mitochondrial DNA **3**(4): 203-212.

Hirotsune, Shinji, Noriyuki Yoshida, Amy Chen, Lisa Garrett, Fumihiro Sugiyama, Satoru Takahashi, Ken-ichi Yagami, Anthony Wynshaw-Boris and Atsushi Yoshiki (2003). "An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene." Nature **423**(6935): 91-96.

http://projects.tcag.ca/variation/. Retrieved 15/04/11.

http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi. Retrieved 04/01/11.

Hutter, Gero, Daniel Nowak, Maximilian Mossner, Susanne Ganepola, Arne Mussig, Kristina Allers, Thomas Schneider, Jorg Hofmann, Claudia Kucherer, Olga Blau, Igor W. Blau, Wolf K. Hofmann and Eckhard Thiel (2009). "Long-Term Control of HIV by CCR5 Delta32/Delta32 Stem-Cell Transplantation." N Engl J Med **360**(7): 692-698.

Iafrate, A. J., L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer and C. Lee (2004). "Detection of large-scale variation in the human genome." Nat Genet **36**(9): 949-51.

International Human Genome Sequencing Consortium (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Jamieson, S. E., E. N. Miller, G. F. Black, C. S. Peacock, H. J. Cordell, J. M. M. Howson, M. A. Shaw, D. Burgner, W. Xu, Z. Lins-Lainson, J. J. Shaw, F. Ramos, F. Silveira and J. M. Blackwell (2004). "Evidence for a cluster of genes on chromosome 17q11-q21 controlling susceptibility to tuberculosis and leprosy in Brazilians." Genes Immun **5**(1): 46-57.

Javanbakht, Hassan, Ping An, Bert Gold, Desiree C. Petersen, Colm O'Huigin, George W. Nelson, Stephen J. O'Brien, Gregory D. Kirk, Roger Detels, Susan Buchbinder, Sharyne Donfield, Sergey Shulenin, Byeongwoon Song, Michel J. Perron, Matthew Stremlau, Joseph Sodroski, Michael Dean and Cheryl Winkler (2006). "Effects of human *TRIM5α* polymorphisms on antiretroviral function and susceptibility to human immunodeficiency virus infection." Virology **354**(1): 15-27.

Jeffreys, Alec J., Liisa Kauppi and Rita Neumann (2001). "Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex." Nat Genet **29**(2): 217-222.

Johnson, Gillian C. L., Laura Esposito, Bryan J. Barratt, Annabel N. Smith, Joanne Heward, Gianfranco Di Genova, Hironori Ueda, Heather J. Cordell, Iain A. Eaves, Frank Dudbridge, Rebecca C. J. Twells, Felicity Payne, Wil Hughes, Sarah Nutland, Helen Stevens, Phillipa Carr, Eva Tuomilehto-Wolf, Jaakko Tuomilehto, Stephen C. L. Gough, David G. Clayton and John A. Todd (2001). "Haplotype tagging for the identification of common disease genes." Nat Genet **29**(2): 233-237.

Jorde, L. B., A. R. Rogers, M. Bamshad, W. S. Watkins, P. Krakowiak, S. Sung, J. Kere and H. C. Harpending (1997). "Microsatellite diversity and the demographic history of modern humans." Proceedings of the National Academy of Sciences of the United States of America **94**(7): 3100-3103.

Kim, Philip M., Hugo Y.K. Lam, Alexander E. Urban, Jan O. Korbel, Jason Affourtit, Fabian Grubert, Xueying Chen, Sherman Weissman, Michael Snyder and Mark B. Gerstein (2008). "Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history." Genome Research **18**(12): 1865-1874.

Kim, Sobin and Ashish Misra (2007). "SNP Genotyping: Technologies and Biomedical Applications." Annual Review of Biomedical Engineering **9**(1): 289-320.

Kimpton, Colin P., Nicola J. Oldroyd, Stephanie K. Watson, Rachael R. E. Frazier, Peter E. Johnson, Emma S. Millican, Andrew Urquhart, Becky L. Sparkes and Peter Gill (1996). "Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification." ELECTROPHORESIS **17**(8): 1283-1293.

Klein, Robert J., Caroline Zeiss, Emily Y. Chew, Jen-Yue Tsai, Richard S. Sackler, Chad Haynes, Alice K. Henning, John Paul SanGiovanni, Shrikant M. Mane, Susan T. Mayne, Michael B. Bracken, Frederick L. Ferris, Jurg Ott, Colin Barnstable and Josephine Hoh (2005). "Complement Factor H Polymorphism in Age-Related Macular Degeneration." Science **308**(5720): 385-389.

Kulkarni, H, BK Agan, VC Marconi, RJ O'Connell, JF Camargo, W He, J Delmar, KR Phelps, G Crawford, RA Clark, MJ Dolan and SK Ahuja (2008). "CCL3L1-CCR5 genotype improves the assessment of AIDS Risk in HIV-1-infected individuals." PLoS ONE **3**(9): e3165.

Lederman, Michael M., Adam Penn-Nicholson, Michael Cho and Donald Mosier (2006). "Biology of CCR5 and Its Role in HIV Infection and Treatment." JAMA **296**(7): 815-826.

Locke, Devin P., Andrew J. Sharp, Steven A. McCarroll, Sean D. McGrath, Tera L. Newman, Ze Cheng, Stuart Schwartz, Donna G. Albertson, Daniel Pinkel, David M. Altshuler and Evan E. Eichler (2006). "Linkage Disequilibrium and Heritability of Copy-Number Polymorphisms within Duplicated Regions of the Human Genome." American journal of human genetics **79**(2): 275-290.

MacDonald, Marcy E., Andrea Novelletto, Carol Lin, Dan Tagle, Glenn Barnes, Gillian Bates, Sherry Taylor, Bernice Allitto, Michael Altherr, Richard Myers, Hans Lehrach, Francis S. Collins, John J. Wasmuth, Marina Frontali and James F. Gusella (1992). "The Huntington's disease candidate region exhibits many different haplotypes." Nat Genet **1**(2): 99-103.

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

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

Mamtani, Manju, Tomoyo Matsubara, Chisato Shimizu, Susumu Furukawa, Teiji Akagi, Yoshihiro Onouchi, Akira Hata, Akihiro Fujino, Weijing He, Sunil K. Ahuja and Jane C. Burns (2010). "Association of *CCR2-CCR5* Haplotypes and *CCL3L1* Copy Number with Kawasaki Disease, Coronary Artery Lesions, and IVIG Responses in Japanese Children." PLoS ONE **5**(7): e11458.

Manolio, Teri A. and Francis S. Collins (2009). "The HapMap and Genome-Wide Association Studies in Diagnosis and Therapy*." Annual Review of Medicine **60**(1): 443-456.

Martin, Maureen P., Michael Dean, Michael W. Smith, Cheryl Winkler, Bernard Gerrard, Nelson L. Michael, Benhur Lee, Robert W. Doms, Joseph Margolick, Susan Buchbinder, James J. Goedert, Thomas R. O'Brien, Margaret W. Hilgartner, David Vlahov, Stephen J. O'Brien and Mary Carrington (1998). "Genetic Acceleration of AIDS Progression by a Promoter Variant of CCR5." Science **282**(5395): 1907-1911.

Martinson, Jeremy J., Nicola H. Chapman, David C. Rees, Yan-Tat Liu and John B. Clegg (1997). "Global distribution of the CCR5 gene 32-basepair deletion." Nat Genet **16**(1): 100-103.

Migueles, Stephen A., M. Shirin Sabbaghian, W. Lesley Shupert, Maria P. Bettinotti, Francesco M. Marincola, Lisa Martino, Clair W. Hallahan, Sara M. Selig, David Schwartz, John Sullivan and Mark Connors (2000). "HLA B*5701 is highly associated with restriction of virus *replication in a subgroup of* HIV-infected long term nonprogressors." Proceedings of the National Academy of Sciences of the United States of America **97**(6): 2709-2714.

Modi, W. S. (2004). "CCL3L1 and CCL4L1 chemokine genes are located in a segmental duplication at chromosome 17q12." Genomics **83**(4): 735-738.

Modi, WS, J Lautenberger, P An, K Scott, JJ Goedert, GD Kirk, S Buchbinder, J Phair, S Donfield, SJ O'Brien and C Winkler (2006). "Genetic variation in the CCL18-CCL3-CCL4 chemokine gene cluster influences HIV Type 1 transmission and AIDS disease progression." Am J Hum Genet **79**(1): 120 - 128.

Möller, Marlo, Erika De Wit and Eileen G. Hoal (2010). "Past, present and future directions in human genetic susceptibility to tuberculosis." FEMS Immunology & Medical Microbiology **58**(1): 3-26.

Mummidi, Srinivas, Seema S. Ahuja, Enrique Gonalez, Stephanie A. Anderson, Elvin N. *Santiago*, Kevin T. Stephan, Fiona E. Craig, Peter O'Connell, Victor Tryon, Robert A. Clark, Matthew J. Dolan *and* Sunil K. Ahuja (1998). "Genealogy of the CCR5 locus and chemokine system gene variants associated with altered rates of HIV-1 disease progression." Nat Med **4**(7): 786-793.

Musunuru, Kiran, Alanna Strong, Maria Frank-Kamenetsky, Noemi E. Lee, Tim Ahfeldt, Katherine V. Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M. Ruda, James P. Pirruccello, Brian Muchmore, Ludmila Prokunina-Olsson, Jennifer L. Hall, Eric E. Schadt, Carlos R. Morales, Sissel Lund-Katz, Michael C. Phillips, Jamie Wong, William Cantley, Timothy Racie, Kenechi G. Ejebe, Marju Orho-Melander, Olle Melander, Victor Koteliansky, Kevin Fitzgerald, Ronald M. Krauss, Chad A. Cowan, Sekar Kathiresan and Daniel J. Rader (2010). "From noncoding variant to phenotype *via SORT1 at the 1p13* cholesterol locus." Nature **466**(7307): 714-719.

Myers, Simon, Leonardo Bottolo, Colin Freeman, Gil McVean and Peter Donnelly (2005). "A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome." Science **310**(5746): 321-324.

Nakajima, Toshiaki, Hitoshi Ohtani, Taeko Naruse, Hiroki Shibata, Jun-ich Mimaya, Hiroshi Terunuma and Akinori Kimura (2007). "Copy number variations of CCL3L1 and long-term prognosis of HIV-1 infection in asymptomatic HIV-infected Japanese with hemophilia." Immunogenetics **59**(10): 793-798.

Nakao, M, H Nomiyama and K Shimada (1990). "Structures of human genes coding for cytokine LD78 and their expression." Mol. Cell. Biol. **10**(7): 3646-3658.

Nibbs, Robert J. B., Jinying Yang, Nathaniel R. Landau, Jian-Hua Mao and Gerard J. Graham (1999). "LD78β, A Non-allelic Variant of Human MIP-1α (LD78α), Has Enhanced Receptor Interactions and Potent HIV Suppressive Activity." Journal of Biological Chemistry **274**(25): 17478-17483.

Niu, Wenquan, Yue Qi, Shuqin Hou, Xiaoyan Zhai, Wenyu Zhou and Changchun Qiu (2009). "Haplotype-based association of the renin-angiotensin-aldosterone system genes polymorphisms with essential hypertension among Han Chinese: the Fangshan study." Journal of Hypertension **27**(7): 1384-1391 10.1097/HJH.0b013e32832b7e0d.

Paximadis, M., N. Mohanlal, G. E. Gray, L. Kuhn and C. T. Tiemessen (2009). "Identification of new variants within the two functional genes CCL3 and CCL3L encoding the CCL3 (MIP-1α) chemokine: implications for HIV-1 infection." International Journal of Immunogenetics **36**(1): 21-32.

Payseur, Bret A., Michael Place and James L. Weber (2008). "Linkage Disequilibrium between STRPs and SNPs across the Human Genome." The American Journal of Human Genetics **82**(5): 1039-1050.

Perry, George H., Fengtang Yang, Tomas Marques-Bonet, Carly Murphy, Tomas Fitzgerald, Arthur S. Lee, Courtney Hyland, Anne C. Stone, Matthew E. Hurles, Chris Tyler-Smith, Evan E. Eichler, Nigel P. Carter, Charles Lee and Richard Redon (2008). "Copy number variation and evolution in humans and chimpanzees." Genome Research **18**(11): 1698-1710.

Piacentini, L., M. Biasin, C. Fenizia and M. Clerici (2009). "Genetic correlates of protection against HIV infection: the ally within." Journal of Internal Medicine **265**(1): 110-124.

Pinnaduwage, Dushanthi and Laurent Briollais (2005). "Comparison of genotype- and haplotype-based approaches for fine-mapping of alcohol dependence using COGA data." BMC Genetics **6**(Suppl 1): S65.

Popat, S., R. Hubner and R.S. Houlston (2005). "Systematic Review of Microsatellite Instability and Colorectal Cancer Prognosis." Journal of Clinical Oncology **23**(3): 609-618.

Reich, David E., Michele Cargill, Stacey Bolk, James Ireland, Pardis C. Sabeti, Daniel J. Richter, Thomas Lavery, Rose Kouyoumjian, Shelli F. Farhadian, Ryk Ward and Eric S. Lander (2001). "Linkage disequilibrium in the human genome." Nature **411**(6834): 199-204.

Ruvolo, M. (1997). "GENETIC DIVERSITY IN HOMINOID PRIMATES." Annual Review of Anthropology **26**(1): 515-540.

Satake, Wataru, Yuko Nakabayashi, Ikuko Mizuta, Yushi Hirota, Chiyomi Ito, Michiaki Kubo, Takahisa Kawaguchi, Tatsuhiko Tsunoda, Masahiko Watanabe, Atsushi Takeda, Hiroyuki Tomiyama, Kenji Nakashima, Kazuko Hasegawa, Fumiya Obata, Takeo Yoshikawa, Hideshi Kawakami, Saburo Sakoda, Mitsutoshi Yamamoto, Nobutaka Hattori, Miho Murata, Yusuke Nakamura and Tatsushi Toda (2009). "Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease." Nat Genet **41**(12): 1303-1307.

Saukkonen, Jussi J., Beth Bazydlo, Michael Thomas, Robert M. Strieter, Joseph Keane and Hardy Kornfeld (2002). "β-Chemokines Are Induced by Mycobacterium tuberculosis and Inhibit Its Growth." Infect. Immun. **70**(4): 1684-1693.

Scott, Laura J., Karen L. Mohlke, Lori L. Bonnycastle, Cristen J. Willer, Yun Li, William L. Duren, Michael R. Erdos, Heather M. Stringham, Peter S. Chines, Anne U. Jackson, Ludmila Prokunina-Olsson, Chia-Jen Ding, Amy J. Swift, Narisu Narisu, Tianle Hu, Randall Pruim, Rui Xiao, Xiao-Yi Li, Karen N. Conneely, Nancy L. Riebow, Andrew G. Sprau, Maurine Tong, Peggy P. White, Kurt N. Hetrick, Michael W. Barnhart, Craig W. Bark, Janet L. Goldstein, Lee Watkins, Fang Xiang, Jouko Saramies, Thomas A. Buchanan, Richard M. Watanabe, Timo T. Valle, Leena Kinnunen, Gonçalo R. Abecasis, Elizabeth W. Pugh, Kimberly F. Doheny, Richard N. Bergman, Jaakko Tuomilehto, Francis S. Collins and Michael Boehnke (2007). "A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants." Science **316**(5829): 1341-1345.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg and M. Wigler (2004). "Large-scale copy number polymorphism in the human genome." Science **305**(5683): 525-8.

Sheppard, Haynes W., Connie Celum, Nelson L. Michael, Stephen O'Brien, Michael Dean, Mary Carrington, Dale Dondero and Susan P. Buchbinder (2002). "HIV-1 Infection in Individuals With the CCR5-[DELTA]32/[DELTA]32 Genotype: Acquisition of Syncytium-Inducing Virus at Seroconversion." JAIDS Journal of Acquired Immune Deficiency Syndromes **29**(3): 307-313.

Shostakovich-Koretskaya, Ludmila, Gabriel Catano, Zoya A Chykarenko, Weijing He, German Gornalusse, Srinivas Mummidi, Racquel Sanchez, Matthew J Dolan, Seema S Ahuja, Robert A Clark, Hemant Kulkarni and Sunil K Ahuja (2009). "Combinatorial content of CCL3L and CCL4L gene copy numbers influence HIV-AIDS susceptibility in Ukrainian children." AIDS **23**(6): 679-688 10.1097/QAD.0b013e3283270b3f.

Shrestha, Sadeep, Mawuli Nyaku and Jeffrey Edberg (2010). "Variations in CCL3L gene cluster sequence and non-specific gene copy numbers." BMC Research Notes **3**(1): 74.

Shrestha, Sadeep, Jianming Tang and Richard A. Kaslow (2009). "Gene copy number: learning to count past two." Nat Med **15**(10): 1127-1129.

Sibley, Charles G. and Jon E. Ahlquist (1984). "The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization." Journal of Molecular Evolution **20**(1): 2-15.

Sibley, Charles G. and Jon E. Ahlquist (1987). "DNA hybridization evidence of hominoid phylogeny: Results from an expanded data set." Journal of Molecular Evolution **26**(1): 99-121.

Silva, Eric and Michael P.H. Stumpf (2004). "HIV and the CCR5-D32 resistance allele." FEMS Microbiology Letters **241**(1): 1-12.

Simón-Sánchez, Javier, Claudia Schulte, Jose M. Bras, Manu Sharma, J. Raphael Gibbs, Daniela Berg, Coro Paisan-Ruiz, Peter Lichtner, Sonja W. Scholz, Dena G. Hernandez, Rejko Kruger, Monica Federoff, Christine Klein, Alison Goate, Joel Perlmutter, Michael Bonin, Michael A. Nalls, Thomas Illig, Christian Gieger, Henry Houlden, Michael Steffens, Michael S. Okun, Brad A. Racette, Mark R. Cookson, Kelly D. Foote, Hubert H. Fernandez, Bryan J. Traynor, Stefan Schreiber, Sampath Arepalli, Ryan Zonozi, Katrina Gwinn, Marcel van der Brug, Grisel Lopez, Stephen J. Chanock, Arthur Schatzkin, Yikyung Park, Albert Hollenbeck, Jianjun Gao, Xuemei Huang, Nick W. Wood, Delia Lorenz, Gunther Deuschl, Honglei Chen, Olaf Riess, John A. Hardy, Andrew B. Singleton and Thomas Gasser (2009). "Genome-wide association study reveals genetic risk underlying Parkinson's disease." Nat Genet **41**(12): 1308-1312.

Skwor, Troy A., Shannon Sedberry Allen, John T. Mackie, Karen Russell, Luc R. Berghman and David N. McMurray (2006). "BCG vaccination of guinea pigs modulates Mycobacterium tuberculosis-induced CCL5 (RANTES) production in vitro and in vivo." Tuberculosis **86**(6): 419-429.

Slatkin, Montgomery (2008). "Linkage disequilibrium - understanding the evolutionary past and mapping the medical future." Nat Rev Genet **9**(6): 477-485.

Sobti, Ranbir, Nega Berhane, Salih Mahdi, Rupinder Kler, Seyed Hosseini, Vijish Kuttiat and Ajay Wanchu (2010). "Impact of ERCC2 gene polymorphism on HIV-1 disease progression to AIDS among North Indian HIV patients." Molecular Biology Reports.

Staiti, N., D. Di Martino and L. Saravo (2004). "A novel approach in personal identification from tissue samples undergone different processes through STR typing." Forensic Science International **146**(Supplement 1): S171-S173.

The International HapMap 3 Consortium (2010). "Integrating common and rare genetic variation in diverse human populations." Nature **467**(7311): 52-58.

The International HapMap Consortium (2003). "The International HapMap Project." Nature **426**(6968): 789-796.

The International HapMap Consortium (2005). "A haplotype map of the human genome." Nature **437**: 1299-1320.

The International HapMap Consortium (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-861.

The International SNP Map Working Group (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature **409**(6822): 928-933.

The Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-678.

The Wellcome Trust Case Control Consortium (2010). "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls." Nature **464**(7289): 713-720.

Thomas, Rasmi, Richard Apps, Ying Qi, Xiaojiang Gao, Victoria Male, Colm O'HUigin, Geraldine O'Connor, Dongliang Ge, Jacques Fellay, Jeffrey N. Martin, Joseph Margolick, James J. Goedert, Susan Buchbinder, Gregory D. Kirk, Maureen P. Martin, Amalio Telenti, Steven G. Deeks, Bruce D. Walker, David Goldstein, Daniel W. McVicar, Ashley Moffett and Mary Carrington (2009). "HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C." Nat Genet **41**(12): 1290-1294.

Thomson, J. A., V. Pilotti, P. Stevens, K. L. Ayres and P. G. Debenham (1999). "Validation of short tandem repeat analysis for the investigation of cases of disputed paternity." Forensic Science International **100**(1-2): 1-16.

Tregouet, David-Alexandre, Sandrine Barbaux, Sylvie Escolano, Nadia Tahri, Jean-Louis Golmard, Laurence Tiret and FranÃ§ois Cambien (2002). "Specific haplotypes of the P-selectin gene are associated with myocardial infarction." Human Molecular Genetics **11**(17): 2015-2023.

Tung, Jenny, Susan C. Alberts and Gregory A. Wray (2010). "Evolutionary genetics in wild primates: combining genetic approaches with field studies of natural populations." Trends in genetics : TIG **26**(8): 353-362.

Turner, Daniel J. and Matthew E. Hurles (2009). "High-throughput haplotype determination over long distances by haplotype fusion PCR and ligation haplotyping." Nat. Protocols **4**(12): 1771-1783.

Urban, Thomas J., Amy C. Weintrob, Jacques Fellay, Sara Colombo, Kevin V. Shianna, Curtis Gumbs, Margalida Rotger, Kimberly Pelak, Kristen K. Dang, Roger Detels, Jeremy J. Martinson, Stephen J. O'Brien, Norman L. Letvin, Andrew J. McMichael, Barton F. Haynes, Mary Carrington, Amalio Telenti, Nelson L. Michael and David B. Goldstein (2009). "CCL3L1 and HIV/AIDS susceptibility." Nat Med **15**(10): 1110-1112.

Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, Robert A. Holt, Jeannine D. Gocayne, Peter Amanatides, Richard M. Ballew, Daniel H. Huson, Jennifer Russo Wortman, Qing Zhang, Chinnappa D. Kodira, Xiangqun H. Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul D. Thomas, Jinghui Zhang, George L. Gabor Miklos, Catherine Nelson, Samuel Broder, Andrew G. Clark, Joe Nadeau, Victor A. McKusick, Norton Zinder, Arnold J. Levine, Richard J. Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel

Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina Di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei E. Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas J. Heiman, Maureen E. Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen A. Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady V. Merkulov, Natalia Milshina, Helen M. Moore, Ashwinikumar K Naik, Vaibhav A. Narayan, Beena Neelam, Deborah Nusskern, Douglas B. Rusch, Steven Salzberg, Wei Shao, Bixiong Shue, Jingtao Sun, Zhen Yuan Wang, Aihui Wang, Xin Wang, Jian Wang, Ming-Hui Wei, Ron Wides, Chunlin Xiao, Chunhua Yan, Alison Yao, Jane Ye, Ming Zhan, Weiqing Zhang, Hongyu Zhang, Qi Zhao, Liansheng Zheng, Fei Zhong, Wenyan Zhong, Shiaoping C. Zhu, Shaying Zhao, Dennis Gilbert, Suzanna Baumhueter, Gene Spier, Christine Carter, Anibal Cravchik, Trevor Woodage, Feroze Ali, Huijin An, Aderonke Awe, Danita Baldwin, Holly Baden, Mary Barnstead, Ian Barrow, Karen Beeson, Dana Busam, Amy Carver, Angela Center, Ming Lai Cheng, Liz Curry, Steve Danaher, Lionel Davenport, Raymond Desilets, Susanne Dietz, Kristina Dodson, Lisa Doup, Steven

Ferriera, Neha Garg, Andres Gluecksmann, Brit Hart, Jason Haynes, Charles Haynes, Cheryl Heiner, Suzanne Hladun, Damon Hostin, Jarrett Houck, Timothy Howland, Chinyere Ibegwam, Jeffery Johnson, Francis Kalush, Lesley Kline, Shashi Koduru, Amy Love, Felecia Mann, David May, Steven McCawley, Tina McIntosh, Ivy McMullen, Mee Moy, Linda Moy, Brian Murphy, Keith Nelson, Cynthia Pfannkoch, Eric Pratts, Vinita Puri, Hina Qureshi, Matthew Reardon, Robert Rodriguez, Yu-Hui Rogers, Deanna Romblad, Bob Ruhfel, Richard Scott, Cynthia Sitter, Michelle Smallwood, Erin Stewart, Renee Strong, Ellen Suh, Reginald Thomas, Ni Ni Tint, Sukyee Tse, Claire Vech, Gary Wang, Jeremy Wetter, Sherita Williams, Monica Williams, Sandra Windsor, Emily Winn-Deen, Keriellen Wolfe, Jayshree Zaveri, Karena Zaveri, Josep F. Abril, Roderic GuigÃ³, Michael J. Campbell, Kimmen V. Sjolander, Brian Karlak, Anish Kejariwal, Huaiyu Mi, Betty Lazareva, Thomas Hatton, Apurva Narechania, Karen Diemer, Anushya Muruganujan, Nan Guo, Shinji Sato, Vineet Bafna, Sorin Istrail, Ross Lippert, Russell Schwartz, Brian Walenz, Shibu Yooseph, David Allen, Anand Basu, James Baxendale, Louis Blick, Marcelo Caminha, John Carnes-Stine, Parris Caulk, Yen-Hui Chiang, My Coyne, Carl Dahlke, Anne Deslattes Mays, Maria Dombroski, Michael Donnelly, Dale Ely, Shiva Esparham, Carl Fosler, Harold Gire, Stephen Glanowski, Kenneth Glasser, Anna Glodek, Mark Gorokhov, Ken Graham, Barry Gropman, Michael Harris,

Jeremy Heil, Scott Henderson, Jeffrey Hoover, Donald Jennings, Catherine Jordan, James Jordan, John Kasha, Leonid Kagan, Cheryl Kraft, Alexander Levitsky, Mark Lewis, Xiangjun Liu, John Lopez, Daniel Ma, William Majoros, Joe McDaniel, Sean Murphy, Matthew Newman, Trung Nguyen, Ngoc Nguyen, Marc Nodell, Sue Pan, Jim Peck, Marshall Peterson, William Rowe, Robert Sanders, John Scott, Michael Simpson, Thomas Smith, Arlan Sprague, Timothy Stockwell, Russell Turner, Eli Venter, Mei Wang, Meiyuan Wen, David Wu, Mitchell Wu, Ashley Xia, Ali Zandieh and Xiaohong Zhu (2001). "The Sequence of the Human Genome." Science **291**(5507): 1304-1351.

Virgos, Carmen, Lourdes Martorell, Joaquln Valero, Fernando Civeira, Jorge Joven, Antonio Labad, Elisabet Vilella and Lldia Figuera (2001). "Association study of schizophrenia with polymorphisms at six candidate genes." Schizophrenia Research **49**(1-2): 65-71.

Vynnycky, Emilia and Paul E. M. Fine (2000). "Lifetime Risks, Incubation Period, and Serial Interval of Tuberculosis." American Journal of Epidemiology **152**(3): 247-263.

Vyshkina, Tamara and Bernadette Kalman (2006). "Analyses of a MS-associated haplotype encompassing the CCL3 gene." Journal of Neuroimmunology **176**(1-2): 216-218.

Walker, Susan (2009). Fine-scale copy number polymorphism in the human genome, University of Nottingham. **PhD**.

Walker, Susan, Somwang Janyakhantikul and John A. L. Armour (2009). "Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs." Genomics **93**(1): 98-103.

Wallace, Chris, Stephen J. Newhouse, Peter Braund, Feng Zhang, Martin Tobin, Mario Falchi, Kourosh Ahmadi, Richard J. Dobson, Ana Carolina B. Marçano, Cother Hajat, Paul Burton, Panagiotis Deloukas, Morris Brown, John M. Connell, Anna Dominiczak, G. Mark Lathrop, John Webster, Martin Farrall, Tim Spector, Nilesh J. Samani, Mark J. Caulfield and Patricia B. Munroe (2008). "Genome-wide Association Study Identifies Genes for Biomarkers of Cardiovascular Disease: Serum Urate and Dyslipidemia." American journal of human genetics **82**(1): 139-149.

World Health Organization (2010). WHO report 2010: Global Tuberculosis Control 2010. Geneva, Switzerland.

Zhang, Feng, Wenli Gu, Matthew E. Hurles and James R. Lupski (2009). "Copy Number Variation in Human Health, Disease, and Evolution." Annual Review of Genomics and Human Genetics **10**(1): 451-481.

Zhang, Zhaolei, Paul M. Harrison, Yin Liu and Mark Gerstein (2003). "Millions of Years of Evolution Preserved: A Comprehensive Catalog of the Processed Pseudogenes in the Human Genome." Genome Research **13**(12): 2541-2558.

Zheng, Deyou, Adam Frankish, Robert Baertsch, Philipp Kapranov, Alexandre Reymond, Siew Woh Choo, Yontao Lu, France Denoeud, Stylianos E. Antonarakis, Michael Snyder, Yijun Ruan, Chia-Lin Wei, Thomas R. Gingeras, Roderic Guigó, Jennifer Harrow and Mark B. Gerstein (2007). "Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution." Genome Research **17**(6): 839-851.

Zheng, Deyou, Zhaolei Zhang, Paul M. Harrison, John Karro, Nick Carriero and Mark Gerstein (2005). "Integrated Pseudogene Annotation for Human Chromosome 22: Evidence for Transcription." Journal of Molecular Biology **349**(1): 27-45.

Zhivotovsky, Lev A., Noah A. Rosenberg and Marcus W. Feldman (2003). "Features of Evolution and Expansion of Modern Humans, Inferred from Genomewide Microsatellite Markers." The American Journal of Human Genetics **72**(5): 1171-1186.

Appendix

Table 1: Prediction of *CCL3L1/CCL4L1* copy number in ECACC samples (panel 1) by using SNP rs16972085 and SNP rs8064426.

Sample	SNP rs16972085	SNP rs8064426	Predicted CN	PRT- CN
C0744	N/A*	GG	N/A*	2
C0147	AA	AG	1	2
C0888	AA	AG	1	1
C0909	AA	GG	2	2
C0938	AA	GG	2	3
C0027	AA	AG	1	1
C0184	AA	GG	2	2
C0969	N/A*	AA	N/A*	2
C0145	AA	GG	2	2
C0096	AA	AA	0	2
C0007	AA	GG	2	3
C0029	AA	GG	2	2
C0156	AA	GG	2	2
C0143	AG	GG	3	2
C0917	AA	GG	2	1
C0093	AG	AG	2	3
C0140	AA	GG	2	2
C0877	AA	GG	2	4
C0185	AG	AG	2	4
C0063	AA	GG	2	2
C0937	AA	GG	2	4
C0861	AA	GG	2	2
C0207	AA	GG	2	2
C0202	AA	GG	2	2
C0068	AA	GG	2	2
C0884	AA	GG	2	2
C0961	AA	GG	2	2
C0195	AA	GG	2	2
C0187	AA	GG	2	1
C0725	AA	GG	2	3
C0960	AA	GG	2	2
C0748	AA	GG	2	2
C0208	AA	GG	2	1
C0100	AG	GG	3	3
C0090	AG	GG	3	3
C0870	AA	GG	2	2
C0188	AA	GG	2	2
C0786	AA	GG	2	2
C0097	AA	GG	2	2
C0176	AA	GG	2	2
C0858	AA	AG	1	1
C0864	AA	GG	2	3
C0735	AA	GG	2	3
C0088	GG	GG	4	4
C0739	AG	GG	3	3
C0204	AA	GG	2	2
C0194	AA	GG	2	1
C0126	AA	GG	2	1
C0978	AA	GG	2	3
C0183	AA	GG	2	2

Table 1: Prediction of *CCL3L1/CCL4L1* copy number in ECACC samples (panel 1) by using SNP rs16972085 and SNP rs8064426 (continued).

Sample	SNP rs16972085	SNP rs8064426	Predicted CN	PRT- CN
C0053	AA	GG	2	2
C0766	AA	AG	1	2
C0034	AA	AG	1	2
C0856	AA	GG	2	2
C0896	AA	GG	2	1
C0215	AA	GG	2	2
C0136	AA	GG	2	2
C0913	AA	GG	2	2
C1006	GG	AG	4**	4
C0863	AA	AG	1	1
C0152	AA	AG	1	1
C0150	AA	GG	2	2
C0899	AA	GG	2	2
C0747	AA	GG	2	2
C0095	AA	GG	2	2
C0075	AA	AG	1	1
C0906	AA	GG	2	2
C0901	AA	GG	2	2
C0107	AG	AG	2	3
C0855	AA	GG	2	2
C0157	AA	GG	2	2
C0182	AA	GG	2	2
C0857	AA	GG	2	3
C0939	AA	GG	2	2
C0898	AA	GG	2	1
C0189	AA	GG	2	2
C0722	AA	GG	2	2
C0908	AA	GG	2	1
C0907	AG	GG	3	3
C0849	AA	GG	2	2
C0121	AG	GG	3	3
C0045	AA	GG	2	3
C0073	AA	AG	1	1
C0940	AA	AG	1	1
C0080	AA	GG	2	1
C0997	AA	GG	2	2
C0904	AA	GG	2	2
C0066	AA	GG	2	1
C0108	AA	AG	1	2
C0941	AG	GG	3	3
C0996	AA	GG	2	1
C0210	AA	GG	2	3
C0878	AG	GG	3	2
C0953	AA	GG	2	3
C0081	AA	GG	2	2
C0036	AG	GG	3	2

Note: *N/A = not available

**The prediction is based on the allele of SNP rs16972085

Table 2: Prediction of *CCL3L1/CCL4L1* copy number in ECACC samples (panel 2) by using SNP rs16972085 and SNP rs8064426.

Code	SNP rs16972085	SNP rs8064426	Predicted CN	PRT- CN
C0871	AA	GG	2	2
C0098	AA	GG	2	3
C0880	AA	GG	2	2
C0061	AA	GG	2	2
C0018	AA	GG	2	3
C0724	AA	GG	2	2
C0892	AA	GG	2	2
C0164	AG	AG	2	2
C0008	AA	GG	2	2
C0897	AA	GG	2	2
C0006	AA	GG	2	2
C0124	AA	GG	2	2
C0106	AA	GG	2	2
C0755	AA	GG	2	2
C0848	AG	GG	3	3
C0723	AA	GG	2	2
C0958	AA	GG	2	2
C0186	AA	AA	0	0
C0058	AA	GG	2	3
C0862	AA	GG	2	3
C0990	AA	GG	2	2
C0030	AG	GG	3	3
C0123	AG	GG	3	3
C1011	AA	GG	2	1
C0113	AA	GG	2	2
C0015	AA	GG	2	2
C0178	AA	AG	1	1
C0728	AG	GG	3	3
C0035	AA	GG	2	2
C0166	AA	GG	2	2
C0968	AA	GG	2	2
C0040	AA	GG	2	2
C0001	AG	AG	2	2
C0882	AA	GG	2	2
C0055	AA	GG	2	2
C0846	AA	GG	2	3
C0967	AA	GG	2	2
C0016	AA	GG	2	2
C0167	AA	GG	2	2
C0051	AG	GG	3	3
C0085	AA	GG	2	2
C0203	AA	GG	2	3
C0881	AG	GG	3	3
C0850	AA	GG	2	2
C0060	AA	AG	1	1
C0191	AA	GG	2	3
C0894	AA	GG	2	2
C0192	AA	AG	1	2
C0065	AA	AG	1	2
C0040	AA	GG	2	2

Table 2: Prediction of *CCL3L1/CCL4L1* copy number in ECACC samples (panel 2) by using SNP rs16972085 and SNP rs8064426 (continued).

Code	SNP rs16972085	SNP rs8064426	Predicted CN	PRT- CN
C0921	AA	AG	1	1
C0920	AA	AG	1	2
C0752	AA	GG	2	2
C0052	AA	AG	1	0
C0731	AG	AG	2	2
C0959	AA	GG	2	2
C0002	AA	AG	1	1
C0135	AA	GG	2	2
C0161	AA	GG	2	2
C0038	AA	GG	2	2
C0084	AG	GG	3	2
C0091	AA	AG	1	1
C0977	AA	AG	1	1
C0047	AA	GG	2	2
C0886	AA	GG	2	3
C0139	AA	GG	2	2
C0201	AA	GG	2	2
C0956	N/A*	GG	N/A*	2
C0902	AA	GG	2	2
C0168	AG	GG	3	2
C0854	AA	AG	1	1
C0137	AA	GG	2	2
C0149	AA	GG	2	3
C0874	AA	AG	1	1
C0738	AA	GG	2	2
C0180	AG	AG	2	2
C0160	AG	GG	3	3
C0009	AA	GG	2	2
C0172	AA	AG	1	1
C0190	AA	GG	2	3
C0154	AA	GG	2	2
C1010	AG	GG	3	2
C0741	AA	AG	1	1
C1008	AA	GG	2	2
C0994	AA	GG	2	2
C0109	AA	GG	2	2
C0111	AA	GG	2	2
C0022	AA	AG	1	1
C0868	AA	GG	2	2
C0851	AA	GG	2	2
C0891	AA	AG	1	1
C0883	AA	AG	1	1
C0197	AG	GG	3	3
C0196	AA	GG	2	2
C0893	AA	GG	2	2
C0159	AA	GG	2	2

Note: *N/A = not available

Table 3: Prediction of *CCL3L1/CCL4L1* copy number in Basque samples by using SNP rs16972085 and SNP rs8064426.

Code	SNP rs16972085	SNP rs8064426	Predicted CN	PRT- CN
GK1	AA	GG	2	2
GK2	AA	AG	1	3
GK3	AA	GG	2	1
GK4	AA	GG	2	2
GK5	AA	GG	2	2
GK6	AG	AG	2	1
GK7	AA	GG	2	2
GK8	AA	GG	2	2
GK9	AG	GG	3	3
GK10	AA	GG	2	3
GK11	AA	AG	1	1
GK12	AA	GG	2	2
BM1	AA	AG	1	1
BM2	AA	GG	2	2
BM3	AA	GG	2	2
BM4	AA	GG	2	2
BM5	AA	GG	2	2
BM6	AA	AG	1	1
BM7	AA	GG	2	2
BM8	AA	GG	2	2
BM9	AA	GG	2	2
ZK1	AA	GG	2	N/A*
ZK2	AA	GG	2	2
ZK3	AA	GG	2	2
ZK4	AA	GG	2	2
ZK5	AA	AG	1	1
ZK6	AA	GG	2	3
ZK7	AA	AG	1	1
ZK8	AA	GG	2	3
ZK9	AA	GG	2	2
ZK10	AA	AG	1	1
ZK11	AA	AG	1	1
ZK12	AG	GG	3	2
ZK13	AA	AG	1	3
ZK14	AA	AG	1	1
ZK15	AA	GG	2	2
ZK16	AA	GG	2	2
ZK17	AA	AG	1	1
ZK18	AA	GG	2	2
ZK19	AA	GG	2	2
ZK20	AA	GG	2	2
ZK21	AA	GG	2	2
DT1	AA	GG	2	2
DT2	AA	GG	2	3
DT3	AA	AG	1	1
DT4	AA	GG	2	2
DT5	AA	GG	2	2
DT6	AG	GG	3	3
DT7	AA	GG	2	2
DT8	AA	GG	2	2

Table 3: Prediction of *CCL3L1/CCL4L1* copy number in Basque samples by using SNP rs16972085 and SNP rs8064426 (continued).

Code	SNP rs16972085	SNP rs8064426	Predicted CN	PRT- CN
DT9	AA	GG	2	2
DT10	AA	AG	1	2
DT11	AA	GG	2	N/A*
DT12	AA	AG	1	1
DT13	AA	GG	2	1
DT14	AA	AG	1	2
DT15	AA	GG	2	2
DT16	AA	GG	2	2
DT17	AG	GG	3	3
DT18	AA	GG	2	1
TX1	AG	GG	3	2
TX2	AA	GG	2	2
TX3	AA	GG	2	2
TX4	AA	GG	2	2
TX5	AA	GG	2	2
TX6	AA	GG	2	2
TX7	AA	GG	2	2
TX8	AA	AG	1	N/A*
TX9	AA	GG	2	2
TX10	AA	GG	2	3
TX11	AA	GG	2	2
TX12	AA	GG	2	3
BS1	AA	AA	0	0
BS2	AA	GG	2	4
BS3	AA	GG	2	3
BS4	AA	GG	2	1
BS5	AA	GG	2	2
BS6	AG	GG	3	2
VI1	AA	GG	2	2
VI2	AA	AG	1	1
VI3	AA	GG	2	0
VI4	AG	GG	3	2
VI5	AA	GG	2	2
VI6	AA	GG	2	2
MR1	AA	GG	2	2
MR2	AA	GG	2	2
MR3	AA	AG	1	1
MR4	AA	GG	2	2
MR5	AA	GG	2	2
MR6	AA	GG	2	2
MR7	AA	GG	2	2
MR8	AA	GG	2	2
MR9	AA	GG	2	3
MR10	AA	GG	2	2
MR11	AA	GG	2	2
MR12	AA	GG	2	3
MR13	AA	GG	2	2
LK1	AA	GG	2	1
LK2	AA	AG	1	1
LK3	AA	GG	2	1
LK4	AA	GG	2	2

Table 3: Prediction of *CCL3L1/CCL4L1* copy number in Basque samples by using SNP rs16972085 and SNP rs8064426 (continued).

Code	SNP rs16972085	SNP rs8064426	Predicted CN	PRT- CN
LK5	AA	GG	2	2
LK6	AA	GG	2	2
LK7	AA	AG	1	1
LK8	AA	GG	2	2
LK9	AA	GG	2	3
LK10	AA	GG	2	3
LK11	AA	GG	2	3
LK12	AA	GG	2	2
LK13	AA	GG	2	3
LK14	AA	GG	2	3
LK15	AA	GG	2	3
LK16	AA	AA	0	3
LK17	AA	GG	2	2
LK18	AA	GG	2	2
LK19	AA	GG	2	2
LK20	AA	GG	2	2
LK21	AA	GG	2	2
LK22	AA	AG	1	1
CA1	AA	AG	1	1
CA2	AA	GG	2	2
CA3	AA	GG	2	2
CA4	AG	GG	3	4
CA5	AA	GG	2	2
CA6	AA	AA	0	0
CA7	AG	GG	3	2
CA8	AA	AG	1	3
CA9	AA	GG	2	2
CA10	AA	GG	2	2
CA11	AA	GG	2	2
CA12	AA	AG	1	1
CA13	AA	GG	2	2
CA14	AA	GG	2	1
CA15	AG	GG	3	3
CA16	AA	GG	2	2
CA17	AA	AG	1	1
CA18	AA	GG	2	3
CA19	AG	GG	3	3
CA20	AA	AG	1	1
CA21	AA	AG	1	1
CA22	AA	GG	2	2
CA23	AA	GG	2	2
CA24	AA	GG	2	2
AM1	AA	GG	2	2
AM2	AA	GG	2	2
AM3	AA	GG	2	3
AM4	AA	GG	2	3
AM5	AA	GG	2	3
AM6	AA	AG	1	0
IZ1	AA	GG	2	2
IZ2	AA	GG	2	2
IZ3	AA	GG	2	3

Table 3: Prediction of *CCL3L1/CCL4L1* copy number in Basque samples by using SNP rs16972085 and SNP rs8064426 (continued).

Code	SNP rs16972085	SNP rs8064426	Predicted CN	PRT- CN
IZ4	AA	AG	1	1
IZ5	AA	GG	2	2
IZ6	AA	GG	2	3
IZ7	AA	GG	2	2
IZ8	AA	GG	2	2

Note: *N/A = not available