# Investigation of Putative Susceptibility Regions

# to Inflammatory Bowel Disease

by

Bryan Michael Dechairo

A thesis submitted for the degree

Doctor of Philosophy

in the University of London

May 2001

Molecular Medicine Unit

Institute of Child Health

London

&

Oxagen Ltd.

Abingdon

ProQuest Number: U642129

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.
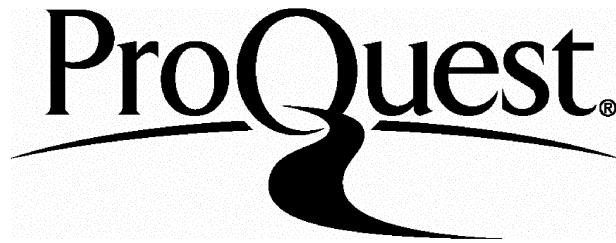


ProQuest U642129

Published by ProQuest LLC(2015). Copyright of the Dissertation is held by the Author.

**Dedicated to Richard Brown**

# Abstract

Inflammatory Bowel Disease (IBD) is a chronic inflammatory disease of the intestine, commonly diagnosed as either ulcerative colitis (UC) or Crohn's disease (CD). Epidemiological studies have consistently shown that both genetic and environmental factors influence the pathogenesis of IBD. Recently, a number of linkage and association studies have been conducted in cohorts of IBD families to try and identify the genetic component of IBD. Several putative linked loci and associated polymorphisms have been identified, with varying significance and consistency between studies.

In the present research, to further investigate and hopefully clarify previous linkage findings, a replication study on chromosome 6p (*IBD3*) and an extension study on chromosomes 3p were conducted. Short tandem repeat (STR) markers across each region were genotyped in 284 IBD affected sibling pairs (ASPs) from 234 families. A nonparametric peak multipoint LOD score of 3.0 was observed near D6S291, replicating the previous linkage to *IBD3*. No linkage was detected to the 3p locus.

Concurrently, due to the initial detection of linkage to the chromosome 3 locus using a subset of the present IBD cohort, association analysis was conducted across the 3p region and 4 candidate genes within it, GNAI2, CCR5, CCR2 and CCRL2. STR markers used in the present 3p linkage study and several biallelic polymorphisms identified across each gene were genotyped in 333 additional IBD simplex families. Significant positive association (P = 0.005) was detected between UC and the CCR5Δ32 polymorphism, previously associated with HIV. No other significant associations were detected.

# Acknowledgements

I am extremely thankful to my supervisor, Alisoun Carey, for her guidance, her many lesson such as tactfulness and humility, and most of all her belief in me which brought me across the Atlantic and continues to propel me forward. I would like to thank my supervisor Pete Scambler for his sponsorship and for his fresh and neutral spin on perplexing problems. I am grateful to Trevor Nichols and Mark Edwards for creating a unique and novel position for me at Oxagen and providing the financial support for my research.

I have enjoyed my time at Oxagen and am looking forward to the years to come. I would like to thank my lab mates – Max, Claire, John, Jo, Emma, Denise, Vicki, Marcela and James - for continuously putting up with my incessant babbling and singing, bad motivational habits that get me through my day. I also thank Gordy, Dave, Nige, Mark, and Gary for our regular morning conferences out back.

I continue to be enthralled with the theories and concepts of statistical genetics, and would like to thank both Lon and Mac for their patience and long conversations. I know the preceding comment may be biased yet, I do believe it is accurate.

Finally, I would like to acknowledge my family. Kathryn, who I met, fell in love with, and married during the first two years of my PhD has kept me motivated and brought me many coffees during my late nights of writing up. My Dad, who taught me perseverance and to stand up for what I believe in. My Mom, who continued to infuse and massage my emotional side so that I did not alienate to many people along the way. Thank you.

# Abbreviations

| | |
|---|---|
| A | adenosine deoxyribonucleotide monophosphate residue |
| ABI | Applied Biosystems, Inc. |
| APB | Amersham/Pharmacia Biotech |
| APS | ammonium persulphate |
| ARP | affected relative pair |
| ASP | affected sibling pair |
| BAC | bacterial artificial chromosome |
| BDKRB1 | bradykinin receptor B1 |
| bp | base pair |
| BSA | bovine serum albumin |
| C | cytosine deoxyribonucleotide monophosphate residue |
| °C | degrees Celsius |
| CCR | chemokine receptor |
| CD | Crohn's disease |
| cDNA | complementary DNA |
| cM | centiMorgan |
| cm | centimeter |
| cR | centiRay |
| CSBP1 | cytokine suppressive anti-inflammatory binding protein 1 |
| dATP | adenosine deoxyribonucleotide triphosphate |
| dCTP | cytosine deoxyribonucleotide triphosphate |
| dHPLC | denaturing high-performance liquid chromatography |
| dGTP | guanine deoxyribonucleotide triphosphate |
| DNA | deoxyribonucleic acid |

| | |
|---|---|
| dNTP | equimolar mixture of dATP, dCTP, dGTP and dTTP |
| dTTP | thiamine deoxyribonucleotide triphosphate |
| EDTA | ethylenediaminotetra-acetic acid |
| EST | expressed sequence tag |
| G | guanine deoxyribonucleotide monophosphate residue |
| g | gram for measures and gravity for centrifugation |
| G3 | Stanford G3 RH panel |
| G4 | MIT/Whitehead Genebridge 4 RH panel |
| gDNA | genomic DNA |
| GNAI2 | guanine nucleotide-binding protein $\alpha$i2 |
| HCl | hydrochloric acid |
| HLA | human leukocyte antigen |
| hr | hour |
| IBD | inflammatory bowel disease |
| IC | indeterminate colitis |
| IL | interleukin |
| kb | kilobase |
| $\lambda_s$ | sibling risk |
| LD | linkage disequilibrium |
| LOD | logarithm of the odds |
| LT | lymphotoxin |
| M | molar |
| Mb | megabase |
| $\mu$g | microgram |
| $\mu$l | microliter |
| mg | milligram |

| | |
|---|---|
| MgCl$_2$ | magnesium chloride |
| MHC | major histocompatibility complex |
| min | minute |
| ml | milliliter |
| mM | millimolar |
| NCBI | National Center for Biotechnology Information |
| ng | nanogram |
| nmol | nanomole |
| NRAMP | natural resistance associated macrophage protein |
| p | chromosomal short arm |
| P-ANCA | perinuclear antineutrophil cytoplasmic antibodies |
| PCR | polymerase chain reaction |
| PEP | primer extension pre-amplification |
| pmol | picomole |
| PPAR | peroxisome proliferator activated receptor |
| q | chromosomal long arm |
| RCA | rolling circle amplification |
| RFLP | restriction fragment length polymorphism |
| RH | radiation hybrid |
| RNA | ribonucleic acid |
| rpm | revolutions per minute |
| RR | relative risk |
| RT | room temperature |
| s | seconds |
| SAP | shrimp alkaline phosphatase |
| SNP | single nucleotide polymorphism |

| | |
|---|---|
| STR | short tandem repeat |
| STS | sequence tag site |
| T | thiamine deoxyribonucleotide monophosphate residue |
| $T_A$ | temperature of annealing |
| TBE | tris boric acid EDTA |
| TDT | transmission disequilibrium test |
| TEAA | triethylammonium acetate |
| $T_M$ | temperature of melting |
| TNF | tumour necrosis factor |
| u | units |
| UC | ulcerative colitis |
| V | volts |
| X | times |
| YAC | yeast artificial chromosome |

# Contents

13

# Figures

# Tables

# Chapter 1

# 1. Introduction

## 1.1. Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) is the medical term for any idiopathic, non-infectious inflammation of the intestinal tract. Many diseases fall under the IBD heading yet it is predominately comprised of the diseases: ulcerative colitis (UC) and Crohn's disease (CD). UC and CD, though clinically different, are similar enough to remain under the same medical heading and cause frequent misdiagnosis, especially when CD is limited to the colon (Parkes *et al.*, 1997). Difficulty in sub-diagnosing IBD is such that up to 10% of cases are labelled as indeterminate colitis (IC) following a review of clinical, endoscopic, and histological data (Satsangi *et al.*, 1998). UC normally onsets in and can be limited to the rectum (proctitis) but frequently extends into the more proximal colon and can include the whole colon. UC is characterised by a continuous pattern of inflammation and ulceration of the colonic mucosa and submucosa. In contrast, CD can affect any part of the gastrointestinal tract but is seen most in the terminal ileum and proximal colon. A segmental distribution of transmural inflammation, deep ulceration, and fissures, strictures, and fistulae characterise CD and may lead to a 'cobblestone' appearance upon dissection. Both UC and CD cause similar symptoms, diarrhoea, fever, cramping and abdominal pain, and similar complications, aphthous mouth ulcers, peripheral and central arthropathies, myocarditis, anterior uveitis, pericholangitis, sclerosing cholangitis and acute toxic megacolon (McConnell and Vadheim, 1992; Sorensen and Fonager, 1997).

The following chapter will summarise previous IBD epidemiological and genetic studies and will provide support and justification for the research undertaken for this thesis.

## 1.2. Epidemiology

Epidemiological studies are conducted to investigate what factors may cause or influence a disease. Epidemiological statistics may provide information about whether a disease is caused by a local phenomenon or more global factors, whether it is purely environmental, has a genetic component or both, and whether a disease has been correctly classified as a single, distinct disease. In the following section, a summary of data from IBD epidemiological studies addressing these questions will be presented.

### 1.2.1. Incidence and prevalence

IBD is a common disease. The incidence of IBD varies between epidemiological studies, normally ranging from 6 to 28 per 100,000 in European and North American Caucasian populations (Bjornsson *et al.*, 1998; Cipolla *et al.*, 1996; Fonager *et al.*, 1997; Manousos *et al.*, 1996; Russel and Stockbrugger, 1996). The incidence of IBD is mainly divided into the UC and CD disease types, with an incidence of UC from 5 to 18 per 100,000 and of CD from 1 to 10 per 100,000. In the same population, the prevalence of IBD ranges from 126 to 215 per 100,000, again with UC being more common with a prevalence between 90 and 161 per 100,000 compared to CD with a range of 33 to 76 per 100,000 (Cipolla *et al.*, 1996; Orholm *et al.*, 1991; Probert *et al.*, 1993; Satsangi *et al.*, 1998). IBD prevalence and incidence rates vary between continents with Europe and North America having the highest rates followed by Africa, South America, Australia and Asia (Andres and Friedman, 1999). For example, Oman

has an incidence of UC of only 1.35 per 100,000 (Radhakrishnan *et al.*, 1997) and Japan

even lower at 0.49 per 100,000 (Kitahora *et al.*, 1995). The geographical differences in

incidence and prevalence are not limited to continents but are also seen between

different latitudes. There is a trend of higher latitudes having increased rates of CD and

UC over more equatorial regions (Sandler, 1998). Differences in prevalence and

incidence are not only seen between regions but are also seen between different ethnic

groups within regions. For example, British Europeans have a higher prevalence of CD

(p < 0.001) and a lower prevalence of UC (p < 0.001) than British Hindus (Probert *et*

*al.*, 1993). Still, within ethnic groups rates of IBD continue to be influenced by

geography. African Americans, Ashkenazi Jews (American and European-born), and

British Southern Asians all show a higher incidence of UC and CD than Africans

(McConnell and Vadheim, 1992), Sephardic Jews (Israeli-born) (Lashner, 1995), and

Southern Asians (Probert *et al.*, 1993), respectively. Overall, incidence and prevalence

data from IBD epidemiological studies illustrate the commonality of the disease and

suggest that both geography (environment) and ethnicity (genetics) cause the

differences in rates between and within IBD cohorts.


## 1.2.2. Relative risk

The risk of IBD for relatives of patients with IBD (RR) has also been examined in

epidemiological studies. Increases in the RR of a disease can be caused either by

environmental factors, genetic factors or a combination of both. In Europe, studies have

shown a 6 to 15-fold increase in the risk of IBD for first-degree relatives (siblings,

parents and offspring) of patients with IBD (Cipolla *et al.*, 1996; Orholm *et al.*, 1991;

Russel *et al.*, 1997). Although UC is more common than CD (Section 1.2.1.), CD has a

larger RR. The first-degree RR of IBD in CD patients ranges from 15 to 38 while the

risk of IBD in UC is only between 11 and 15 (Binder and Orholm, 1996; Peeters *et al.*,

1996; Probert *et al.*, 1993). For relatives of CD patients the increase risk is mainly for

developing CD and for relatives of UC patients for the risk of developing UC. Only in

CD patients is there a significant increase in the first-degree RR of developing the

discordant disease, ranging from 3 to 4. Research does not show significant

geographical differences in RR but does show slight ethnic variation. For example,

Ashkenazi Jewish families with an IBD patient have a higher IBD RR than similar non-

Jewish families (Yang *et al.*, 1993). Looking at first-degree RR with more depth by

dividing it into parental, offspring and sibling risk ($\lambda_s$), studies have shown an increase

in these risks of approximately 15, 17 and 30 respectively for CD (Probert *et al.*, 1993;

Satsangi *et al.*, 1994) and 13, 11 and 12 respectively for UC (Kitahora *et al.*, 1995;

Probert *et al.*, 1993). Again, only in relatives of CD patients is there an increase

observed in the risk of developing the discordant disease ($\lambda_s = 16.6$) (Satsangi *et al.*,

1998). Studies have also looked at spouse risk, since an increase in RR caused solely by

environmental factors should have a corresponding increase in spouse risk. No

statistically significant increase in the risk of IBD for spouses of IBD patients has been

observed (McConnell and Vadheim, 1992). Therefore, together RR studies suggest that

the increase in risk observed in relatives of IBD patients is more likely caused by

genetic factors and that CD with a higher RR than UC may have a larger genetic

component than UC.


### 1.2.3. Concordance

Conducting concordance studies between and within families with a particular disease

helps to establish whether sub-classification of the disease into different disease types is

merited. At present, physicians and scientists agree that IBD should be sub-divided into

more than one disease but disagree to the extent of division warranted. For IBD the

concordance studies conducted have predominately dealt with concordance between twins, age or onset, and disease type and location.

## 1.2.3.1. Twin Studies

Concordance studies in cohorts of twins have shown that between 17 to 25 percent of monozygotic twins and 2 to 5 percent of dizygotic twins are concordant for IBD (Thompson *et al.*, 1996; Tysk *et al.*, 1988). Hence, a monozygotic twin of an IBD patient has a five-fold higher risk of developing IBD than a dizygotic twin of an IBD patient. Thus, since it is assumed that twins are normally raised in the same environment, the data would again suggest that genetics play a strong role in the aetiology of IBD. If the twin IBD cases are divided into CD and UC, CD again exhibits a stronger genetic influence with 20 to 44 percent and 4 to 7 percent concordance between monozygotic and dizygotic twins, respectively, compared to UC with 6 to 16 percent and 1 to 3 percent concordance. The different rates of concordance between UC and CD support the sub-classification of IBD into UC and CD. However, one case study was published where a pair of monozygotic twins was concordant for IBD but phenotypically discordant in type (Breslin *et al.*, 1997). The existence of such a family suggests that either the genetic factors of the two diseases are not completely exclusive or that the current sub-classification is not extensive enough and that another type of disease has been included together with either CD or UC or both (Section 1.3.1.2.). Overall, twin studies support genetic susceptibility to IBD, since a purely environmental disease should have even rates of concordance between monozygotic and dizygotic twins. However, with approximately 21% instead of 100% concordance between monozygotic twins, the data suggests that IBD is not purely caused by genetic factors either.

## 1.2.3.2. Age of onset

Consistently research shows that there is a strong concordance in the age of onset of IBD within a generation (p < 0.0001), but a strong discordance between different generations (p < 0.0001) with offspring having a younger age of diagnosis (Peeters *et al.*, 1996; Polito *et al.*, 1996; Satsangi *et al.*, 1996a). This discordance in age of diagnosis between generations is known as genetic anticipation and is seen in other genetic diseases (e.g. Huntington's disease). Genetic anticipation may be caused by a form of bias where children whose parents have a disease may be looking for signs that they too have the disease and are thus diagnosed earlier then normal when the disease manifests. However, counter to this argument of bias, it has been noted by scientists conducting these studies that children are frequently diagnosed with IBD prior to the diagnosis of their parents. Even so, the resulting concordance could also be explained by a bias where children living in a more health conscious culture may go to the doctor more frequently and for smaller irritations than would their parents. Today, genetic scientists continue to debate whether genetic anticipation is a real phenomenon caused by genetic (e.g. CAG repeat expansions) and/or environmental (e.g. fast-food culture) factors, or whether it is an artefact caused by diagnosis bias.

## 1.2.3.3. Disease type and location

Epidemiological studies show concordance in IBD disease type within families, especially between siblings (p < 0.001) (Lee and Lennard-Jones, 1996), again suggesting the existence of different genetic and environmental factors between CD and UC and again supporting their present sub-classification. Less consistent and weaker concordance within families for both extent and location of the disease is also reported (Bayless *et al.*, 1996; Colombel *et al.*, 1996; Peeters *et al.*, 1996), suggesting that further

division still may be warranted. Overall, data supports the division of IBD into UC and CD, but further division within the disease types may be justified.

## 1.3. Genetic models

After several epidemiological studies concluded that genetic factors influence IBD susceptibility, the focus of IBD research shifted towards studies that may discover these genetic factors. However, before trying to locate the genetic factors of a disease, it is important to understand the mode of inheritance of the disease. A genetic disease can either be caused by a single gene (Mendelian) or several genes (polygenic), with the mode of inheritance ranging from simple to complex in either model. A discussion of both of these genetic models and results from studies investigating which model best fits the inheritance of IBD will be presented in this section.

### 1.3.1. Mendelian

The simplest genetic model is the Mendelian model where a mutation in a single gene causes a disease. A Mendelian disorder can be categorised by one of four modes of inheritance, either dominant, recessive, sex-linked or imprinted. The mode of inheritance is usually easily deduced from studying affected pedigrees. However, this deduction becomes harder if environmental factors also influence the disease aetiology.

Environmental factors can influence the inheritance of a disease either by causing incomplete penetrance or due to environmental heterogeneity. Incomplete penetrance of a single gene disorder is observed when exposure to specific environmental factors, as well as having the genetic mutation, is required for the disease to manifest. In genetic diseases with incomplete penetrance, the mode of inheritance is difficult to deduce because all individuals a pedigree with the disease mutation cannot be

identified. Environmental heterogeneity occurs when the same disease phenotype can be caused either solely by environmental factors or solely by genetic factors. In genetic diseases with environmental heterogeneity, the mode of inheritance is difficult to deduce because pedigrees with the environmental disease are included in the segregation analysis (Section 1.3.1.1.) causing inconsistency in the mode of inheritance assumed from the pedigrees with the genetic disease.

## 1.3.1.1. Segregation analysis

Segregation analysis is the easiest way to uncover whether a single gene causes a disease, especially when its mode of inheritance is cloaked by incomplete penetrance. Segregation analysis is performed using software that analyses several affected pedigrees simultaneously looking for models of disease inheritance which best fit the observed transmission of disease. Segregation analysis was conducted across multiple extended pedigrees containing IBD patients to try to determine whether IBD is a monogenic trait and if so, what mode of inheritance best fit the patterns of segregation. The studies concluded that one mode of inheritance could not account for all cases of IBD. However, segregation patterns did indicate the following possibilities: 9 to 13 percent of UC patients carry a dominant disease gene with incomplete penetrance; a recessive gene of incomplete penetrance is responsible for all cases of CD; and 7 percent of CD patients carry a recessive gene with complete penetrance (Kuster et al., 1989; Orholm et al., 1993). Overall, segregation analysis on IBD affected families indicate that no simple Mendelian models fit IBD as a whole and that within the disease types, UC and CD, more than one gene may play a role. The studies concluded that IBD is either a polygenic disease or several Mendelian diseases masked by genetic and/or environmental heterogeneity.

## 1.3.1.2. Genetic heterogeneity

Genetic heterogeneity occurs when several genetic diseases are grouped together because phenotypically they appear identical or very similar. With approximately 11 percent of UC cases and 7 percent of CD cases showing a Mendelian pattern of inheritance, it is frequently argued that IBD may actually be several monogenic diseases that are phenotypically similar and therefore mistakenly grouped as one. The argument is supported by epidemiological studies previously discussed that show concordance between disease location and extent within families but not between families diagnosed with the same disease type.

## 1.3.2. Polygenic

A polygenic disease is defined as a genetic disease where mutations in two or more genes are required for the disease to manifest. One example of this would be two genes that serve the same function and do not have a dose–dependent phenotype. One gene would compensate a loss of function of the other gene. Mutations would have to occur in both genes for a new phenotype to be observed. Several possible polygenic disease scenarios with combinations of loss and gain of function plus increases and decreases in transcription are possible making the number of genes involved impossible to assume. The polygenic model becomes even more complicated when different mutations in and combinations of disease genes have different affects on disease aetiology. In these cases, the same mutations and genes may not be involved in each diseased individual and thus the gene is classified as a susceptibility locus (Section 1.3.2.1.) rather than a disease gene.

### 1.3.2.1. Susceptibility loci

Susceptibility loci are genes containing polymorphisms where different combination of alleles between loci may increase or decrease a person's susceptibility to a particular disease. If a person has a particular combination of susceptibility alleles from multiple susceptibility loci and encountered any required environmental insults then the disease will manifest itself phenotypically. It is hypothesised that different combinations of susceptibility alleles and loci can cause the same disease and that within the disease, minor phenotypical differences and even major phenotypic division, as in CD and UC for IBD, may be the result of these different combinations.

### 1.3.3. IBD genetic model

The polygenic disease model can resolve the apparent conflicts between resulting data from IBD epidemiological studies. However, the conflicts can also be resolved by describing IBD as several monogenic diseases with genetic heterogeneity or one monogenic disease with environmental heterogeneity. At present it is impossible to distinguish between several single gene disorders with genetic heterogeneity and a polygenic disease with susceptibility loci. For IBD, the segregation analysis supports the existence of a couple of simple Mendelian disease genes masked by genetic heterogeneity. However, the current belief of most IBD geneticists is that the majority of IBD is caused by a complex interaction of susceptibility loci.

## 1.4. Linkage

One method to try and identify chromosomal regions where genes causing susceptibility to a disease are located is to conduct genome wide linkage studies in patient cohorts. In linkage studies, polymorphic markers are analysed across the genome in affected cohorts, to identify specific chromosomal regions where the increase in sharing is

statistically higher than would be expected under the null hypothesis of independent assortment. The hypothesis of linkage maintains that relatives affected with the same disease will share both a disease causing gene and 'linked' specific alleles from adjacent polymorphic markers.

Most linkage analysis programs typically express results as a logarithm of the odds (LOD) score, a log-likelihood ratio of the results under the hypothesis of excess sharing compared to the hypothesis of no excess sharing, where the higher the LOD score the more significant the result. For example, a LOD score of 2.0 means that the genotype results are 100 times more likely to be caused by the linkage hypothesis over the null hypothesis of independent assortment. Linkage results can also be expressed as a p-value, a pointwise probability of observing the deviation under the null hypothesis. The smaller the p-value, the more significant the result. Continuing the example above, a LOD score of 2.0 will have a corresponding p-value of approximately 0.001, which means that there is one in a thousand chance of resulting in a LOD score of 2.0 under the null hypothesis at that specific location.

Universal rules for reporting and confirming linkage results have been proposed by Lander and Kruglyak (1995). When reporting linkage using these guidelines, three categories of linkage exist, suggestive, significant and highly significant. Each category relates to how frequently per genome scan, observed results would be expected by chance. Suggestive linkage occurs one time at random per genome scan, significant linkage 0.05 times and highly significant linkage only 0.001 times. The categories correspond to LOD scores of 2.2 (p = $7x10^{-4}$), 3.6 (p = $2x10^{-5}$) and 5.4 (p = $3x10^{-7}$) respectively, when conducting a human genome scan on affected sibling pairs (ASPs). The guidelines also state that following a report of significant or highly significant

linkage, a second independent study in a separate patient cohort needs to replicate the findings with a nominal p-value of 0.01 or lower, before a putative linkage region is confirmed as credible.

The following section will briefly describe the different linkage analysis methods and summarise the IBD linkage studies to date.

### 1.4.1. Linkage analysis methods

When preparing to undertake a linkage study, it is important that an appropriate population is collected for analysis and that the appropriate type of analysis and parameters are selected. This section discusses the different choices available and when they are typically used.

1.4.1.1. Population choice

Depending upon the type of genetic disease being investigated, different types of populations are collected for linkage analysis. When trying to find a simple Mendelian disease gene, large extended families with several affected individuals are preferable for linkage analysis. For a dominant disease, every affected individual within an extended pedigree will share one chromosomal region that is not present in any of the unaffected members of the pedigree (taking into account age of incidence and penetrance). For a recessive disease, every affected individual will share two chromosomes in the disease gene region and unaffected individuals will only share one or none. Potentially, one large family is sufficient to locate the region where a Mendelian disease gene is located. However, due to the lack of informative recombinants within one pedigree and the lack of affected individuals in some recessive traits, typically more then one family is required to refine the interval to a manageable size for positional cloning.

In theory, almost any population containing a few pedigrees with at least two affected individuals can be used to find a monogenic disease. However, when the aim of a study is to locate genes involved in complex, polygenic diseases such as IBD, many pedigrees containing affected relative pairs (ARPs) are required. Within a pedigree containing multiple affected individuals several disease susceptibility genes could be segregating. Different combinations of disease genes could be causing the same disease in different individuals. Thus, to provide sufficient power to detect linkage to a gene in the presence of interference from affected individuals with other disease genes, multiple pairs of affected individuals sharing the same gene are required. In complex disease it is typically hard to find extended pedigrees containing numerous affected individuals. Yet, due to the increased first-degree relative risk for most genetic disease it is relatively quick and easy to collect nuclear families containing ASPs. Thus, ASP populations have become the population of choice for complex polygenic linkage studies.

Still, some scientists have argued that collection of only ASPs limits the power of the study and ignores data that may be easy to collect and generate from other types of ARPs. Collecting and analysing all available affected individuals from every pedigree increases the power of the study. However, if the appropriate weighting or correction is not applied, large pedigrees with several ARPs can further mask the information from smaller single ASP pedigrees and overestimate the effect caused by genes within the extended pedigrees. Thus, it is more important to collect a linkage population with pedigrees of uniform number and type of ARPs rather than worrying about whether to collect all ARPs or just ASPs.

1.4.1.2. Parametric vs. non-parametric analysis

To perform linkage analysis on genotype data, the appropriate set of analysis conditions needs to be selected. Parametric analysis allows for inputs such as disease mode of inheritance and penetrance to be set prior to analysis. By informing the analysis program about the disease, it weights certain results more than it would without parameters. For example, if the disease were a completely penetrant recessive disorder then the program would increase the significance of regions where two affected siblings share two chromosomes. Thus, for single gene disorders with known modes of inheritance and penetrance values parametric analysis is better for detecting linkage. However, for complex, polygenic disease where the number of genes involved and their associated modes of inheritance and penetrance values are not known, non-parametric analysis is better for detecting valid regions of linkage. Still, some scientists prefer to use several types of parametric analysis rather than non-parametric analysis when analysing polygenic disease. However, if positive linkage is detected after several types of parametric analysis are conducted, it is not known whether the result is actually real and the parameters chosen appropriate or whether it is a false positive generated from multiple testing and chance selection of parameters coinciding with the data inputted.

1.4.1.3. Two-point vs. multipoint analysis

In the past, the only analysis available for scientists conducting linkage studies was two-point analysis, where each marker genotyped in a genome scan is analysed for linkage individually. However, today scientists also have the choice of using multipoint analysis to analyse genotype data. Multipoint analysis uses the data from all markers and the distances between them to calculate the linkage across the area between the markers rather than solely at single marker loci. The results of multipoint analysis of adjacent markers are usually consistent because inferences at one marker are normally

verified with data from adjacent markers. In contrast, the LOD scores from two-point

analysis of neighbouring markers can vary greatly due to unverified inferences of

sharing status. Only when both parents are heterozygous for different alleles is a family

fully informative and no assumptions made about sharing status. Varying

heterozygosities of markers lead to different proportions of families fully informative

for each marker. Thus, analysing data from markers with varying heterozygosities will

give varying results, with the results from more heterozygous markers being more

accurate. Furthermore, since different markers are informative in different families the

varying two-point results from adjacent markers can be even further accentuated.

Consequently, with the recent availability of multipoint analysis, the use of two-point

analysis has decreased sharply.


## 1.4.2. Previous IBD genome scans

Five years ago, the first genome scan across an IBD population was published. Since

then, six more IBD genome scans have been published. Of the seven scans, three

consisted of only CD families (Duerr *et al.*, 2000; Hugot *et al.*, 1996; Ma *et al.*, 1999)

and two others mostly CD and mixed families (ASPs containing one CD patient and one

UC patient) (Cho *et al.*, 1998; Rioux *et al.*, 2000). All scans covered all the autosomes

but only three included the X chromosome (Duerr *et al.*, 2000; Hampe *et al.*, 1999a;

Rioux *et al.*, 2000). Different numbers of families and ASPs were scanned in each

linkage study and therefore each study had a different power to detect linkage. The

power to detect linkage regions, and also to exclude other chromosomal regions,

increases with sample size. The largest genome scan for IBD susceptibility loci

contained 353 ASPs (Hampe *et al.*, 1999a). The other studies contained 183 ASPs

(Rioux *et al.*, 2000), 151 ASPs (Cho *et al.*, 1998), 94 ASPs (Duerr *et al.*, 2000), 89

ASPs (Satsangi *et al.*, 1996b), 65 ASPs (Ma *et al.*, 1999) and 41 ASPs (Hugot *et al.*, 1996).

### 1.4.3. Putative IBD susceptibility regions

The putative susceptibility regions resulting from the IBD genome scans and subsequent

IBD replication and extension studies, are summarised, chromosome by chromosome,

in this section (Figure 1.1. and Table 1.1.). The linkage results are categorised

according to the guidelines proposed by Lander and Kruglyak (Section 1.4.).

#### 1.4.3.1. Chromosome 1

Suggestive linkage was reported at D1S236 (1p) for CD with a p-value of 0.0006 using

singlepoint analysis (Hugot *et al.*, 1996). However, in an extension study by the same

group, linkage to CD was not detected at D1S236 when the number of ASPs was

increased from 41 to 112. Even so, based upon the initial result, another group tested

the same marker in 84 UC relative pairs to see whether the region was potentially linked

to IBD overall rather then solely to CD (Mirza *et al.*, 1998), but linkage to the region

was not detected. To date, significant linkage between any form of IBD and this region

has yet to be reported.

In a later genome scan, suggestive linkage between IBD and chromosome 1 was again

detected. However, this linkage was to a more p-terminal marker, D1S552, with a

multipoint LOD score of 2.7 (Cho *et al.*, 1998). Linkage between this region and IBD

has yet to be extended to the level of significant.

#### 1.4.3.2. Chromosome 3

Suggestive linkage was reported at D3S1573 (3p) with a singlepoint LOD score of 2.7
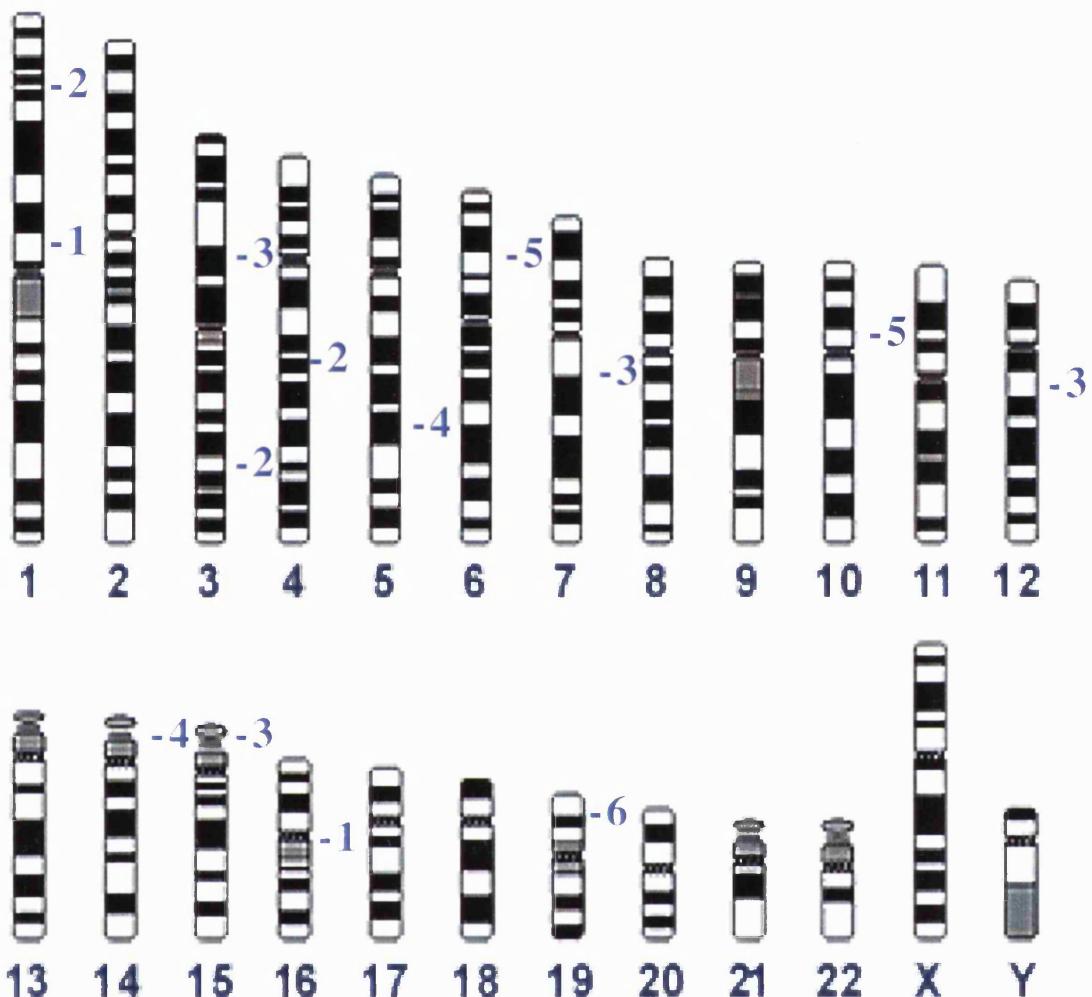
**Figure 1.1. IBD linkage regions from previous genome scans**

The first genome scan to report linkage for each region is indicated by the number next to the region. 1 - Hugot et al, 2 - Cho et al, 3 - Satangi et al, 4 - Ma et al, 5 - Hampe et al, 6 - Rioux et al.

IBD

| study | ASP | 1pter | 1p | 3p | 3q | 4q | 5q | 6p | 7q | 10p | 12q | 14p | 15p | 16q | 19p |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hampe (E) | 428 | | | | | | | 4.2 | | | | | | | |
| Hampe (GS) | 353 | N | N | N | N | N | N | 2.1 | N | N | 1.6 | N | N | N | N |
| Rioux (R) | 200 | | | N | | | | | N | | N | | | N | |
| Rioux (E) | 199 | | | | | 2.4 | | | | | | | | | |
| Satsangi (E) | 186 | | | 2.7T | | | | N | 3.1T | | 5.5T | | N | N | |
| Rioux (GS) | 183 | N | N | 2.4 | N | N | 3 | 2.3 | N | N | N | N | N | N | 4.6 |
| Cho (GS) | 151 | N | 2.7 | N | 2.3 | 2.7 | N | N | N | N | N | N | N | N | N |
| Ohmen (R) | 117 | | | | | | | | | | | | | N | |
| Duerr (R) | 115 | | | | | | | | | | 2.8 | | | | |
| Hampe (CG) | 100 | | | | | | | | | | N | | | | |
| Satsangi (GS) | 89 | N | N | 0.001T | N | N | N | N | 0.0002T | N | 0.002T | N | 0.0004T | N | N |
| Vermeire (R) | 79 | | | N | | | | | N | | N | | | N | |
| Annese (R) | 54 | | | N | | | | N | N | | N | | | 0.003 | |

CD

| study | ASP | 1pter | 1p | 3p | 3q | 4q | 5q | 6p | 7q | 10p | 12q | 14p | 15p | 16q | 19p |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hampe (GS) | 162 | N | N | N | N | N | N | N | N | 2.3 | 1.8 | N | N | 1.7 | N |
| Rioux (E) | 139 | | | | | | 3 | | | | | | | | |
| Lesage | 128 | | | | | | | | | | N | | | | |
| Rioux (GS) | 124 | N | N | N | N | N | | N | N | N | N | N | | N | 3 |
| Rioux (R) | 114 | | | N | | | | | N | | N | | | N | |
| Hugot (E) | 112 | N | | | | | | | | | | | | 0.000015 | |
| Cho (GS) | 97 | N | N | N | N | N | N | N | N | N | N | N | N | 1.8 | N |
| Duerr (GS) | 94 | N | N | N | N | N | N | N | N | N | N | 3.6 | N | N | N |
| Duerr (R) | 84 | | | | | | | | | | 1.8 | | | | |
| Satsangi (E) | 81 | | | 0.0002T | | | | N | 0.0002T | | 0.0003T | | N | 0.004T | |
| Brant (R) | 78 | | | N | | | | | N | | N | | | 0.007T | |
| Ohmen (R) | 75 | | | | | | | | | | | | | 1.7 | |
| Cavanaugh (R) | 72 | | | | | | | | | | | | | 6.3 | |
| Yang (CG) | 70 | | | | | | 0.002T | | | | | | | | |
| Ma (GS) | 65 | N | N | N | N | N | N | N | N | N | 2 | 2.8 | N | N | N |
| Vermeire | 61 | | | N | | | | | N | | N | | | N | |
| Brant (R) | 50(nj) | | | N | | | | | N | | N | | | N | |
| Ohmen (R) | 50(nj) | | | | | | | | | | | | | 2.4 | |
| Rioux (E) | 50(16) | | | | | | 3.9 | | | | | | | | |
| Brant (R) | 50(s) | | | | | | | | | | | | | 3.84 | |
| Rioux (GS) | 46(16) | | | | | | | | | | | | | | 1.5 |
| Satsangi (GS) | 44 | N | N | 0.003T | N | N | N | N | 0.003T | N | N | N | 0.0004T | N | N |
| Hugot (GS) | 41 | 0.0006T | N | N | N | N | N | N | N | N | N | N | N | 0.001T | N |
| Ma (GS) | 39(nj) | N | N | N | N | N | N | 1.7 | N | N | N | N | N | N | N |
| Hampe (CG) | 29 | | | | | | | | | | N | | | | |
| Brant (R) | 28(j) | | | | | | | | | | | | | 0.007T | |
| Ma (GS) | 26(j) | N | N | N | N | N | 2.2 | N | N | N | 1.7 | 2 | N | N | N |
| Annese (R) | 20 | | | N | | | | | N | | N | | | | |

UC

| study | ASP | 1pter | 1p | 3p | 3q | 4q | 5q | 6p | 7q | 10p | 12q | 14p | 15p | 16q | 19p |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hampe (GS) | 114 | N | N | N | N | 1.6 | N | N | N | N | N | N | N | N | N |
| Satsangi (E) | 64 | | | N | | | | N | 0.00006T | | 0.003T | | N | N | |
| Hampe (CG) | 46 | | | | | | | | | | N | | | | |
| Rioux (R) | 36 | | | N | | | | | N | | N | | | N | |
| Satsangi (GS) | 30 | N | N | N | N | N | N | N | 0.00006T | N | 0.0002T | N | N | N | N |
| Mirza (R) | 30 | | | | | | | | | | | | | N | |
| Ohmen (R) | 23 | | | | | | | | | | | | | N | |
| Rioux (GS) | 20 | N | N | N | N | N | N | N | N | N | N | N | N | N | 2.9 |
| Annese (R) | 19 | | | N | | | | N | N | | N | | | 0.007 | |
| Cho (GS) | 18 | N | N | N | N | N | N | N | N | N | N | N | N | N | N |
| Duerr (R) | 15 | | | | | | | | | | 1.8 | | | | |

**Table 1.1.  Comparison of IBD linkage studies by power**

Backgrounds of red, blue, and green surround significant, suggestive, and replicative linkage results, respectively (Lander and Kruglyak, 1995).  LOD scores are > 1 and p-values are < 1.  N means no linkage detected above a LOD of 1.5.  T stands for two-point analysis.  NJ, J, S and 16 stand for non-Jewish, Jewish, severe and early onset, respectively.  Genome scan studies are shown in blue font and with a GS.  R, E and CG stand for replication, extension, and candidate gene studies, respectively.

(p = 0.0002) in IBD (Satsangi *et al.*, 1996b). The linkage was detected in an extension

study (186 ASPs) carried out on loci from the group's genome scan that resulted in

p-values of 0.001 or less. Following the initial report of linkage, four groups tested the

same region for linkage in their IBD populations (Annese *et al.*, 1999; Brant *et al.*,

1998; Rioux *et al.*, 1998; Vermeire *et al.*, 2000). Linkage to IBD was not detected in

these studies. However, subsequently one of the groups published a genome scan using

a subset of their IBD population which, did result in suggestive linkage in the 3p region

with a multipoint LOD of 2.4 between D3S1766 and D3S1285 (Rioux *et al.*, 2000).

Still, a study has yet to detect significant linkage to this region.


In another genome scan, suggestive linkage was also reported between IBD and

chromosome 3, but this time on the q-arm between D3S3053 and D3S2427with a

multipoint LOD score of 2.3 (p = 0.0006) (Cho *et al.*, 1998). A subsequent study on the

q-arm loci has yet to be conducted.


## 1.4.3.3. Chromosome 4

At D4S2623 suggestive linkage was reported with a LOD of 2.76 (p = 0.0002) in a

subgroup of 96 mixed ARPs (Cho *et al.*, 1998). In a later genome scan, linkage above

the level needed to replicate significant linkage was reported in the same region for a

subgroup of 114 UC ASPs with a LOD of 1.56 (p = 0.009) between D4S1575 and

D4S424 (Hampe *et al.*, 1999a). Taking the two results together it appears that the

region may contain a gene causing susceptibility to UC only. The region has yet to be

extended to the level of significant.

## 1.4.3.4. Chromosome 5

Suggestive linkage was reported between D5S393 and D5S673 with a multipoint LOD

of 2.2 in a subgroup of Ashkenazi Jewish CD patients (26 ASPs) (Ma *et al.*, 1999).

Subsequently, another genome scan also resulted in suggestive linkage to the region

with a multipoint LOD of 3.0 between D5S816 and D5S1480 for IBD (Rioux *et al.*,

2000). However, following this genome scan the group investigated the region further

in an extension study with more markers and families (199 ASPs). This study resulted

in less suggestive linkage than the previous genome scan, with a LOD of 2.4 for IBD.

Yet, after phenotypic division, significant linkage was detected in the region with a

LOD of 3.9 around D5S2497 in 50 early age of onset CD families. A replication study

has yet to be conducted in this significant linkage region.


## 1.4.3.5. Chromosome 6

In a genome scan, suggestive linkage to IBD was almost detected between D6S289 and

D6S276 with a multipoint LOD score of 2.1 (p = 0.003) (Hampe *et al.*, 1999a). Prior to

the scan, the identical region was tested for linkage in two subgroups of the same

genome scan population and again linkage was not observed (Hugot *et al.*, 1994; Naom

*et al.*, 1996). However, following the genome scan result, more markers in the region

were genotyped in the same population and suggestive linkage with a multipoint LOD

score of 2.9 at D6S461 was detected (Hampe *et al.*, 1999b). Continuing the

investigation, the same group conducted an extension study in the region increasing the

number of IBD ASPs from 353 to 428. This study resulted in significant linkage to the

region with a LOD of 4.2 at D6S461. Concurrently, another group replicated the

linkage in a study of 70 CD ASPs with a singlepoint p-value of 0.002 with a short

tandem repeat (STR) within the tumour necrosis factor (TNF) gene (Yang *et al.*, 1999b).

Following these reports, a second genome scan also resulted in suggestive linkage in the

region, with a multipoint LOD of 2.3 for IBD between D6S1281 and D6s1019 (Rioux *et al.*, 2000). Consequently, with the independent replication of this significant linkage region, the scientific community designated the region the *IBD3* susceptibility locus.

### 1.4.3.6. Chromosome 7

At D7S669 suggestive linkage was reported with a singlepoint p-value of 0.0002 in IBD (Satsangi *et al.*, 1996b). Following this result an extension study containing 186 ASPs was conducted with the same marker in addition to adjacent markers. The study resulted in a more suggestive p-value of 0.00008 (LOD = 3.1) at D7S669, contributed mostly by their CD families (p = 0.0002). However, from the same study, one of the adjacent markers also resulted in suggestive linkage with a singlepoint LOD of 2.9 (p = 0.0001), but this time contributed mostly by their UC families (p = 0.00006). Combining these results, the linkage interval appears to be contributed by the IBD population as a whole. Following the publication of these results, four groups analysed the same region in their own cohorts but linkage was not detected (Annese *et al.*, 1999; Brant *et al.*, 1998; Rioux *et al.*, 1998; Vermeire *et al.*, 2000).

### 1.4.3.7. Chromosome 10

One genome scan reached suggestive linkage between D10S548 and D10S197 with a LOD of 2.3 (p = 0.001) in only their 162 CD ASPs (Hampe *et al.*, 1999a). Linkage was not observed in their IBD population overall or in their mixed or UC subdivisions. To date, further IBD studies on this putative linkage region have not been published.

### 1.4.3.8. Chromosome 12

Highly significant linkage with a singlepoint LOD of 5.5 (p = 0.0000003) at D12S83 was reported from an extension study, where the number of IBD ASPs increased from

89 to 186 (Satsangi *et al.*, 1996b). In the genome scan conducted by the same group, D12S83 plus adjacent markers had resulted in p-values less than 0.01, which was the criteria used to define regions that progressed to the second stage extension study. Since the publication, the region has been independently replicated in 208 IBD ARPs from 122 families with a LOD of 2.8 (p = 0.0002) between D12S1724 and D12S90 (Duerr *et al.*, 1998). Yet, this same group failed to see linkage in the region in their genome scan using a subset of 127 CD ARPs (Duerr *et al.*, 2000). Not surprisingly, when these two groups combined their data in a meta-analysis, again significant linkage was detected to IBD and UC with LODs of 5.2 and 3.9, respectively, but was not detected to CD (Parkes *et al.*, 2000). Concurrently, other groups also replicated and broadened the linkage to this region with LOD scores of 1.8 (p = 0.005) between D12S303 and D12S326 (Curran *et al.*, 1998) and 2.0 between D12S345 and D12S85 (Ma *et al.*, 1999; Yang *et al.*, 1999a), both only in their CD populations. However, six other groups failed to replicate the linkage in their IBD populations (Annese *et al.*, 1999; Brant *et al.*, 1998; Hampe *et al.*, 1998; Lesage *et al.*, 2000; Rioux *et al.*, 1998; Vermeire *et al.*, 2000). Despite these negative data, with multiple independent replications of the initial highly significant linkage, the scientific community has labelled this region as the *IBD2* susceptibility locus.

### 1.4.3.9. Chromosome 14

Suggestive linkage with a multipoint LOD score of 2.8 resulted between D14S582 and D14S72 from a genome scan comprised of only CD families (Ma *et al.*, 1999). Previously in another genome scan, the same region resulted in linkage at the level needed to replicate significant linkage, with a LOD of 1.53 (p = 0.004) at D14S608 in a subgroup of 96 mixed ARPs (Cho *et al.*, 1998). Following these reports of linkage, a third genome scan extended the level of linkage in the region to significant with a

multipoint LOD of 3.6 at D14S261 for CD (Duerr *et al.*, 2000). A region specific

replication study has yet to be conducted in this IBD linkage region.


## 1.4.3.10. Chromosome 15

Suggestive linkage was achieved with a singlepoint p-value of 0.0004 at D15S128

contributed predominately by CD families (p = 0.0004) (Satsangi *et al.*, 1996b).

However, in a subsequent extension study by the same group linkage was not detected.

Other IBD linkage studies have yet to be conducted in this region.


## 1.4.3.11. Chromosome 16

Significant linkage between D16S409 and D16S419 (p = 0.000015) resulted from an

extension study where 112 CD ASPs, containing the original 41 ASPs, were genotyped

with STR markers around and including D16S408 (Hugot *et al.*, 1996). The extension

study was conducted on chromosomal intervals where a marker from the genome scan,

like D16S408 (p = 0.001), resulted in a singlepoint p-value less than or equal to 0.001.

However in a later extension study which used the same families in addition to others

recruited since (167 CD ASPs total), the linkage fell below the level of suggestive with

a multipoint LOD of 1.7 (p = 0.005) between D16S409 and D16S411 (Curran *et al.*,

1998). Nonetheless, linkage to the region continues to be supported with six other

groups replicating it in their populations. The strongest of these replication studies

achieved highly significant linkage with a multipoint LOD of 6.3 between D16S409 and

D16S756 in 72 CD ASPs (Cavanaugh *et al.*, 1998). The other five studies achieved a

multipoint LOD of 3.84 in 50 'severe' CD ASPs (Brant *et al.*, 2000), a multipoint LOD

of 2.41 at D16S411 in 50 CD ASPs (Ohmen *et al.*, 1996), a single point p-value of

0.004 at D16S407 in 81 CD ASPs (Satsangi *et al.*, 1996b), a multipoint p-value of 0.007

at D16S769 in 148 CD relative pairs (Brant *et al.*, 1998), and a multipoint p-value of

0.003 at D16S408 in 73 IBD relative pairs (Annese *et al.*, 1999). Since, the last group

was the only group to report positive linkage in the interval to phenotypes other than

just CD, the region may contain a CD specific susceptibility gene. However, three other

groups failed to replicate the linkage, with 83 UC ASPs (Mirza *et al.*, 1998), 79 IBD

ASPs (68 CD) (Vermeire *et al.*, 2000) and 200 IBD ASPs (114 CD) (Rioux *et al.*,

1998). Even so, the studies have resulted in sufficient data for the scientific community

to designate the susceptibility region as the *IBD1* locus.


## 1.4.3.12. Chromosome 19

Significant linkage was reported for IBD with a multipoint LOD of 4.6 between

D19S591 and GATA21G05 (Rioux *et al.*, 2000). The linkage was shared equally

between both the CD and UC families with multipoint LOD scores of 3.0 and 2.9

respectively. A replication study has yet to be published for this region.


## 1.4.3.13. Linkage results between a LOD of 1.5 and 2.2

Linkage results greater than the minimum level needed to replicate a region but less

than the minimum level needed to declare suggestive linkage were reported in IBD

ASPs overall at D1S2670-D1S2682, D2S142, D22S315-D22S421, and DXS1202-

DXS1214; in CD ASPs at D1S1597, D2S2952-D2S1400, D8S256. D9S2157,

D11S1999, D16S748-D16S764, D16S516, D19S591, and D19S1034-D19S586; in

Ashkenazi Jewish CD ASPs at D17S925-D17S787, D17S784, and D18S474; in UC

ASPs at DXS1001-DXS1047; and in mixed ASPs at D1S1609 (Cho *et al.*, 1998; Duerr

*et al.*, 2000; Hampe *et al.*, 1999a; Ma *et al.*, 1999; Satsangi *et al.*, 1996b).

## 1.5. Association

Following the identification of a susceptibility locus, association analysis is usually conducted to further refine the genetic interval where a disease mutation may reside. Association analysis is used to detect linkage disequilibrium (LD) between alleles from polymorphic markers and a disease mutation. When a mutation occurs on a chromosome, it will initially be in complete LD with an allele of every polymorphic marker along the chromosome. It is hypothesised that over time polymorphic markers nearer a disease mutation should have alleles in greater LD with the disease mutation and thus exhibit greater positive association with the disease. The hypothesis is based on the fact that fewer recombination events, which breakdown LD, will occur between a disease mutation and a marker the smaller the distance between them provided that the alleles from the surrounding markers existed prior to the mutation.

This section will cover association analysis methods and results from previous published IBD association studies.

### 1.5.1. Association analysis methods

When planning to conduct an association study it is important to determine a number of criteria, for example: what type of population to analyse, which type of association test to perform and which markers to genotype. The choice of population has a direct affect on the choice of association test and vice versa. The focus of the study has a direct impact on the markers that should be genotyped. The different types of association tests and studies are discussed below.

## 1.5.1.1. TDT vs. case/control

There are two types of association tests, a transmission disequilibrium test (TDT) and a case/control test. For both tests, the results are expressed in the form of a p-value (Section 1.4), with positive association resulting from p-values equal to or less than 0.05. Both tests are a form of chi-square analysis used to highlight excess transmission of a polymorphic marker allele within a disease population. In a case/control test excess transmission of an allele is detected by comparing the allele frequencies within a disease population (cases) to the frequencies within a matched control population. The further an allele frequency in the cases differs from the corresponding frequency in the controls, the smaller the resulting p-value and larger the positive association. In a TDT, excess transmission of an allele is detected by comparing the number of transmissions of an allele between parent and offspring to the number expected by independent assortment. If a parent were heterozygous for alleles of a polymorphic marker, then any offspring would have a 50% chance of receiving each allele. Thus, within a random population of simplex families (parents and a proband) the genotyping of each marker allele should result in equal numbers of transmissions and non-transmissions. However, when genotyping a simplex population where all probands have the same disease, alleles in disequilibrium with a disease mutation would be transmitted more frequently and therefore result in positive disease association.

When comparing the two types of association tests it is not apparent which is better since both tests have advantages and disadvantages. The results from a case/control test can suffer from selection bias if cases and controls are not accurately matched for age, sex, geographic origin, and most of all ethnicity. This bias then leads to a bias in the allele frequency between the two populations. For example, if the disease of interest is typified by a late age of onset and if the controls are composed of people below that age,

then some of the controls may actually become affected in the future. Therefore, their

inclusion in the control population could bias the allele frequency. The TDT is free

from selection bias since it utilises the theory of independent assortment as the control.

Another disadvantage to conducting a case/control test is the inability to generate

haplotypes, which may be useful for detecting association in a complex disease.

Without parental genotypes, haplotypes can only be generated by algorithms that are

based upon assumptions that may be invalid. However, the advantage of the

case/control test is that, with equal size populations, it has more power to detect

association than the TDT. In a case/control test all the individuals who are genotyped

can contribute to the analysis. In a TDT only when a parent is heterozygous is a family

useful in the analysis. Overall, if the disease of interest is not prevalent then it is better

to collect simplex families for TDT analysis since the size of the cohort collected for the

study will be small and thus easily influenced by selection bias. Yet, if the disease is

prevalent, a case/control study may be better since it offers more power to detect

association and its main disadvantages become less significant as the size of the case

and control population increases.


## 1.5.1.2. Candidate genes vs. linkage regions

Genetic association tests are normally employed across either candidate disease genes

or disease specific linkage regions, both of which can be used to locate a disease

mutation. With the candidate gene approach, information and assumptions about the

disease aetiology is used to search the database for genes that, if mutated, may

contribute to the disease of interest. Polymorphic markers are located within and

around the candidate genes and subsequently used for association analysis. With the

linkage region approach, bias caused by selection of only genes of known function is

obviated. Instead, association analysis is conducted on evenly spaced polymorphic

markers spanning a disease linkage region. If association is detected with a marker, then bioinformatic information is investigated and experiments conducted to locate genes around the marker. Only then are specific genes analysed for positive association with the disease phenotype. Both approaches offer obvious advantages and disadvantages. However, the favoured study design combines both strategies, where candidate genes within a linkage region are analysed for association. With this positional candidate design, 'favourite gene' bias is diminished while insights from previous scientific findings are used to prioritise the association studies across the region.

### 1.5.2. Results from previous IBD association studies

Numerous association studies have been conducted on IBD cohorts. Presently, due to the relative novelty of the TDT test, combined with the relative ease of collecting cases, the majority of these studies used the case/control test. Some of these studies tested for association between IBD and alleles from polymorphic markers within candidate genes, as described above. However, the remainder tested for association between IBD and either environmental or clinical factors. Results from both types of studies will be presented in this section.

1.5.2.1. Environmental and clinical associations with IBD

Valid associations between a disease and both environmental and clinical factors are useful when conducting genetic linkage and associations studies for a complex disease. Difficulty in locating disease genes increases with genetic heterogeneity (Section 1.3.1.2.). Environmental and clinical associations allow phenotypic divisions to be made within disease cohorts, possibly decreasing heterogeneity and enabling masked genetic linkage and association to be detected.

Several studies have reported positive association between IBD and both environmental

and clinical factors. Repeatedly, association has been reported with perinuclear

antineutrophil cytoplasmic antibodies (P-ANCA). The association is mainly between

increased levels of P-ANCA and UC (Kim *et al.*, 1995; Seibold *et al.*, 1994; Yang *et al.*,

1995). CD is only associated with P-ANCA in patients exhibiting left-sided colitis and

inflammation, which has been considered a 'UC-like' phenotype (Vasiliauskas *et al.*,

1996). Elevation in P-ANCA levels appears to be caused by UC rather than actually a

cause of UC, since unaffected relatives of UC patients do not show association for

elevated levels of P-ANCA (Bansi *et al.*, 1996; Papo *et al.*, 1996). Another association

has been observed between IBD and smoking. CD is associated with increased levels of

smoking while UC is associated with decreased levels (Lee and Lennard-Jones, 1996;

Reif *et al.*, 2000; Tysk *et al.*, 1988). Positive associations are also reported between

IBD and Turner's syndrome, suggesting the possible existence of a susceptibility locus

on the X chromosome (Hayward *et al.*, 1996). Other, weaker, associations have been

reported between IBD and intestinal permeability (Hollander *et al.*, 1986) and between

IBD and measles infection and vaccination (Pardi *et al.*, 2000). However, these two

association have not been replicated in other studies (Feeney *et al.*, 1997; Munkholm *et

al.*, 1994).


1.5.2.2. Genetic associations with IBD

Presently, all IBD genetic association studies have been candidate gene studies. The

majority of these studies analysed genes of the auto-immune response and inflammatory

pathways, including genes from the human leukocyte antigen (HLA) -class II region,

TNFα, lymphotoxin alpha (LTα), interleukin (IL) pathways, natural resistance

associated macrophage proteins (NRAMPs) and bradykinin receptor B1 (BDKRB1).

Other genes selected for analysis included genes involved in intestinal permeability and cancer. These associations will be discussed in more detail in the following section.

### 1.5.2.2.1. HLA class II region

Of the IBD association studies conducted on genes within the HLA class II region on chromosome 6, most studies resulted in positive association with either an allele from a marker within one gene or with a haplotype of alleles from several markers across the region (Bouma et al., 1997; Bouma et al., 1998; Danze et al., 1996; De La Concha et al., 1997; Fernandez Arquero et al., 1998; Folwaczny et al., 1999; Forcione et al., 1996; Heresbach et al., 1996; Hesresbach et al., 1996; Reinshagen et al., 1996; Satsangi et al., 1996c; Trachtenberg et al., 2000; Uyar et al., 1998; Yoshitake et al., 1999). Only two studies reported no evidence of association within their CD populations (Satsangi et al., 1996c; Trachtenberg et al., 2000) and two others within their UC populations (Folwaczny et al., 1999; Forcione et al., 1996). The remaining studies reported positive association within the HLA class II region across all populations tested. The strongest and most consistent association was observed between IBD and alleles from markers within DRB1, with p-values from 0.5 to $1\times10^{-8}$. Analysis of markers within other genes in the region resulted in fewer and less consistent positive associations.

### 1.5.2.2.2. TNFα and LTα

The IBD association studies across TNFα and LTα resulted in positive association with TNFα. Association was not seen in any population with markers from LTα (Bouma et al., 1996; Hampe et al., 1999b). Of the TNFα studies, three studies reported association in CD and another in UC, with p-values between 0.04 and 0.01 for CD and 0.04 for UC (Bouma et al., 1996; Louis et al., 1996; Negoro et al., 1999; Plevy et al., 1996). However, the positive CD studies failed to detect association in UC and vice versa.

Also, two other TNFα IBD association studies resulted in no observed associations (Hampe *et al.*, 1999b; Mansfield *et al.*, 1994). Thus, the IBD association results across this region on chromosome 6 appear less consistent than the associations observed across the more p-telomeric HLA class II region.


1.5.2.2.3. Interleukin pathways

Association studies between IBD and genes involved in IL pathways have been conducted on IL1A, IL1B, IL1RA, IL1R, IL4R, IL4, IL6, IL10 and interferon gamma (INFG). The majority of studies analysed markers within IL1B and IL1RA on chromosome 2, testing for association with single alleles and with haplotypes of alleles from markers spanning both genes (Bioque *et al.*, 1996; Bioque *et al.*, 1995; Carter *et al.*, 2001; Hacker *et al.*, 1998; Hacker *et al.*, 1997; Heresbach *et al.*, 1997; Louis *et al.*, 1996; Mansfield *et al.*, 1994; Mwantembe *et al.*, 2001; Nemetz *et al.*, 1999; Stokkers *et al.*, 1998; Tountas *et al.*, 1999). Positive association was reported between three UC populations and markers within IL1RA, with p-values of 0.04, 0.01 and 0.007 (Carter *et al.*, 2001; Mansfield *et al.*, 1994; Tountas *et al.*, 1999). However, no association was observed between markers within IL1RA and seven other UC populations. In addition, no association was observed between IL1RA and both CD and IBD. Only one of the twelve studies observed positive association between IL1B and IBD (p = 0.00001) (Mwantembe *et al.*, 2001). Haplotype association analysis with markers spanning both genes resulted in positive association with both UC (p = 0.02) and CD (p = 0.01), again with at least one other study reporting no association in a corresponding population (Bioque *et al.*, 1995; Nemetz *et al.*, 1999; Stokkers *et al.*, 1998). Two association studies were conducted on IL4 and IL4R (Aithal *et al.*, 2001; Olavesen *et al.*, 2000). One study resulted in positive allelic association between IL4 and CD (p = 0.002) and haplotype association between markers in both IL4 and IL4R and CD (p = 0.005)

(Aithal *et al.*, 2001). No association was seen for the IBD or UC phenotypes. For IL10

on chromosome 1, positive allelic association was observed in both UC and IBD, with

p-values of 0.04 and 0.03 respectively (Tagore *et al.*, 1999). For IL6 on chromosome 7,

positive allelic association (p = 0.03) was also detected to IBD (Koss *et al.*, 2000). No

association was seen between IBD and IL1A on chromosome 2 (Mansfield *et al.*, 1994)

and INFG on chromosome 12 (Hampe *et al.*, 1998).


1.5.2.2.4. NRAMPs

Both NRAMP on chromosome 2 and NRAMP2 on chromosome 12 have been tested for

association with IBD. A haplotype from markers across NRAMP showed positive

association with CD (p = 0.001) but no association with UC or IBD combined

(Hofmeister *et al.*, 1997). No association was observed between NRAMP2 and IBD,

CD, or UC (Stokkers *et al.*, 2000).


1.5.2.2.5. Bradykinin receptor B1

Bradykinin receptor B1 (BDKRB1) on chromosome 14 is a gene involved in the

inflammation response pathway and was thus selected as a good candidate gene for an

association study with IBD. A single nucleotide polymorphism (SNP) within BDKRB1

was genotyped across 53 IBD patients and 110 unaffected controls. Positive association

was detected with a p-value of 0.0002 (Bachvarov *et al.*, 1998).


1.5.2.2.6. Mucins

The association studies conducted on genes involved in intestinal permeability included

MUC2 on chromosome 11 and MUC3 on chromosome 7. The MUC3 studies resulted

in positive association between a STR and UC (p = 0.0001) and a SNP and CD

(p = 0.01) (Kyo *et al.*, 2001; Kyo *et al.*, 1999). However, no association was detected in the study between MUC2 and UC (Swallow *et al.*, 1999).

### 1.5.2.2.7. MLH1

Alleles from markers within and around MLH1, a DNA repair gene, were analysed for association with IBD. MLH1 was selected as a candidate gene due to the 10-20 fold increase risk of intestinal cancer in people diagnosed with IBD (McConnell and Vadheim, 1992). Positive association was detected between a number of haplotypes and CD, UC and IBD combined with p-values from 0.04 to 0.002 (Pokorny *et al.*, 1997).

## 1.6. Aims and objectives

After reviewing the data presented in this chapter, it is apparent that, although several IBD linkage regions had been identified, further linkage and association analysis within the regions need to be conducted to extend or replicate previous findings and to identify novel genetic associations. To date, the majority of linkage and association studies were conducted using IBD cohorts with less than 200 probands. To better stratify linkage and association findings, studies need to be performed with IBD cohorts of greater power. High-throughput genotyping capabilities combined with the continuing collection of IBD families now allow for this type of powerful genetic study to be carried out in a much shorter time frame than before, using a larger number of samples and a higher degree of accuracy.

The experiments and analysis presented in this thesis further investigate previous linkage results and attempt to identify positive associations within both linkage regions and novel positional candidate genes. With collaborative access to an extended IBD

ASP population, a linkage study was performed with a dense set of STRs across the putative IBD susceptibility locus on chromosome 3p. The study aimed to extend the detected linkage in the region to the level of significant, using the criteria of Lander and Kruglyak (Section 1.4.). The genotype data resulting from the linkage study was further analysed for association with IBD to try and narrow the region where the putative IBD susceptibility gene may reside. Subsequently, to increase the power to detect association and remove any potential bias from the previous study, the identical set of STRs were genotyped across a large simplex population. In parallel, in an attempt to rapidly identify a disease gene, association studies on candidate genes within the region were performed using the same simplex population. The genes selected for analysis were chemokine receptor 2 (CCR2), CCR5, CCRL2, and guanine nucleotide-binding protein αi2 (GNAI2). Finally, due to negative results from the 3p linkage study and the recent identification of linkage between IBD and chromosome 6 by Hampe et al. (1999), a linkage study was conducted on chromosome 6 to try and replicate the *IBD3* linkage in an independent cohort of IBD ASPs.

# Chapter 2

# 2. Materials and Methods

## 2.1. Methods

### 2.1.1. General methods

2.1.1.1. Genomic DNA extraction from whole blood

Whole blood samples were stored at -80°C. Before extraction the samples were thawed

at 37°C in a water bath. To extract the genomic DNA, 3ml of whole blood were added

to 9ml of RBC lysis solution in a 14ml centrifuge tube. The mixture was inverted twice

while incubating for 10min at room temperature (RT). To separate the white cells from

other lysed material, the mixture was centrifuged in a RT-7™ from Sorvall Ltd. at

2,000X g for 10min. Most of the supernatant was removed leaving approximately

100μl of residual solution to resuspend the cells in by vigorously votexing. Following

resuspension 3ml of cell lysis solution were added and mixed by pipetting. To digest

any RNA, 15μl of RNase A solution were added to the lysate, which was then incubated

at 37°C for 35min. After RNA digestion, 1ml of protein precipitation solution was

added and the mixture was votexed for 20s. To separate the proteins from the DNA, the

mixture was again centrifuged at 2,000X g for 10min. The supernatant retaining the

DNA was then poured off into another 14ml centrifuge tube containing 3ml of 100%

isopropanol and mixed by inverting 50X. To separate the DNA from the aqueous

solution, the mixture was centrifuged at 2,000X g for 3min. The resulting supernatant

was removed and 3ml of 70% ethanol were added to remove any residual salt. The

mixture was centrifuged at 2,000X g for 1min, the ethanol was poured off, and the

resulting DNA pellet was allowed to dry at RT for 15min. To hydrate the DNA, 250μl

of DNA hydration solution were added and the mixture was incubated at 65°C for 1hr.

The solution was rocked at RT overnight allowing the DNA to equilibrate. The

resulting DNA samples were transferred to 1.5ml tubes and stored at -80°C.


2.1.1.2. DNA concentration quantification

The concentration of extracted genomic DNA was quantified using the Fluorocount™

from Packard Instruments Co. To quantify the DNA, 2μl of each DNA sample were

diluted with 998μl of DNA hydration solution. Standards of λ DNA were prepared in

1ml aliquots at concentrations of 0, 10, 20, 40, 60, 100, 400, 600, 800 and 1000ng/ml.

The DNA in each diluted sample and standard was labelled with fluorescent

PicoGreen® dye by adding 100μl from each into the appropriate well of a 96-well plate

containing 100μl of PicoGreen® dsDNA quantification reagent dilute 1/200 with DNA

hydration solution. The plate was centrifuged for 30s at 500rpm in a RT-7™ from

Sorvall Ltd. and allowed to incubate in the dark at RT for 2min. The plate was then

inserted into the Fluorocount™ and the fluorescent emissions from each well was

detected and recorded. A curve was calculated from the data recorded from each

standard, plotting the concentration of each standard against the amount of fluorescent

emission detected. The resulting curve was used to convert the amount of fluorescent

emission detected for each diluted sample into a DNA concentration. The concentration

of the stock tube was calculated by multiplying the concentration of the diluted DNA

aliquot by 500.

## 2.1.1.3. DNA dilution and storage

In the assigned well of a deep 96-well plate, 40µg of DNA were mixed with the required amount of water to bring the final volume to 1ml. This 40ng/µl plate was designated the 'stock plate'. To make the 4ng/µl 'working plate', 100ul from each well of the stock plate were added to the corresponding well in another plate containing 900ul of water. The stock plate was stored at -80°C and the working plate at -20°C.

## 2.1.1.4. PCR DNA plate dispensing and storage

Dispensing DNA from a working plate into several 96-well plates for subsequent polymerase chain reaction (PCR) amplification reactions was performed using the Hydra™ 96 from Robbins Scientific, Inc. The Hydra™ was programmed to dispense 5µl of DNA from each well of a 'working plate' into the corresponding wells of the required number of PCR plates (Figure 2.1.). Before dispensing, the Hydra™ was washed with water. Following dispensing the Hydra™ was sterilised by washing with water, with 2% chloros industrial, and again with water. The resulting PCR plates, containing 20ng of DNA in each well, were dried at 80°C for 1hr. The plates were stored at -20°C.

## 2.1.1.5. Oligonucleotide design

The sequences from which primers were designed were initially analysed for similarities to sequences within the ALU and NR databases at the National Center for Biotechnology Information (NCBI) Internet site. The sequences were submitted in FASTA format (Figure 2.2.). The sequences were compared using BLAST 2.1 and the BLASTn algorithm. Any regions of the sequences where an ALU repeat or discordant genomic DNA fragment aligned with an alignment score greater than 40 were excluded,

| | | |
|---|---|---|
| Dispense mode | - | DIS MODE |
| Dispense volume (μl) | - | DV 5.0 |
| Dispense Height | - | DH 4000 |
| Fill volume (μl) | - | FV X (X = N * 5) |
| Fill height | - | FH 4050 |
| Empty height | - | EH 4000 |
| Wash volume | - | WV full |
| Wash Height | - | WH 4000 |
| Wash cycles | - | WASH 3 |

**Figure 2.1. Hydra™ 96-well dispensing parameters**

Parameters programmed into a Hydra™ to dispense 5μl of DNA from each well of a

'working plate' into the corresponding wells of N 96-well PCR plates. N equals the

number of 96-well PCR plates required.

>gi|393571|emb|Z23382.1|HS126XG7 H. sapiens (D3S1573) DNA segment containing

(CA) repeat; clone AFM126xg7; single read

CTGTGTTTTTCATTTTTGCTTATTAAGATATGCATATACATACCCCTCCAACG

CGCGCGCACACACACACACACACACACACACACACACACACAAACATAT

TTCATTGAGAAAAGTCATGAAATTGGAATAGCCCTGTNATTTACTGGCAG

GTATTTATAATGTTCTGAGCCTTTTCAAGTACAATTTGAACTTGTTTCTAAG

CT

**Figure 2.2. A sequence in FASTA format**

when possible, from primer design analysis. Within available regions, a 'forward and reverse' primer was designed to amplify each sequence tag site (STS) or expressed sequence tag (EST) using certain criteria. Primers were designed with a 3'-clamp of GC or, decreasing in preference, CC, GG, CG, C or G. Whenever possible, primers were designed with a length of 20bp, 50:50 GC/AT content, and a total of 10 G's and C's combined.

## 2.1.1.6. Melting and annealing temperature calculation

The melting temperature ($T_M$) in degrees Celsius for each oligonucleotide was calculated using the formula $T_M = 64.9 + 41*GC/L-500/L$, where GC equals the total number of G's and C's and L equals the length of the primer in basepairs. The annealing temperature ($T_A$) was calculated by subtracting 5°C from the $T_M$.

## 2.1.1.7. Polymerase chain reaction

PCR was used to amplify ESTs, STSs and other DNA fragments. Each PCR reaction consisted of 20ng of genomic DNA, 0.5u of AmpliTaq Gold™ DNA polymerase, 6pmol of 'forward' oligonucleotide, 6pmol of 'reverse' oligonucleotide, 1X PCR buffer II, 1mM dNTP, and 2.5mM $MgCl_2$, in a final volume of 20-50µl. Amplification of each reaction was performed in a 96-well plate using a PTC-100™ or PTC-225™ from MJ Research, Inc. The DNA was denatured and the polymerase activated by incubating the reaction for 12min at 95°C. Initial denaturing was followed by 30-40 cycles consisting of DNA denaturing for 15s at 94°C, oligonucleotide annealing for 15s at the calculated $T_A$, and oligonucleotide extension for 30s or 1min at 72°C depending on the length of the amplicon. Amplicons below 500bp were extended for 30s and amplicons 500bp or larger were extended for 1min. Cyclic amplification was follow by a final extension of 5min at 72°C. Each reaction was then cooled to RT.

## 2.1.1.8. Agarose gel electrophoresis

DNA fragments of various sizes were separated and prepared for visualisation by agarose gel electrophoresis. Since all DNA fragments analysed were smaller than 1kb, only 1X TBE, 2% agarose gels were utilised. To prepare the gel, the gel mixture was heated in a microwave until homogenous and 0.5mg/ml of ethidium bromide was added. The heated gel mixture was poured into a frame containing combs and allowed to cool for 30min at RT. After solidification, the combs were removed and the gel place into a tank containing 1X TBE. DNA was mixed with 5X loading buffer in a 4:1 ratio and transferred into a well on the gel. To size the DNA fragments, 500ng of 100bp ladder were run at the same time in a parallel lane. DNA was pulled through the gel from cathode to anode with a voltage gradient of 2V/cm until all fragments were sufficiently separated.

## 2.1.1.9. Visualisation and imaging of DNA fragments

During agarose gel electrophoresis, ethidium bromide binds to the various DNA fragments as they migrate through the gel. To visualise the DNA fragments, the agarose gel was exposed to UV irradiation causing the ethidium bromide bound to the DNA to fluoresce. Electronic images of each gel were captured using an AlphaImager™ 2000 from Alpha Innotech Co. Graphic copies of each gel were printed using a VP-1500 II and video printer paper both from Seiko Precision, Inc.

## 2.1.1.10. Primer extension pre-amplification

The amount of DNA received for each patient from the Oxford linkage cohort (Section 2.2.1.) was insufficient to complete all of the planned linkage studies. Therefore, in order to increase the concentration 60X, primer extension pre-amplification (PEP) was

conducted (Zhang *et al.*, 1992). Each PEP reaction consisted of 20ng of genomic DNA,

8.3u of AmpliTaq™ DNA polymerase, 3.336nmol of PEP primer, 1X PCR buffer II,

1mM $MgCl_2$, 0.4mM dNTP, and 0.001% gelatin, in a final volume of 100µl.

Amplification of each reaction was performed in a 96-well plate using a PTC-100™

from MJ Research, Inc. The DNA was initially denatured for 10min at 94°C. Initial

denaturing was followed by 50 cycles consisting of DNA denaturing for 1min at 94°C,

oligonucleotide annealing for 2min at 37°C, initial extension starting at 37°C and

increasing to 55°C at 1°C every 10s, and further extension for 4min at 55°C. After

cyclic amplification each reaction was cooled to 4°C.


2.1.1.11. Importing data into a Discovery Manager™ database

The pedigree, phenotype, marker, genotype, and map information for all linkage and

association studies was stored within a Discovery Manager™ database from Genomica,

Inc. All information to be imported into the database was written in either comma or

tab delimited format, where by commas or tabs separate fields specific to the type of

data being imported (Figure 2.3.).


## 2.1.2. Radiation Hybrid mapping methods


2.1.2.1. Amplification of RH panels

To construct radiation hybrid (RH) maps of the linkage regions analysed in this study,

each marker to be mapped was amplified by PCR across the Stanford G3 (G3) RH panel

and the MIT/Whitehead Genebridge 4 (G4) RH panel. The G3 and G4 RH panels are

comprised of 83 and 93 human-hamster hybrid, respectively. Both panels also include

both a human and hamster positive control. Both panels are supplied with 5µg of DNA

from each hybrid clone and control at a concentration of 5µg/ml. The DNA from each

a). OAG.SIBPAIR,103,3,2,1,2,,,,SP01A11,OAG-IBD-001023

b). PopID,KinID,SubID,IBD,CD,UC

OAG.SIBPAIR,103,3,2,2,1

c). D3S1573,homo,sapiens,3,AFM126xg7

d). 38,B,SP01A11,D3S1573,124,126,,RAM Disk:ibd6 r1:ibd6 r1

e). D3S1573    2.3

**Figure 2.3. Examples of Discovery Manager™ data import formats**

a). Pedigree data is written in comma delimited format without a header. The data is divided into the following fields: population identification, kindred identification, subject identification, mother identification, father identification, sex, date of birth, date of death, proband, alias 1, alias 2, and alias $n^{th}$. The first five fields are required. Sex is a coded variable where 0 is unknown, 1 is male and 2 is female.

b). Phenotype data is written in comma delimited format with a required header. The data is divided into the following fields: population identification, kindred identification, subject identification, phenotype 1, phenotype 2, and phenotype $n^{th}$. Phenotypes used in this study were coded variables where 0 was unknown, 1 was unaffected and 2 was affected.

c). Marker data is written in comma delimited format without a header. The data is divided into the following fields: marker name, genus, species, chromosome, alias 1, alias 2, and alias $n^{th}$.

d). Genotype data is written in comma delimited format without a header. The data is divided into the following fields: lane, dye, subject alias, marker, allele 1, allele 2, overflow, and gelfile.

e). Map data is written in tab delimited format without a header. The data is divided into the following fields: marker name and distance to previous marker.

panel was transferred to a deep 96-well plate in the order specified for submission of results, starting with wells A1 - A12 and then B1 - B12, etc. For both panels, the human control was placed in well A4 and the hamster control in A8. Amplification of each marker was performed using PCR (Section 2.1.1.7.) and 25ng of each DNA (5µl) from each panel. The amplification of each marker was performed in duplicate to verify results.

## 2.1.2.2. Visualisation of RH panel amplification

After amplification of a marker across one of the RH panels, the resulting reactions were fractionated and stained with ethidium bromide by agarose gel electrophoresis (Section 2.1.1.8.). The agarose gels were constructed using 52-well combs. When loading the PCR reactions onto the agarose gel each specific DNA amplification was loaded next to its duplicate reaction. Thus, for each maker and RH panel four lanes of 52 wells were used. Lane one contained PCR reactions from row A and B of a 96-well plate, loaded: 100bp ladder, A1, A1 duplicate, A2, A2 duplicate, etc. Lanes two, three, and four contained rows C and D, E and F, and G and H, respectively. After electrophoresis the DNA fragments were visualised, imaged and stored (Section 2.1.1.9.).

## 2.1.2.3. Analysis of RH data and map construction

Each image of an amplification of a marker across an RH panel was analysed for validity and then scored. To validate the RH mapping data, the negative controls, wells H2 – H12 for the G3 panel and H12 for the G4 panel, and positive controls, wells A4 and A8 for both panels, were analysed. The data was considered valid if the negative controls were blank, the human controls had one DNA band present of the correct size for the particular marker amplified, and the hamster controls were either blank or

contained a band(s) of a different size to the positive control which would not interfere

with the calling of data from human-hamster hybrid PCR. The remainder of the

duplicate data was scored in the sequence of clones specified for each RH panel. For

each clone, two positive results (the presence of a band the same size as the human

control band) were scored as a 1, two negative results (no band matching the human

control band) were scored as a 0, and one positive and one negative result were scored

as a 2 for the G4 panel or as a R for the G3 panel. The data was used to place the

marker on a pre-existing framework RH map and to construct new RH maps solely from

RH data generated from this project.


### 2.1.2.3.1. Stanford RHServer

Data generated from the amplification of a marker across the G3 RH panel was

submitted through the Stanford RHServer Internet site for analysis. The chromosomal

assignment expected was selected. The closest five markers with a resulting LOD score

of greater than 4.0 were designated for retrieval. After submission, results were

returned by email (Figure 2.4.). By comparing the resulting distances between the

marker submitted and the markers retrieved to the order of the retrieved markers on the

G3 RH map, the new marker could be placed on the map (see example – Figure 2.4.).

After all markers from a particular chromosome were placed on the map, the distances

between each submitted marker were calculated using the distances given on the G3 RH

map and a G3 RH map consisting of only the submitted markers was constructed.


### 2.1.2.3.2. MIT/Whitehead Institute

Data generated from the amplification of a marker across the G4 RH panel was

submitted to the Whitehead Institute/MIT Center for Genome Research RH server

Internet site for analysis. The official Genebridge4 order was selected and a LOD score

a).

G3.D3S1573    000000011000100000000000000000010000000000000000100100000

0000000R0001R01000001010111


b).

| # | SHGCNAME | CHROM# | LOD_SCORE | DIST.(cRs) |
|---|---|---|---|---|
| 1 | SHGC-1705 | 3 | 9.56 | 13 |

Vector:0000000110000000000000000000000000110000100000010000000000000011

0010000000001000001

| 2 | SHGC-56838 | 3 | 8.71 | 14 |

Vector:0000000000000000000000000000000000110000100000010000000000000011

0010000000001000001

| 3 | SHGC-21615 | 3 | 8.65 | 18 |

Vector:0000000100000000000000000000100000011000010000001000000000000011

0110000000001000001


**Figure 2.4. G3 RH data submission and results via the Stanford RHServer**

a). An example of the format required for submission of G3 RH data.

b). An example of the results received after submission of G3 RH data (only three of

the five markers retrieved are shown). In the above example, if on the G3 RH map

marker SHGC-1705 is 1cR above marker SHGC-56838, then the marker submitted

could be placed on the same map 13cR above marker SHGC-1705.

of greater than 3 designated as the cut off for placing the marker on one of the G4 RH

framework maps. After submission, results were returned by email (Figure 2.5.). The

entire chromosome map was returned with all the submitted markers placed on the map

and distanced between markers calculated (see example Figure 2.5.). After all markers

from a particular chromosome were placed on the map, the distances between each

submitted marker were calculated and a G4 RH map consisting of only the submitted

markers was constructed.

### 2.1.2.3.3. RHMAP

To construct an RH map without framework markers and derived only from RH data

generated through this project, RHMAP (version 2.01) (Boehnke *et al.*, 1991) was used.

The data for analysis was imported into the program (Figure 2.6.) and analysed using a

multiple point maximum-likelihood test. Possible marker orders were written to a

'results' file by descending LOD score. The LOD score associated with each order was

derived from the likelihood of that order calculated from the number of breaks required

to generated the RH data imported. The program could only analyse eight markers at a

time so multiple rounds of analysis with different combinations of markers were

conducted for each region. From the multiple results the most consistent marker order

was chosen and the distances associated with it used for the final map construction.

### 2.1.2.4. Conversion of RH maps to genetic maps

The RH maps were converted to genetic maps for linkage analysis. The G3 and G4 RH

panels were generated with different amounts of irradiation, 10,000rad and 3,000rad

respectively, and thus have a different factor for conversion from centiRays (cR) to

centiMorgans (cM). For this project, conversion factors of $1cR_{10,000} = 25kb$, $1cR_{3,000} =$

270kb and $1Mb = 1cM$ were used (Carey, 1997).

a).

D14S789 10101 01000 10001 11010 01001 10000 10110 00101 01011 01101 00001

10000 01000 00100 00020 10111 00100 00010 000

b).

D14S789     Chromosome Chr14

            Places 1.71cR from D14S872 (lod > 3.0)

| Name | Dist | Type | Vector |
|------|------|------|--------|
| D14S872 | 1.7 | F | |

101010100010001110100000110010101100010101011011010000110000010000100

00010101110010000010000

| D14S789 | 11.6 | P | >3.0 |

101010100010001110100100110000101100010101011011010000110000010000100

00020101110010000010000

| D14S288 | 7.7 | F | |

101010100010121110100000110000101000010111001011010000110000010000001

00010101110000000010000

Figure 2.5. G4 RH data submission and results via the MIT/Whitehead RH server

a). An example of the format required for submission of G4 RH data.

b). An example of the results received after submission of G4 RH data (only an excerpt of the entire map is shown). From the excerpt it can be read that, from the data submitted for D14S787, D14S789 is located 1.7cR below D14S872 and 11.6cR above D14S288.

```
1  8 83  1  2
1.2 1.4 1.5 1.8 2.5 2.6 3.5 3.13
(A2,8(1X,A1),T3,I1)
102
 1 0 0 0 0 0 0 0 0
 2 0 0 0 0 0 0 0 0
 3 0 0 0 0 0 0 0 0
 4 0 0 0 0 0 0 0 0
 5 0 0 0 0 0 0 0 0
 6 0 0 0 0 0 0 0 0
 7 0 0 0 0 0 0 0 0
 8 1 1 1 1 1 1 1 1
 9 1 1 0 0 0 0 0 0
10 0 0 0 0 0 0 0 0
11 0 0 0 0 0 0 0 0
12 0 0 0 0 0 0 0 0

72 0 0 0 0 0 0 0 0
73 0 0 0 0 0 0 0 0
74 0 0 0 0 0 1 0 0
75 0 0 0 0 0 0 0 0
76 0 0 0 0 0 0 0 0
77 1 1 1 1 1 1 1 1
78 0 0 0 0 0 0 0 0
79 0 0 0 0 0 0 0 0
80 0 0 0 0 0 0 0 0
81 0 0 0 0 0 0 0 0
82 0 0 0 0 0 0 0 0
83 1 1 0 0 0 0 0 0
 8  1  4  1  0  1  1  1.00
1.2 1.4 1.5 1.8 2.5 2.6 3.5 3.13
 0  0.00  3.00  3.00
1.2 1.4 1.5 1.8 2.5 2.6 3.5 3.13
```

**Figure 2.6. An example of the import format for RHMAP**

The first line indicates the program to be used, the number of markers to be analysed, the number of hybrids assayed, the number of problems to be performed and whether the species being analysed is haploid or diploid. The second line is a list of the markers to be analysed. The third line contains the run parameters. The fourth line indicates how the information is coded 1-positive, 0-negative, and 2-ambigious. The first four lines are followed by the data (hybrids 13-71 would normally be included) and the problem parameters.

71

## 2.1.3. Polymorphism detection methods

### 2.1.3.1. Fragment selection and amplification

The complete coding sequence of each candidate gene was retrieved from the nucleotide

database within the Entrez browser at the NCBI Internet site by searching for entries

that included the specific gene name. The 5' and 3' UTR from each sequence was used

to search for significant matches to genomic DNA sequence in either the NR or HTGS

database by using BLAST 2.1 and the BLASTn algorithm at the NCBI Internet site.

After a genomic DNA sequence was retrieved, it was analysed for matches to human

DNA repeats using RepeatMasker2 at the University of Washington Internet site.

Oligonucleotides were designed (Section 2.1.1.5.) in non-repetitive genomic sequence

to amplify DNA fragments between 500bp and 800bp around and within the candidate

genes. The fragments were then amplified across 12 control DNA samples using PCR

(Section 2.1.1.7.) in preparation for polymorphism detection.

### 2.1.3.2. Calculation of fragments' melting domains and temperatures

The sequences of DNA fragments to be used in polymorphism detection were imported

into Wavemaker (version 3.4.4) from Transgenomics, Inc. A melting curve analysis

was conducted comparing the percentage of fragment denaturing to an increasing

temperature gradient. A fragment that gave a melting curve that transitioned from

completely annealed to completely denatured without stabilising momentarily in a

partially denatured state was designated as having only one melting domain. Any

plateaus of partial denaturing over small temperature gradients were designated as

divisions between separate melting domains. The temperatures at which 30% of each

domain was denatured were selected for use in polymorphism detection.

2.1.3.3. Running fragments on the WAVE™

The detection of polymorphisms within fragments was conducted by denaturing high-performance liquid chromatography (dHPLC) using the WAVE™ from Transgenomics, Inc. A 96-well plate containing amplified fragments was placed in the cold block within the WAVE™. The capillary was heated to the chosen temperature for polymorphism detection and equilibrated with the programmed percentages of buffer A and buffer B. Since, buffer A and buffer B were the only solutions used their percentages always combined to 100% and thus further reference to buffer A concentration will be excluded. After 0.1 seconds of equilibration, the percentage of buffer B was increased by 5% and 5μl of amplified fragment were injected into the capillary. After injection, the percentage of buffer B was increased at a rate of 2% per 1min for 4min. The percentage of buffer B was then decreased back down to the equilibration conditions for 2min to clear the capillary of any residual fragment and prepare it for a subsequent injection. During the increasing buffer B gradient if the polymorphism detection gradient and temperature were selected correctly, the binding of the fragment to the resin in the capillary is broken and the fragment is released from the capillary and detected. Any heterodueplex DNA strands, caused during amplification of a fragment containing a polymorphism, bind at a lower affinity to the resin in the capillary and are released and detected at a lower percentage buffer B concentration than the complimentary homoduplex strands. Thus, any fragments that produced a WAVE™ detection trace with a shoulder or separate peak prior to the homoduplex peak were designated as containing a polymorphism.

To construct a WAVE™ polymorphism detection program for a fragment, the percentage of buffer B at which the fragment would be released from the capillary at a set temperature was calculated. On average, a 700bp fragment in a completely annealed

73

state will be released from the capillary at a 70% buffer B concentration. The same

fragment, completely denatured, will usually be released at a 52% buffer B

concentration. Using this information, a linear correlation was drawn between the

percentage of buffer B and fragment denaturing. This linear correlation was used to

extrapolate the percentage of buffer B at which a fragment would be released from the

capillary from the percent of fragment denaturing at the selected polymorphism

detection temperature (Section 2.1.3.2.). The calculated percentage of buffer B

concentration was then used to construct a polymorphism detection program for the

fragment (Figure 2.7.)

## 2.1.3.4. Exo/SAP digestion of oligonucleotides

To prepare amplified fragments for sequencing, an Exo/SAP digestion was conducted to

remove all oligonucleotides remaining from the previous amplification. To a 20µl

fragment PCR reaction, 1u of exonuclease I and 1u of shrimp alkaline phosphatase

(SAP) were added. The reaction was incubated at 37°C for 1hr in a PTC-100™ from

MJ Research, Inc. After digestion, the temperature was increased to 80°C for 15s to

heat inactivate the exonuclease. The reaction was stored at -20°C until used in a

subsequent sequencing reaction.

## 2.1.3.5. End terminator cycle sequencing reactions

To verify the presence of polymorphisms detected by dHPLC, end-terminator cycle

sequencing of the potentially polymorphic fragments and a non-polymorphic control

fragments was conducted. A sequencing reaction was set-up by combining 5µl of the

exo/SAP digested fragment PCR reaction with 8µl of sequencing reagent premix and

6pmol of either the 'forward' or 'reverse' oligonucleotide used in the original

| Step | Time | % buffer B |
|------|------|-----------|
| Equilibration | 0.0 | 56 |
| Injection and gradient initiation | 0.1 | 61 |
| Gradient termination | 4.1 | 69 |
| Wash initiation | 4.2 | 56 |
| Wash termination | 6.2 | 56 |

**Figure 2.7. An example of a WAVE™ polymorphism detection program**

The program is for a DNA fragment, which at the chosen polymorphism detection

temperature will be released from the capillary at a calculated buffer B concentration of

65%. The calculated buffer B concentration always occurs in the middle of the

gradient.

amplification of the fragment. Water was added to the reaction to increase the volume to 20μl. Cycle sequencing was conduced in a PTC-100™ from MJ Research, Inc. with 30 cycles of denaturing at 96°C for 30s, oligonucleotide annealing at 50°C for 20s and single strand extension at 60°C for 1min. After cycle sequencing the reaction was cooled and stored at 4°C.

## 2.1.3.6. Purification of sequencing reactions

All sequencing reactions were purified to remove proteins, dNTPs, fluorescent ddNTPs and salt. To a 96-well centrifuge plate taped onto a 96-well collection plate, 200μl of homogenous P10 gel solution were added to each well and centrifuged for 2min at 2,000rpm in a RT-7™ from Sorvall Ltd. All water collected in the collection plate was removed. Again, 200μl of homogenous P10 gel solution were added to each well, centrifuged for 2min at 2,000 rpm, and any water collected was discarded. A 20μl sequencing reaction and 20μl of water were then added to each well. Underneath each well of the centrifuge plate containing a sequencing reaction, 80μl of ethanol and 8μl of 3M sodium acetate were added to the corresponding well in the collection plate. The plate was centrifuged again for 2min at 2,000rpm. The collection plate, now containing the sequenced DNA, was placed at -20°C for 10min and then centrifuged at 3,800 rpm for 35min at 4°C to pellet the DNA. The plate was removed from the centrifuge and the ethanol solution was poured off. To remove more salt, 100μl of 70% ethanol were added to each well and the plate centrifuged for 10min. The ethanol solution was again poured off and any residual liquid remove with a pipette. The DNA pellet was allowed to dry for 1hr at RT. To resuspend the DNA and prepare it for loading on a polyacrylamide gel, 4μl of formamide loading dye were added to each well. The plate was stored at 4°C.

2.1.3.7. Preparation of a 48cm polyacrylamide sequencing gel

To separate sequencing fragments, 6M urea, 1X TBE polyacrylamide gels were

prepared. For each gel, approximately 28ml of sterile water were added to 18g urea and

5ml of Sequagel XR to give a final volume of 45ml. To de-ionise the mixture, 0.5g of

mixed bed resin were added and the mixture stirred for 30s. The mixture was placed at

RT and allowed to de-ionise while 5ml of 10X TBE were vacuumed through a $0.2\mu C$

filter into a 100ml collection jar. The resin was removed from the mixture and the

mixture added to the filtered 10X TBE by also passing it through the $0.2\mu C$ filter. The

resulting gel mixture was de-gassed for 10min by placing a lid onto the filter unit to

create a vacuum. Following de-gassing, $250\mu l$ of 10% ammonium persulfate (APS) and

$70\mu l$ of temed were added to the gel mixture to cross-link the acrylamide. Before the

cross-linking was complete, the gel mix was poured between two 48cm long glass plates

separated by 0.2cm wide spacers. A 48-well comb was inserted between the plates at

the top of the gel and the gel was allowed to solidify for 2hr.


2.1.3.8. 48cm sequencing gel electrophoresis

Sequence fragments generated from cycle sequencing reactions were separated by

electrophoresis through a 48cm polyacrylamide sequencing gel. Simultaneously, the

fluorescent labels attached to the ddNTPs incorporated in the sequence fragments were

excited using a laser and the subsequent emissions were detected and stored as a gel

image. This electrophoresis, laser excitation, and detection of sequence data was

performed using a 377 from PE Applied Biosystems, Inc. Before running a sequencing

gel, the 48-well comb and all debris in the scan region were removed. The gel was then

clamped into a 377 along with an upper and lower chamber for electrophoresis buffer.

Both chambers were filled with 1X TBE. Purified sequencing samples in a 96-well

plate were denatured for 6min at 95°C in a PTC-100™ from MJ Research, Inc. and then placed on ice. Starting with the samples in odd numbered wells, a 0.5µl aliquot of each sample was then loaded onto the gel into the corresponding lane. Prior to loading the even samples, the odd samples were run into the gel for 2min by electrophoresis. After all the samples were loaded onto the gel, the sample sheet with the corresponding lane to sample information, run module 48A-1200, plate check A, and the latest sequencing matrix were selected within the 377 run software. The sequencing gel run was then initiated and the gel image data collected. Following the run, each sample's sequence data was automatically tracked and extracted from the gel image using sequence analysis software from PE Applied Biosystems, Inc.

2.1.3.9. Analysing sequence data

Sequence files created during the extraction of sequence gel data were imported and analysed within Sequencher (version 4.0.5b5) from Gene Codes Co. Sequence of poor quality, defined as more than 1 ambiguous base every 20bp, was auto-trimmed from both ends of sequences analysed. After trimming, all sequences were ran through an algorithm, which aligned sequences that were 75% similar over at least 30bp. Corresponding 'forward' and 'reverse' sequence data from a fragment amplified in subjects showing both polymorphic and non-polymorphic WAVE™ traces were joined into contigs by this alignment process. The sequence data within contigs was then analysed for differences between subjects. By comparing chromatograms of sequence data between subjects and 'forward' and 'reverse' sequence reads, differences detected were determined to be either polymorphisms or sequence calling errors. All polymorphisms detected were correlated back to the data generated by dHPLC (Section 2.1.3.3.).

## 2.1.4. SNP, insertion and deletion genotyping methods

2.1.4.1. Assay design

Assays were developed to genotype selected polymorphisms within and around

candidate genes. The polymorphisms were found by polymorphism detection methods

(Section 2.1.3.) or by using Sequencer (version 4.0.5b5) from Gene Codes Co. to align

cDNA sequences retrieved from the nucleotide database at the NCBI Internet site.

Polymorphisms previously reported in other studies were also selected for genotyping.

For the genotyping of insertion and deletions 30bp or larger, PCR assays were designed.

For smaller insertions and deletions and for SNPs, PCR restriction fragment length

polymorphism (PCR-RFLP) assays were designed. Designer Mismatch™ from Oxagen

Ltd. was used to analyse the polymorphisms to be genotyped by PCR-RFLP for

restriction endonucleases that would discriminate between alleles. If a naturally

occurring discriminating restriction site was not present, the output from the program

highlighted bases around the polymorphism that could be changed to create a

discriminating restriction site. The output also included the number of bases separating

the polymorphism and the next natural restriction site both 5' and 3' of the

polymorphism. This information was used to select a restriction endonuclease and

oligonucleotides that when used together in a PCR-RFLP assay produced results that

made the discrimination of all genotypes possible. All assays were designed so that the

non-deleted alleles of the PCR assays and the undigested alleles of the PCR-RFLP

assays were smaller than 800bp and always at least 30bp larger than their respective

deleted alleles and largest digested allelic fragments.

## 2.1.4.2. Oligonucleotide design

When possible, oligonucleotides previously designed to amplify polymorphism detection fragments (Section 2.1.3.1.) were used for the insertion, deletion and SNP genotyping assays. For assays where the oligonucleotides designed previously were not useable due to the absence of either a natural discriminating restriction site or discriminating restriction pattern, new oligonucleotides were designed. All new oligonucleotides, except those needed to create a restriction site, were designed using the guidelines presented previously (Section 2.1.1.5.). For assays where a naturally occurring discriminating restriction site was not present, oligonucleotides were designed with a mismatch, or non-homologous base, towards the 3' end, which when used in amplification would create a restriction site that could be used to discriminate alleles of the polymorphisms. When designing a mismatch oligonucleotide, of the possible mismatches, the mismatch furthest away from the polymorphic site was chosen in order to stabilise the annealing of the oligonucleotide. Each mismatch oligonucleotide was also designed so that the 3' end would terminate at the base prior to the polymorphic site. All mismatch oligonucleotides designed were of sufficient length to produce a digested product greater than 30bp.

## 2.1.4.3. Amplification by PCR

Genomic DNA to be genotyped for an insertion, deletion, or SNP was amplified in a 20µl PCR reaction with a 60C $T_A$ and 35 cycles (Section 2.1.1.7.).

## 2.1.4.4. Restriction endonuclease digestion of PCR amplified DNA

Restriction endonuclease digestion was used to discriminate between alleles from insertion, deletion, and SNP polymorphisms. To a 20µl PCR reaction, 5u of the restriction endonuclease specific to the particular genotype assay being conducted, 2µl

of the corresponding 10X restriction endonuclease buffer, and 2μl of 100X BSA, if

required, were added. The new reaction was mixed by pipetting and then incubated for

1hr in a PTC-100™ from MJ Research, Inc. at the temperature at which the specific

restriction endonuclease is active. The digestion reaction was then cooled to RT and

stored at 4°C.

## 2.1.4.5. Visualisation, analysis and storage of genotype results

Fragments generated in insertion, deletion and SNP PCR and PCR-RFLP assays were

separated by electrophoresis through an agarose gel (Section 2.1.1.8.). The results were

then visualised and captured (Section 2.1.1.9.). The allele corresponding to the largest

DNA fragment was labelled 1 and the other allele was labelled 2. The resulting

genotype data was stored in a Discovery Manager™ database from Genomica, Inc.

(Section 2.1.1.11.) for later analysis.

## 2.1.5. STR genotyping methods

### 2.1.5.1. Marker selection

STR markers to be used for genotyping were selected from the Marshfield sex-average

genetic maps within the Discovery Manger™ reference database from Genomica Inc.

If, when selecting markers, a choice between possible markers was required, then the

marker with the highest heterozygosity and where oligonucleotides could be designed in

non-repetitive sequence was selected.

### 2.1.5.2. Oligonucleotide design

For the STR markers selected, oligonucleotides were designed (Section 2.1.1.5.) on

either side of the STR to be used for amplification and subsequent genotyping. To the

5' end of all 'reverse' oligonucleotides a 'pig-tail' of GTTTCTT was added so that

adenylation of the 3' end of the forward strand would occur during PCR amplification

(Brownstein *et al.*, 1996). To the 5' end of all 'forward' oligonucleotides a FAM, TET,

or HEX fluorophor was added.


2.1.5.3. Panel design

All selected markers for a specific study were designed into panels, so that several

markers could be analysed simultaneously. The size of the amplicon generated for each

marker during PCR was calculated by adding the number of base-pairs present in the

marker's database entry, starting from the 5'end of the 'forward' oligonucleotide and

ending at the 5' end of the 'reverse' oligonucleotide. The markers were then sorted by

size from smallest to largest. Starting at the top, each marker was assigned a fluorophor

in a repetitive cycle of FAM, then TET, and then HEX. Next, the markers were

separated into groups by fluorophor and again sorted by size from smallest to largest. If

two markers with the same label were less than 40bp apart another panel was designed

by repeatedly numbering from the top of each group with the new number of panels

desired until reaching the bottom of each group. The markers were then separated into

panels and examined again for potential overlaps. This process was repeated until all

adjacent markers labelled with the same fluorophor in a panel were more than 40bp

apart.


2.1.5.4. PCR amplification and pooling

The amplification of STR markers by PCR was set-up and run either manually (Section

2.1.1.7.) or robotically using a RapidGene™ automated system from Oxagen Ltd. The

subsequent pooling of the amplified products into panels was solely performed using a

RapidGene™. To prepare the automated system for use, the number of markers to be

amplified and their panel, pooling, and PCR program information was entered into a spreadsheet within a computer interface. For each marker, a 96-well PCR plate was placed into a 'hotel' and a 1.5ml tube placed into a cold block set at 4°C. The PCR plates contained 20ng of dried genomic DNA in each well. The 1.5ml tubes contained 5µM of appropriate 'forward' oligonucleotide, 5µM of the corresponding 'reverse' oligonucleotide, 1X PCR buffer II, 2.5mM $MgCl_2$, 1mM dNTP, and 0.06u/µl AmpliTaq Gold™ DNA polymerase in final volume of 1.2ml. The marker specific plates and tubes were inserted into the 'hotel' and cold block in the same order as entered into the spreadsheet. A 96-well pool plate for each panel was also placed into the 'hotel' in the order entered into the spreadsheet. Once the system was set-up the run was initiated. During the run, 10µl of the appropriate PCR reaction mixture and 90µl of mineral oil were dispensed into the wells of the plates. The DNA in each plate was then amplified using a PTC-225™ from MJ Research, Inc. according to the associated PCR program. All programs contained the thermal cycling parameters described previously (Section 2.1.1.7.). Following amplification the reactions were pooled into panels according to the ratios entered into the spreadsheet and dispensed into the appropriate pool plate. The pool plates were then manually sealed and stored at 4°C.

2.1.5.5. Preparation of a 36cm polyacrylamide genotyping gel

To separate pooled fluorescently labelled DNA fragments, 6M urea, 1X TBE polyacrylamide gels were prepared. For each gel, approximately 28ml of sterile water were added to 18g urea and 5ml of acrylamide/bis 29:1 4% C to give a final volume of 45ml. To de-ionise the mixture, 0.5g of mixed bed resin were added and the mixture stirred for 30s. The mixture was placed at RT and allowed to de-ionise while 5ml of 10X TBE were vacuumed through a 0.2µC filter into a 100ml collection jar. The mixed bed resin was removed from the mixture and the mixture added to the filtered 10X TBE

by also passing it through the 0.2µC filter. The resulting gel mixture was de-gassed for 10min by placing a lid onto the filter unit to create a vacuum. Following de-gassing, 250µl of 10% APS and 35µl of temed were added to the gel mixture to cross-link the acrylamide. Before the cross-linking was complete, the gel mix was poured between two 36cm long glass plates separated by spacers which graduated from 0.4cm wide at the top to 0.2cm wide at the bottom. The trough side of a 96-lane 'shark tooth' comb was inserted between the plates at the top of the gel and the gel was allowed to solidify for 2hr.

## 2.1.5.6. 36cm genotyping gel electrophoresis

Pooled DNA fragments generated from the amplification of STR markers were separated by electrophoresis through a 36cm polyacrylamide genotyping gel. Simultaneously, the fluorescent labels attached to the 5' ends of the fragments were excited using a laser and the subsequent emissions were detected and stored as a gel image. This electrophoresis, laser excitation, and detection of DNA fragments was performed using a 377 from PE Applied Biosystems, Inc. Before running a genotyping gel, all debris in the scan region were removed and the 96-lane 'shark tooth' comb was inverted. The gel was then clamped into a 377 along with a cooling plate and an upper and lower chamber for electrophoresis buffer. The chambers were filled with 0.5X TBE and 1X TBE, respectively. In each well of a 96-well plate, 2µl of pooled genotyping samples were added to 3µl of TAMRA loading buffer and denatured for 6min at 95°C in a PTC-100™ from MJ Research, Inc. Following denaturing, the samples were placed on ice. Starting with the samples in odd numbered wells, a 0.5µl aliquot of each sample was then loaded on the gel into the corresponding lane. Prior to loading the even samples, the odd samples were run into the gel for 2min by electrophoresis. After all samples were loaded onto the gel, the sample sheet with the corresponding lane to

sample information, the latest TAMRA matrix, plate check C, and the 36C-2400 pre-run

and run module were selected within the 377 run software. The genotyping gel run was

then initiated and the gel image data collected.


## 2.1.5.7. Tracking and extracting genotype data

Tracking and extracting of sample genotype data from gel images was performed using

Genescan (version 3.0) from PE Applied Biosystems, Inc. The tracking of each lane

was done automatically. Following the tracking of lanes, the TAMRA size standard

peaks from the first two lanes were analysed and labelled with the appropriate sizes in

basepairs according to the literature accompanying the size standard. The size data,

from lanes one and two, were then applied to rest of the odd and even lanes,

respectively. Analysis parameters were set so that peaks with intensities greater than 50

were analysed. The four fluorescent channels were selected for each lane and the

genotype data was extracted. The peaks in the FAM, TET and HEX channels were

converted into basepairs relative to the TAMRA size standard.


## 2.1.5.8. Allele calling

The genotypes for each sample and marker were called and labelled using Genotyper

(version 2.0) from PE Applied Biosystems, Inc. The extracted data and a template

corresponding to the panel of markers being analysed were imported into the program.

A template was composed of a list of markers and their fluorescent label, size range,

and allelic bins. Following the import, a macro was run, which correlated the data

within each lane to the data within the template. Within a marker size range, the macro

labelled the peak of greatest intensity with the allele name associated to the bin

encompassing the peak's size in basepairs. Once the first peak was labelled, the macro

labelled any other peaks in the marker size range that had a greater size in basepairs and

were at least 30% the intensity of the first marker. If multiple secondary peaks were labelled then the macro would remove the labels from all peaks except the most intense one. If no secondary peaks were found then only the initial peak was labelled. After the peaks were called, the data was written to a table. For each sample and marker, the table included the two alleles called, their size in basepairs relative to the TAMRA size standard, and the peak intensity. Each genotype was then visually checked to correct any mistakes made by the macro. After all corrections, the table was automatically updated with any changes and the genotype data exported. The exported data was imported into a Discovery Manager™ database from Genomica, Inc. (Section 2.1.1.11.).

2.1.5.9. Designing templates

For each STR used for genotyping, a template for calling each allele was constructed using the extracted data corresponding to the first 95 individuals amplified with each marker. For each marker, the respective extracted data was imported into Genotyper (version 2.0) from PE Applied Biosystems, Inc. and the appropriate colour and range for the marker was selected. Manually, by selecting and clicking with a computer 'mouse', the peaks of each allele for each sample were labelled with their relative size in basepairs compared to the TAMRA size standard. The maximum and minimum relative size for each unique allele within the 95 samples was then found using the statistics window. The average of the maximum and minimum size for each allele was designated as the midpoint size for each allele, respectively. The spacing between consecutive alleles was determined by subtracting the corresponding lower midpoint value from the higher value. The average spacing between consecutive alleles was then used to assume the midpoint of alleles not present in the first 95 samples but potentially existing in other samples due to a gap between the alleles present. For example, if the

STR marker A was a di-nucleotide repeat, then between the smallest and largest allele present in the first 95 samples, other alleles could exist approximately every 2 relative basepairs. Thus if alleles with relative midpoints of 122.56, 124.53 126.50 and 130.44 were the only alleles present in the first 95 samples amplified with marker A, then another allele at 128.47 may exist and would therefore be created. After all observed and potential midpoints were calculated, a specific whole number was then designated for each allele usually in a pattern corresponding to the type of repeat and the whole numbers that the majority of midpoints rounded to. Continuing the example above, because the marker contains a di-nucleotide repeat and thus should have alleles separated by 2bp and because the majority of the allelic midpoints round to an odd number then all the alleles would be designated as odd numbers separated by 2bp. Therefore, the allele midpoints 122.56, 124.53, 126.50, 128.47 and 130.44 would be designated allele 123, 125, 127, 129, and 131, respectively. The midpoint and corresponding designated whole number calculated for each observed and potential allele of a marker were then used to construct a template to automatically and consistently call and label the alleles from the amplification of future samples with that marker (Figure 2.8.). For all alleles a range of +/- 0.4bp relative to the TAMRA size standard was used as the possible allele to allele and gel to gel deviation. FAM, TET, and HEX were always designated as blue, green and yellow, respectively.

2.1.5.10. Pooling optimisation

Before pooling amplified STR markers into their corresponding panels, a pooling strategy was constructed to optimise the intensity of each marker relative to the others. Two control DNA samples were individually amplified for each marker in a panel (Section 2.1.1.7). Each amplified sample was then run on a 36cm genotyping gel

Marker A

| Label as | 123 | all peaks between | 122.56 | +/- 0.4 | in blue. |
|----------|-----|-------------------|--------|----------|----------|
| Label as | 125 | all peaks between | 124.53 | +/- 0.4 | in blue. |
| Label as | 127 | all peaks between | 126.50 | +/- 0.4 | in blue. |
| Label as | 129 | all peaks between | 128.47 | +/- 0.4 | in blue. |
| Label as | 131 | all peaks between | 130.44 | +/- 0.4 | in blue. |

**Figure 2.8. An example of a Genotyper template for marker A labelled with FAM**

(Section 2.1.5.6.), one sample and marker per lane, and analysed using Genescan

(version 3.0) from PE Applied Biosystems, Inc. (Sections 2.1.5.7.). The extracted data

was imported into Genotyper (version 2.0) from PE Applied Biosystems, Inc. and all

alleles present for each sample and marker were labelled by manually clicking with the

computer 'mouse'. From the intensity data, the average undiluted intensity was

calculated for each marker. The resulting intensity was entered into a column in Excel

97 from Microsoft, Inc. For pooling using the Rapidgene™ from Oxagen Ltd. (Section

2.1.5.4.), any whole integer volume between 2µl and 8µl could be pooled for each

marker. Thus, to start calculating the optimal pooling strategy, the average pooling

volume of 5µl was inserted into the column next to each undiluted intensity. In the next

column a calculation was performed multiplying the average intensity of each marker

and the inserted pool volume and then dividing the results by the total volume of all

samples pooled. This gave the average intensity of each marker in a panel after pooling

if pooled in the same ratio. The volumes associated with each marker were then altered

up and down between 2µl and 8µl until the average intensity of each marker in a panel

was as similar as possible. The resulting volumes for each marker were used when

pooling as the optimal pooling conditions.


## 2.1.6. Statistical genetics methods


2.1.6.1. Pedigree and genotype data quality checking by inheritance analysis

The pedigree and genotype data imported into the database was analysed for errors

using Pedcheck (version 1.1) (O'Connell and Weeks, 1998). Pedcheck analyses the

genotypes associated with each subject of each pedigree and writes to a file any data

inconsistent with a Mendelian mode of inheritance. Before running Pedcheck, the data

to be analysed was selected within the database and exported in Pedcheck file format.

After running Pedcheck, any detected errors in the pedigree information were corrected. For example, if one child was consistently showing misinheritance with the father but the other children were consistent with the father, then a new father would be created for the inconsistent child and thus the child would only be a half-sibling of the other children. After correcting all pedigree errors, the data was re-exported and re-analysed. Any remaining inheritance errors were considered genotype errors or real mutations. For each error detected by Pedcheck, the genotypes most likely causing the error were also written to the output file. These genotypes were subsequently removed and not used in any further analysis.

2.1.6.2. Genotype data quality checking by recombination analysis

Potential errors in genotype data were detected using CRI-MAP (Lander and Green, 1987). CRI-MAP analyses the pedigree and genotype data for recombination points between markers. The genotype data was exported from the database in CRI-MAP format and imported into CRI-MAP with the chromosomal order of the markers designated from p to q. During the analysis, CRI-MAP produced an output file that contained a diagram of each subject's parental chromosomes and whether the allele of each marker on the chromosome was grand-paternal or grand-maternal. Since a recombination event occurs approximately every 100cM per generation and since all makers analysed were less than 20cM away from the next analysed marker, any genotype causing a recombination point on both sides of a marker was considered false, removed and not used in any further analysis.

2.1.6.3. Genetic map construction using genotype data

To verify the marker order of the genetic or RH map being used for CRI-MAP and linkage analysis and to check for markers with poor quality genotype data, genetic maps

were constructed from the genotype data using MultiMap (version 2.0) (Matise *et al.*, 1994). Multimap analyses genotype data for the number of recombination points between markers and uses the data to construct sex-average maps with marker order and inter-marker distances represented in Kosambi centiMorgans. Before running MultiMap, the genotype and pedigree data were exported from the database in MultiMap format. Extended maps were constructed with markers added into inter-marker intervals only if placement into the interval resulted in a maximum likelihood score greater than a LOD of 3.0. 'Flips' analysis was conducted, where the order of every three markers was inverted to see whether the output order was more likely than any other order. The output order had to be at least a LOD of 3.0 more likely than any other order or the markers 'flipped' were removed from the final map output. Any markers not placed in the final map were placed into bins between the mapped markers showing their potential positions. The final map was compared to other maps for overall distance, inter-marker distance, and marker order.

## 2.1.6.4. Linkage analysis

Nonparametric multipoint linkage analyses across all regions was performed using MAPMAKER/SIBS implemented within GENEHUNTER (version 2.0) (Daly *et al.*, 1998). Prior to analysis, the genotype, pedigree and map information were selected within the database, exported in GENEHUNTER format and imported into GENEHUNTER. During analysis, all affected sib pairs in each pedigree were counted, using the "weighted" option. Analysis was performed under conditions that included the possibility of dominance variance. All data were analysed under three phenotype models: 1) any form of IBD designated as affected, 2) only CD patients designated as affected and 3) only UC patients designated as affected. Allele frequencies for each STR marker were calculated using available genotypes from all individuals in every

pedigree, both affected and unaffected. The allele frequencies were consistent with those generated by genotyping the same markers in other independent populations. The RH map generated for each region was used for marker order and spacing. The GENEHUNTER program was used to calculate the information content for each marker.

### 2.1.6.5. Association analysis

Association analysis was performed using GENEHUNTER (version 2.0) (Daly *et al.*, 1998). Association of a single marker with a disease was tested using the TDT function. Association analysis of haplotypes and inter-marker LD of adjacent markers were performed using the TDT2 function. To test how frequently the results could be achieved by chance, 1000 randomly generated equivalent data sets were analysed using the Perm1 and Perm2 functions after the TDT and TDT2 functions, respectively.

## 2.2. Subjects

### 2.2.1. Oxford linkage cohort

Ethical approval was obtained from the Central Oxford Research and Ethics Committee (COREC) for the collection of IBD families and their subsequent participation in genetic studies. A total of 234 UK Caucasian families were identified for the present study, 151 of that were included in a previous two-stage genome scan (SIBS1&2) (Satsangi *et al.*, 1996b). The families consisted of 107 CD, 78 UC, and 49 mixed families. Mixed families were defined as those containing any first- or second-degree relative with a discordant form of IBD. Two of the families comprised two generations of affected siblings, giving a total of 236 nuclear families. Of the nuclear families, 212

contained two affected siblings and 24 contained three affected siblings. In total, there

were 284 IBD ASPs, consisting of 143 CD ASPs, 90 UC ASPs and 51 mixed ASPs

(Table 2.1.). Blood samples from the parents of affected siblings were collected when

possible, resulting in 101 nuclear families with DNA collected from two parents, 64

families with DNA from only one parent and 71 families without DNA from either

parent. Of the families without parental DNA, 27 included a DNA sample from at least

one unaffected sibling. Diagnosis of IBD and sub-classification as CD or UC were

determined by the use of standard diagnostic criteria (Lennard-Jones, 1989).


## 2.2.2. Oxford TDT cohort

The Oxford TDT cohort was collected under the same ethical approval and diagnostic

criteria as the Oxford linkage cohort (Section 2.2.1.). The population contained 333

IBD simplex families, consisting of 167 UC families, 156 CD families and 10 IC

families. A subgroup (TDT1) of 124 families from the cohort, consisting of 83 UC

families, 38 CD families and 3 IC families, was used for the association study with

chromosome 3 STR markers. Another subgroup (TDT2) of 185 families from the

cohort, consisting of 121 UC families, 58 CD families and 6 IC families, was used for

the first stage of the GNAI2 association study. The full cohort (TDT3) was used for the

second stage of the GNAI2 association studies and all other candidate gene association

studies. Each subsequent, larger population was an extension of the previous

population.

| Family Type/ Affected Sibs | CD (ASPs) | UC (ASPs) | Mixed (ASPs) | IBD (ASPs) |
|---|---|---|---|---|
| CD/2 | 94 (94) | - | - | 94 (94) |
| CD/3 | 13 (39) | - | - | 13 (39) |
| UC/2 | - | 72 (72) | - | 72 (72) |
| UC/3 | - | 5 (15) | - | 5 (15) |
| UC/2G-2 | - | 1 (2) | - | 1 (2) |
| Mixed/2 | 2 (2) | - | 40 (40) | 42 (42) |
| Mixed/3 | 5 (7) | 1 (1) | 5 (10) | 6 (18) |
| Mixed/2G-2 | 1 (1) | - | 1 (1) | 1(2) |
| Total | 115 (143) | 79 (90) | 46 (51) | 234 (284) |

**Table 2.1.  Summary of IBD Cohort Used for Linkage Analysis**

"2G-2" denotes an extended family with two generations of ASPs.  Mixed family type refers to any family with a least one first- or second-degree relative with a discordant form of IBD.  Mixed ASPs denotes any sibling pair with discordant forms of IBD.

## 2.3. Reagents

### 2.3.1. General

Ethanol and isopropanol used for DNA extraction and sequence fragment clean up were purchased for EM Industries, Inc. The 10X TBE used for all forms of gel preparation and electrophoresis was from National Diagnositics Ltd.

### 2.3.2. DNA extraction, quantification and storage

RBC lysis solution, cell lysis solution, Rnase A solution, protein precipitation solution, and DNA hydration solution were part of the Purgene® DNA isolation Kit from Gentra Systems, Inc. PicoGreen® dsDNA quantification reagent was from Molecular Probes, Inc. The chloros industrial used to clean the DNA dispensing robot was purchased from Woburn Chemicals Ltd.

### 2.3.3. Oligonucleotides

Fluorescent labelled oligonucleotides including their corresponding unlabeled 'reverse' oligonucleotide were synthesised at a 0.2μM scale by Oswel Ltd. All other oligonucleotides were synthesised either by Oswel Ltd. at a 0.2μM or 40nM scale or by MWG Biotech, Inc. at a 0.01μmol scale.

### 2.3.4. PCR

AmpliTaq Gold™ DNA polymerase, AmpliTaq™ DNA polymerase, GeneAmp® 10X PCR buffer II, and $MgCl_2$ solution were part of the GeneAmp® PCR kits from PE Applied Biosystems, Inc. The dNTPs were purchased from Amersham Pharmacia Biotech UK Ltd. Gelatin and mineral oil were purchased from Sigma-Aldrich Company Ltd.

### 2.3.5. Agarose gel preparation

UltraPURE® electrophoresis grade agarose from Life Technologies Ltd. was used for the preparation of all agarose gels. The ethidium bromide was purchased from Sigma-Aldrich Company Ltd.

### 2.3.6. Polyacrylamide gel preparation

Acrylamide/bis 29:1 4% C for 36cm genotyping gels and urea were purchased from Bio-Rad Laboratories Ltd. National Diagnostics Ltd supplied Sequagel XR for 48cm sequencing gels. Temed and mixed bed resin were received from Sigma-Aldrich Company Ltd. The APS was purchased from Amresco, Inc.

### 2.3.7. Size standards and loading buffers

Glycerol and bromophenol blue used to make 5X loading buffer for agarose gel electrophoresis was from Sigma-Aldrich Company Ltd. The 100bp ladder was purchased from Life Technologies Ltd. PE Applied Biosystems, Inc. supplied the Genescan – 500 [TAMRA] size standard and loading buffer for STR fluorescent genotyping. Demonised formamide was purchased from Amresco, Inc. Formamide loading dye for the electrophoresis of sequencing samples was part of the DYEnamic™ ET terminator cycle sequencing kit from Amersham Pharmacia Biotech UK Ltd.

### 2.3.8. Polymorphism detection

The acetonitril and 10X TEAA used to make buffer A and B for dHPLC were purchased from Fisher Scientific UK Ltd. and Transgenomic, Inc., respectively.

## 2.3.9. Cycle sequencing including pre and post sequencing purification

Exonuclease I, SAP, and the DYEnamic™ ET terminator cycle sequencing kit including the sequencing reagent premix were purchased from Amersham Pharmacia Biotech UK Ltd. Fisher Scientific UK Ltd. and EM Industries, Inc. respectively supplied the tris and hydrochloric acid used to make Tris HCL. The fine Bio-Gel® P-10 polyacrylamide gel used to make P-10 gel solution was from Bio-Rad Laboratories Ltd. Sodium acetate used in the ethanol precipitation of sequencing fragments was purchased from Sigma-Aldrich Company Ltd.

## 2.3.10. Restriction endonucleases

New England Biolabs Ltd supplied all restriction endonucleases used for genotyping RFLPs.

## 2.3.11. RH panels

The Stanford G3 and MIT/Whitehead Genebridge 4 RH panels were purchased from Research Genetics, Inc.

# 2.4. Solutions

**5X loading buffer**

0.05%          bromophenol blue

30%          glycerol

**1M Tris HCl (pH 7.6)**

1M          tris

pH 7.6          hydrochloric acid

**P-10 gel solution**

7.5g            Bio-Gel® P-10 Polyacrylamide Gel, fine

up to 100ml    1M Tris HCl (pH 7.6)


**Buffer A**

1X            TEAA


**Buffer B**

25%            acetonitril

1X            TEAA


## 2.5. Accession numbers and URLs


**Online Mendelian Inheritance in Man (OMIM)**

http://www.ncbi.nlm.nih.gov/Omin/

*IBD1* [MIM 266600]

*IBD2* [MIM 601458]

*IBD3* [MIM 604519]


**National Center for Biotechnology Information**

http://www.ncbi.nlm.nih.gov


**Marshfield Medical Research Foundation, Center for Medical Genetics**

http://www.marshmed.org/genetics/

**Stanford RH Server**

http://www.shgc.stanford.edu/RH/index.html

**Whitehead Institute/MIT Center for Genome Research RH server**

http://carbon.wi.mit.edu:8000/cgi-bin/contig/rhmapper.pl

**RHMAP**

http://carbon.wi.mit.edu:8000/ftp/distribution/software/rhmapper/RHMAPPER-

1.1/rhmap/

**RepeatMasker2**

http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker

**Pedcheck (version 1.1)**

http://watson.hgen.pitt.edu/register/soft_doc.html

**MultiMap and CRI-MAP**

ftp://ftp.ebi.ac.uk/pub/software/linkage_and_mapping/MULTIMAP/

**GENEHUNTER (version 2.0)**

http://www.genome.wi.mit.edu/ftp/distribution/software/genehunter

# Chapter 3

# 3. Results

## 3.1. Chromosome 3p extension study with STR markers

### 3.1.1. Markers and oligonucleotides

STR markers were selected across a 40cM area around D3S1573, which was the marker

showing suggestive linkage to IBD in the Satsangi et al. (1996) genome scan (Section

1.4.3.2.). In total, 48 markers were chosen (Section 2.1.5.1.), consisting of 24 markers

p-terminal to D3S1573, 23 markers p-centromeric to D3S1573 and D3S1573 itself

(Table 3.1.). A pair of oligonucleotides was designed (Section 2.1.5.2.) to amplify

across the repetitive element of each marker (Table 3.2.). In order to choose a

fluorescent label for each marker, the markers were placed into putative panels (Section

2.1.5.3.) based upon the length of each amplicon according to the associated GenBank

sequence (Table 3.3.). The $T_A$ of each oligonucleotide for each marker was calculated

(Section 2.1.1.6.). The $T_A$ at which each marker was initially amplified by PCR was  –

calculated by taking the lowest $T_A$ of a pair of oligonucleotides and rounding down to

the nearest even integer. The $T_A$ of each oligonucleotide and the marker PCR $T_A$ are

shown in Table 3.2.

### 3.1.2. PCR parameter optimisation and amplification across RH panels

The 48 STR markers were amplified by PCR across the G4 and G3 RH panels using the

PCR $T_A$ calculated in the previous section and 30 cycles (Section 2.1.2.1.). The results

| Marker | Map position | Het | Marker | Map position | Het | Marker | Map position | Het |
|--------|-------------|-----|--------|-------------|-----|--------|-------------|-----|
| D3S1266 | 0 | 0.73 | D3S3559 | 15.34 | 0.85 | D3S3717 | 21.73 | 0.79 |
| D3S3547 | 2.51 | 0.74 | D3S3624 | 15.87 | 0.72 | D3S2408 | 21.73 | 0.69 |
| D3S1609 | 2.51 | 0.64 | D3S3563 | 15.87 | 0.77 | D3S1606 | 21.73 | 0.78 |
| GATA8A07 | 2.51 | 0.71 | D3S3678 | 15.87 | 0.76 | D3S3588 | 22.26 | 0.63 |
| D3S3727 | 3.47 | 0.79 | D3S3597 | 15.87 | 0.74 | D3S3724 | 22.79 | 0.75 |
| D3S3567 | 4.09 | 0.73 | D3S3647 | 15.87 | 0.73 | D3S3721 | 22.79 | 0.71 |
| D3S2432 | 5.32 | 0.82 | D3S3582 | 16.58 | 0.67 | D3S3621 | 22.79 | 0.76 |
| D3S1211 | 5.32 | nil | D3S1767 | 17.29 | 0.71 | D3S3048 | 22.79 | 0.64 |
| GATA112E02 | 8.39 | 0.63 | D3S1581 | 18 | 0.88 | D3S3616 | 23.86 | 0.85 |
| D3S3518 | 8.39 | 0.78 | D3S3640 | 18 | 0.65 | D3S3532 | 24.39 | 0.65 |
| D3S1619 | 8.39 | 0.73 | D3S3615 | 18 | 0.62 | D3S2402 | 24.39 | 0.76 |
| D3S1611 | 8.92 | 0.66 | D3S1578 | 18 | 0.88 | D3S1540 | 24.39 | nil |
| D3S3623 | 8.92 | 0.76 | D3S3561 | 18 | 0.73 | D3S1236 | 24.39 | 0.53 |
| D3S3512 | 8.92 | 0.68 | D3S3560 | 18 | 0.66 | D3S1295 | 24.39 | 0.62 |
| D3S2417 | 8.92 | 0.68 | D3S1588 | 18 | 0.75 | D3S1592 | 24.39 | 0.67 |
| D3S1561 | 8.92 | 0.63 | D3S1568 | 18 | 0.88 | D3S3553 | 24.39 | 0.86 |
| D3S1612 | 8.92 | 0.68 | D3S1621 | 18 | 0.73 | D3S1514 | 24.39 | nil |
| D3S3718 | 8.92 | 0.71 | D3S2321 | 18 | nil | D3S3722 | 26.02 | 0.82 |
| D3S1277 | 8.92 | 0.82 | D3S3604 | 18 | 0.67 | D3S1766 | 26.02 | 0.71 |
| D3S1768 | 8.92 | 0.7 | D3S2420 | 18 | 0.77 | D3S1547 | 26.02 | 0.69 |
| D3S1298 | 9.45 | 0.88 | D3S3629 | 18 | 0.77 | D3S1234 | 26.02 | 0.68 |
| D3S2411 | 9.45 | 0.39 | D3S3667 | 18 | 0.7 | D3S3577 | 26.02 | 0.74 |
| D3S3572 | 10.52 | 0.74 | D3S1573 | 18 | 0.79 | D3S1313 | 26.02 | 0.68 |
| D3S3639 | 10.52 | 0.62 | D3S2456 | 18 | 0.74 | D3S2452 | 26.02 | 0.61 |
| D3S3605 | 10.52 | 0.73 | D3S3729 | 18 | 0.44 | D3S1300 | 27.7 | 0.83 |
| D3S3527 | 10.52 | 0.84 | D3S3688 | 18 | 0.64 | D3S3631 | 29.62 | 0.64 |
| D3S1260 | 10.52 | 0.65 | D3S3648 | 18 | 0.25 | D3S1239 | 29.62 | 0.29 |
| D3S3593 | 10.52 | 0.74 | D3S2384 | 18 | 0.75 | D3S1312 | 29.62 | 0.75 |
| D3S3521 | 10.52 | 0.83 | D3S1235 | 18 | 0.43 | D3S3566 | 31.48 | 0.5 |
| D3S3522 | 12.66 | 0.56 | D3S3026 | 18 | 0.72 | D3S3698 | 32.29 | 0.71 |
| D3S3658 | 12.66 | 0.58 | D3S1289 | 18.8 | 0.81 | D3S1600 | 33.33 | 0.72 |
| D3S3564 | 15.34 | 0.78 | D3S3666 | 19.6 | 0.85 | D3S1287 | 35.96 | 0.67 |
| D3S3685 | 15.34 | 0.89 | D3S3672 | 19.6 | 0.66 | D3S4542 | 37.27 | 0.76 |
| D3S2319 | 15.34 | nil | D3S1582 | 19.6 | 0.8 | D3S3571 | 37.27 | 0.91 |
| D3S3687 | 15.34 | 0.4 | D3S3660 | 19.6 | 0.73 | D3S3635 | 37.27 | 0.69 |
| D3S2304 | 15.34 | nil | D3S1613 | 19.6 | 0.83 | D3S3644 | 38.53 | 0.64 |
| D3S2407 | 15.34 | 0.75 | D3S3719 | 19.6 | 0.82 | D3S1285 | 38.53 | 0.74 |

**Table 3.1.  STR markers selected for the 3p extension study**

The peak linkage marker (D3S1573) from the Satsangi et al (1996) genome scan is indicated in blue.  The markers selected for the extension study include D3S1573 and the markers indicated in red.  The marker order, distances and heterozygosity were retrieved from the chromosome 3 Marshfield sex-average genetic map.  The map position of each marker is relative to its distance from the most p-terminal marker chosen for the study.

| Marker | Forward Oligonucleotide | $T_A$ (C) | Reverse Oligonucleotide | $T_A$ (C) | PCR $T_A$ (C) |
|---|---|---|---|---|---|
| D3S1234 | CCTGTGAGACAAAGCAAGAC | 55 | GACATTAGGCACAGGGCTAA | 55 | 54 |
| D3S1266 | ATGAGGGGTAATGTTGGAGG | 55 | GCTATGAAGGAAAATTAAGCATC | 52 | 52 |
| D3S1277 | CCAGATTAGGTGATCTGAGC | 55 | AACATTNCTTCTGCCTCCCC | 56 | 54 |
| D3S1285 | GAAAACCCATACAGCATGGC | 55 | TGTTCATCACAGGGGTAGC | 55 | 54 |
| D3S1295 | ACACACATACATGCATGCCC | 55 | AAGTTTTGATACCCACCCATG | 54 | 54 |
| D3S1298 | GTTGGTCTCTCCAGATACC | 55 | AGCCTTGTGGTCTCTTTACC | 55 | 54 |
| D3S1547 | CCTTTGTTACACACGCATGC | 55 | GGATTTGGCATTCTTTATAAGGG | 54 | 54 |
| D3S1568 | ATGAACAGAACCTCCCTAGG | 55 | AGGGACCTCTNAAGAGAACC | 56 | 54 |
| D3S1573 | TATACATACCCCTCCAACGC | 55 | CTGCCAGTAAATNACAGGGC | 56 | 54 |
| D3S1578 | GCCAAATCAGAGACAACAGC | 55 | GCCTGAAAGCCTTGATGAC | 55 | 54 |
| D3S1581 | CTGACAGAACTGCCAAACC | 55 | AGTTCAAGGCTGCAGTGAG | 55 | 54 |
| D3S1582 | GGTACTATGAAAGCCTGTAGC | 56 | GAAAATCCCAGAGGAACAGC | 55 | 54 |
| D3S1588 | TACGATGGTCAAAACAGAAGC | 54 | TCATTGGGTAGGGACTGC | 55 | 54 |
| D3S1592 | CAGTATTGGAAGTCCTAGCC | 55 | GTGTGTTGATTTCGTATCCTGC | 56 | 54 |
| D3S1600 | ATCACCATCATCTGCCTGTC | 55 | AAGAGGTGAAGGTTGCTTGC | 55 | 54 |
| D3S1606 | GCGCAACATAGTGAAATCCC | 55 | ACCATGCCTGGCTACTTTG | 55 | 54 |
| D3S1613 | CTGTGATAAGGACCAAGGC | 55 | TTCAGTGAACAGAGCAAATTGC | 54 | 54 |
| D3S1619 | TCTCTCTGTGGAAACCTTCC | 55 | GTTTTCACACGTGAAGGG | 53 | 52 |
| D3S1621 | TGGAGCCCTTTGAGAAATGC | 55 | CTCCCAAGGAAGGGTTTTAC | 55 | 54 |
| D3S1766 | ACTTCTCAGCCTCCATGAC | 55 | GGTGTTGTTACCTAGAAGGG | 55 | 54 |
| D3S1767 | GCTGTATTTCAACACACAGGC | 56 | TCCCAGCTACTTGAGAGG | 55 | 54 |
| D3S1768 | TATTTGTGCTGGTTGCTGCC | 55 | CCACTGTGATTTGCTGTTGG | 55 | 54 |
| D3S2384 | GTGACAGAGACTCTTGTCTC | 55 | ATCCATTCTGGTTGGGGAAC | 55 | 54 |
| D3S2408 | CTGCTAACACTGACAGTGC | 55 | ACACAGATACCAATGGGTGG | 55 | 54 |
| D3S2420 | ACAAGTGCGAAACTCTGCC | 55 | ATGACAATAGTCCAGGAGCC | 55 | 54 |
| D3S2452 | TTGCAGATAGCAGATCGTGG | 55 | TAAGATTAGTCAGGGCTCGC | 55 | 54 |
| D3S2456 | GGATATCCAGTTTTCCCAGC | 55 | TGATTGTGCCATTGCACTCC | 55 | 54 |
| D3S3518 | AAGGCTCCTTTCGGGATATG | 55 | ATTTGTGCAGGGTCCTAGAG | 55 | 54 |
| D3S3521 | TCACACAGTAAGTCCATGCC | 55 | GTTAGCAGCAGTTACCTTGG | 55 | 54 |
| D3S3527 | GATCGGCCATATGTTGGTAC | 55 | GTGGTCTGAGAAGATACTTG | 53 | 52 |
| D3S3559 | AGCTCTACATCAGGCAACC | 55 | TTCATGGCTTTGGGTGGTTC | 55 | 54 |
| D3S3564 | GCTAAACACAGTCTAACTGC | 53 | CTCTCTTCCTTACATGAGGC | 55 | 52 |
| D3S3571 | CTGAGGCTCTGCAATGTTTAG | 56 | TTGCCAGGCAAGGTTATCTG | 55 | 54 |
| D3S3577 | AAGCACATTCCCTCCATCC | 55 | TTGCCATCCCTGACTGAC | 55 | 54 |
| D3S3616 | TCAGAGACACTCAAGTGACC | 55 | TTAAGCCCCTGGGAAGTTAC | 55 | 54 |
| D3S3621 | GTTTCTGAGAGAGAAGAGTGC | 56 | GGGACCTTGAGCCTAAAAAG | 55 | 54 |
| D3S3623 | CCATGTTGGTTAAAGGCAAGG | 56 | TGAACTGAAGTGACCTCCTG | 55 | 54 |
| D3S3629 | TTCAAGCAGATGTTGCTGCC | 55 | TGGAATTACCTGGATTGCCC | 55 | 54 |
| D3S3635 | TTCATTTTCCAAACATGGTGAG | 52 | ATGTGCATAGCCAGTTTTAGC | 54 | 52 |
| D3S3640 | GATCGCGTGACATTCC | 52 | GCTACTTGCTATTTATCAGACC | 54 | 52 |
| D3S3647 | ATCTCAGGCTCAGAGCAG | 55 | TAAAATGTGTTCCCAGNCCC | 54 | 54 |
| D3S3660 | AGGCCTCTAGACAGTAAGC | 55 | CACATGCTGCTTTTGTTGGC | 55 | 54 |
| D3S3666 | TTTCACTAGTNGGGGACTGG | 56 | CAAATACCCTGTAAGGCACTG | 56 | 56 |
| D3S3672 | GAATCATGTGTGCGTGATTGC | 56 | GACATTACTAGGGTTTTGTGG | 54 | 54 |
| D3S3678 | TTTCCCTCAGCAGGACCAC | 57 | AGGTGGGCTCATTCTCATGC | 57 | 56 |
| D3S3685 | AACAACCCAGCTCTTTGTGG | 55 | GACAATACAGGTAACCCTGC | 55 | 54 |
| D3S3719 | CATGCAGAATCTCATGGGTTC | 56 | ATAACTCTGCTCCTAACGGC | 55 | 54 |
| D3S3727 | TAAAATNCCTAGTACGGGCC | 53 | TGGTAGGTAGTTCCAGTGTG | 55 | 52 |

**Table 3.2.  The sequence and annealing temperature of oligonucleotides designed for amplification of chromosome 3p STR markers**

| Panel 1 | | | | Panel 3 | | |
|---|---|---|---|---|---|---|
| Marker | Est. Length | Label | | Marker | Est. Length | Label |
| D3S1547 | 88 | FAM | | D3S1606 | 98 | FAM |
| D3S1573 | 122 | FAM | | D3S2456 | 140 | FAM |
| D3S2408 | 186 | FAM | | D3S1298 | 191 | FAM |
| D3S3647 | 217 | FAM | | D3S1285 | 226 | FAM |
| D3S1277 | 268 | FAM | | D3S3577 | 297 | FAM |
| D3S1621 | 107 | HEX | | D3S1234 | 111 | HEX |
| D3S3672 | 143 | HEX | | D3S1578 | 147 | HEX |
| D3S1766 | 232 | HEX | | D3S1768 | 208 | HEX |
| D3S3719 | 298 | HEX | | D3S3629 | 255 | HEX |
| D3S2452 | 151 | TET | | D3S2384 | 122 | TET |
| D3S1588 | 208 | TET | | D3S3621 | 217 | TET |
| D3S3571 | 260 | TET | | | | |

| Panel 2 | | | | Panel 4 | | |
|---|---|---|---|---|---|---|
| Marker | Est. Length | Label | | Marker | Est. Length | Label |
| D3S3635 | 98 | FAM | | D3S3527 | 103 | FAM |
| D3S3640 | 136 | FAM | | D3S1581 | 141 | FAM |
| D3S1767 | 220 | FAM | | D3S3559 | 198 | FAM |
| D3S1592 | 273 | FAM | | D3S3521 | 236 | FAM |
| D3S1295 | 107 | HEX | | D3S3727 | 124 | HEX |
| D3S3666 | 147 | HEX | | D3S3518 | 183 | HEX |
| D3S1600 | 203 | HEX | | D3S3564 | 216 | HEX |
| D3S3660 | 244 | HEX | | D3S3678 | 240 | HEX |
| D3S1266 | 304 | HEX | | D3S1619 | 135 | TET |
| D3S2420 | 113 | TET | | D3S3685 | 180 | TET |
| D3S1582 | 173 | TET | | D3S3623 | 214 | TET |
| D3S3616 | 210 | TET | | D3S1568 | 252 | TET |
| D3S1613 | 263 | TET | | | | |

**Table 3.3. Putative panels for the fluorescent genotyping of 3p STR markers**

The estimated length of each marker was calculated using the Genebank sequence and the oligonucleotides in table 3.2. The colour of the background for each fluorescent label corresponds to the colour used to display the respective dye in both ABI Genescan and Genotyper software when using filter set C.

were visualised and scored as described in Sections 2.1.2.2. and 2.1.2.3. The scored

results for the G4 and G3 RH panels appear in Figures 3.1 and 3.2, respectively.


All markers exhibited high-fidelity amplification across the RH panels (Figure 3.3.) at

the predicted PCR conditions except D3S1592, D3S2456, D3SS518, D3S3564,

D3S3666, and D3S3727. D3S3564 failed to amplify even at a $T_A$ of 48C and 40 cycles

and was thus removed from the study. The amplification of D3S1592 and D3S2456

resulted in spurious fragments ranging from 100bp to 400bp. Increasing the $T_A$ for the

amplification of these markers failed to remove the non-specific bands and thus the

markers were also removed from the study. The initial amplification of the other three

markers was of either weak intensity or failed completely. These markers were

subsequently amplified with altered PCR parameters of lower $T_A$'s in increments of 2C

down to 48C and increased cycles in increments of 5 additional cycles up to 40 cycles.

After optimisation, D3S3666 amplified well with a PCR $T_A$ of 52C and 40 cycles and

D3S3518 and D3S3727 amplified well with a PCR $T_A$ of 50C and 40 cycles.


### 3.1.3. RH map construction and comparison

The G3 and G4 data from the 45 remaining 3p STR markers were sent to external RH

servers for analysis against existing framework maps (Section 2.1.2.3.1.-2.). Based

upon the positions and distances of the submitted markers amongst the framework

markers, RH maps including only the submitted markers were generated. All markers

submitted were mapped to the framework maps except D3S3727. The G4 data for

marker D3S3727 was insufficient to map it to a single chromosome and it was therefore

excluded from the resulting G4 framework map used in this project.

```
D3S1234   0000001010101000011011000010110100100010001011100010101000010110000100101010010000000011001
D3S1266   0000000010200000011010000001000100000000000102000000001000000202000000102000000000010001011
D3S1277   1000000010100000001110000011200100001000000101100010001000000001000000000010000000000000011001
D3S1285   0000001010101000211011000020110100101010201001000101010000101100000100101001000000000011001
D3S1295   0000000010101000011011000010110100101010001011100010101000010110000100101210010000000011001
D3S1298   1000000010100000001210000011000100102000000101100010001000010011000010010101000000000011001
D3S1547   0000002010101000011011000010110100100010001011100010101000010110000100101010010000000011001
D3S1568   1000000010100000011010000001110100110011001010110000100010000101100001001010100100000000011000
D3S1573   1000000010100000011010000012110100111011001011101010001000010110000100111010010000000011000
D3S1578   0000001010101000011111000011110100111010101011100010001000010110000100101001000000000011001
D3S1581   1000000010000000011010000011110100110010001001000010001000100000011000010011011001000000011001
D3S1582   0000000010101000011011000021110100111010021011200010001000010111000010010101001000000011001
D3S1588   0000001010101000011111000011110100111010101011100010001000010110000100101001000000000011001
D3S1600   0000001010100000011011000010110100101010010001011100010101000010110000110101010010000000011001
D3S1606   0010000010101000011011000010110110111011110101110001000100001011200010010101001000000011001
D3S1613   0000000010101000011011000011110100111010000101110001000100001011100010010101001000000011001
D3S1619   1000000010100000001110000011000100001000000101100010001000000011000000010100000000000011001
D3S1621   1000000010100000011010000011101001120110010110001000100010000100110000100111010010000000011000
D3S1766   0000000010101000011011000010110100100010001011100010101000010110000100101010010000000011001
D3S1767   1000000010000000001010000011100100100010001001001000010001000000011000010011011001000000011001
D3S1768   1000000010100000001110000011000100001000100101100010001000000011000000000001000000000000011002
D3S2384   1000000010000000011010000001110100110000010010000010001000000011000010010101101000000011000
D3S2408   0010000010101000011011000010110110111010001011100010001011100010010101001000000011001
D3S2420   1000000010000000011010000011110100120010001001000210001000000011000010011011001000000011001
D3S2452   0000001010101000011011000010110100100010001011100010101000010110000100101010010000000011001
D3S3518   1000000010100000001110000011000100001000000101100010001000000011000000010100000000000011001
D3S3521   1000000010100000001210000011000100100000022000100010001000020110000100100101010000000011001
D3S3527   1000000010100000001110000011000100000000000102100010001000010110000100101010000000000011001
D3S3559   1000000010100000001110000011000100101000000001001010011000001100001001010100000000000011001
D3S3571   0000001000101000011001000000110100101010001011100010100010110001001010100100000000010000
D3S3577   0000001010101000011011000010110100001010101011100010101000010110000100101010010000000011001
D3S3616   0000001010101000011011000010110100101010001011100010101000010110000100101010010000000011001
D3S3621   0000000010101000011011000010110100101010001011100010001000010110000100101010010000000011001
D3S3623   1000000010100000001110000110000100201000001011000100010001000001100000001010100000000000011001
D3S3629   1000000010000000011010000011101001100100010011000100000010000101100010001110110100000000011000
D3S3635   0000001010101000011011000010110100101010001011100010101000010110000100101010010000000011001
D3S3640   1000000010000000011010000011110100110010011001001000010001000000110000100111011001000000011001
D3S3647   1000000010100000001010000011000100010010000001001000010001000000110000100101010000000011001
D3S3660   0000000020101000011021000011110100111010001011100010001000010111000010010121200100000000011001
D3S3666   0000000010100000011011000011101001000000010110001000100010001011000010010101001000000011001
D3S3672   0000000010100000011011000011110100111010001011100010001000010111000010010101001000000011001
D3S3678   1000000010100000001110000110001001010000000100101000100000001100001001010100000000000011001
D3S3685   1000000010100000001110000110001001000100010000000100100110000001100000100101010000000011001
D3S3719   0010001010101000011011000010110100111011110101110001000100001012100010010100001000000011001
D3S3727   0000000010000000010100000100010000000000001001000100010000000010100000010100000000000100011001
```

**Figure 3.1.  G4 RH panel results for 3p STR markers**

For each data vector, the 93 human-hamster hybrids are ordered in the official G4 order.

The results are coded 0 for negative, 1 for positive and 2 for ambiguous.

```
D3S1234 000000011000100000000000000000001000000000001000000010000000000001000100100000101 0001
D3S1266 000000000000000010000000000000000000000001000000110000000000001000000001R000000010
D3S1277 100000000000000010000000000010000000000001100000010010R00000001000000000000001001010000
D3S1285 100000000000000000000000000000001000001000000111000000000000000010101000000001000000
D3S1295 000000011000000010010000000000100001010010100000000000000000000001000000100100101 0000
D3S1298 000000000000000000000000000000001000100001110000011000000001000010000000001001000000
D3S1547 000000011000100000000000000001000000000001000000010000000000001000000100000101 0001
D3S1568 000000011000000000000000000000011000010000010000000000000110010000000001000001
D3S1573 000000011000000010000000000000000011000010000000000000000000110100000000100 0001
D3S1578 000000010000000010000000000000001110R100100000000000000000010101001000001000000
D3S1581 000000010010000000000000000000000010000011000001000000000010000101101100000001000010
D3S1582 000000010000000110000000000000011001100110010100001001100000010101010010010010 01000000
D3S1588 000000010000000010000000000000010001110010010000000101000000000010101001000001000000
D3S1600 000000010000000000100000000000000000000010001111010000000010000100010010010000001010000
D3S1606 000001010000000010010000000000100011010010100000001000000100001010100100110 1000000
D3S1613 000000010000000010000000000000010011100110100000001000000000010101010010010010 01000000
D3S1619 100000000000000110100001001010000011000000000001001010000011000000000000001001010000
D3S1621 000000011000000000000000000000001100001000010000000000000011101000000001000001
D3S1766 00000001R0000000000100000000001000010100001000000010000000000001000000100000101 0001
D3S1767 00000001000001000001000000000000001000001000000000000000R000010100100000001000010
D3S1768 100000000000000010000000000100001000001100000010010100000110000001000001001010000
D3S2384 000000000010000000000000000001000000100001000001001000000010000100101000000 01000001
D3S2408 000001010000000010010000000001000110100101000000010000000100001010100100110 1000000
D3S2420 000000100100010000000000000000000000000001000100000000100001000R10110110000001 000010
D3S2452 000000011000000000100000000010000101000010100000010000000000100000001001010001
D3S3518 000000000000000010000000000000001101000010010000000100000000000011101000000
D3S3521 000000000000000000000000000001000100001110 0R0010000000000000001000000100100 1000000
D3S3527 000000000000000000000000000001000100001110000010000000000000001010000100100 1000000
D3S3559 000001001000010000011000000000000100000111000100000000000001010000100001000010
D3S3571 100000010000000000000000000000000000000000001111010000000100001010100000001000000
D3S3577 000000011000100000000000000010000000000001000000R10100000000001000010100100101 0001
D3S3616 000000010000000010000000000001000010100101000000010000000000001000000100100 1000000
D3S3621 00000111000000001000000000000100011010010100000010101000000010000110011011 01000000
D3S3623 0000000100000000000000000001010100001110000010000000000100001000100000100 1000000
D3S3629 000000010000000000000000001000001100010000010000000010000101101000000 01000001
D3S3635 100000010000000000000000000000000000001111010000000100001010100000001000000
D3S3640 00000001001001001000000000000000000010000000000010000100001111011000000R000010
D3S3647 000001001000000000010000000001000100000110000000000000010000010000001000001000010
D3S3660 000000010000000010000000000010001100R1101000000100000000010101001001001000000
D3S3666 000000010000000010000000000001100110001001000001001000000101010010 1000001000000
D3S3672 00000001000000011000000000001000110010010000000100000000101010010010000 1000000
D3S3678 000001001000010000010000000001000R000011000000000000100001000001000001000010
D3S3685 000001000000000000010000000000001000001110011000000000000001010000100001000010
D3S3719 0000010100000000100100000000010001100010100000001000000100101010010010011 01000000
D3S3727 0000000000000000010000000000000000000010R100010110000000000010000000010000000100
```

## Figure 3.2. G3 RH panel results for 3p STR markers

For each data vector, the 83 human-hamster hybrids are ordered in the official G3 order.

The results are coded 0 for negative, 1 for positive and R for ambiguous.

**Figure 3.3.  A typical high-fidelity PCR amplification across the G3 RH panel**

The marker was amplified in duplicate across the panel with identical lanes loaded

adjacent to each other.  Samples in rows A and B of a 96-well plate were loaded A1-12

and B1-12 on the first row of the agarose gel.  Rows C and D follow, etc.  100bp size

standard was loaded in the first well of each row to size the fragments.

Both sets of RH data were also imported and analysed in the RHMAP program as described in Section 2.1.2.3.3. However, due to the low resolution of the G4 RH panel, only a G3 non-framework RH map of the STR markers could be generated.

The centiRay distances of the RH maps were converted into centiMorgans (Section 2.1.2.4.) for subsequent map comparisons and linkage analysis. The resulting G3 RHMAP and the G3 and G4 framework RH maps are shown in Table 3.4. The three maps were compared to the Marshfield sex-average genetic map for overall length and map position and the Whitehead Institute/MIT YAC contig physical map for marker order (Figure 3.4.). Of the three maps, the G3 RH map generated by analysis within the RHMAP program was closest in size and map position to the genetic map and had the least number of marker inversions compared to the physical map. Thus, the G3 RHMAP RH map was chosen for subsequent use in linkage analysis.

### 3.1.4. Panel re-design and pooling strategy

Based upon a comparison of the intensity of the DNA bands on the RH mapping PCR agarose gel images, the cycling parameter of the PCR protocol was increased to 35 for 18 of the markers and to 40 for 5 others. With these new PCR conditions, the 45 remaining STR markers were analysed by acrylamide gel electrophoresis to design an optimal pooling strategy as described in Section 2.1.5.10. From the acrylamide gel electrophoresis results, it was observed that the intensities of markers D3S2408 and D3S3666 were insufficient for them to be reliably genotyped, even with 40 cycles of PCR and a maximum pooling volume. Thus, these two markers were removed from subsequent experiments.

| Genebridge 4 RH Server | | | Stanford G3 RH Server | | | Stanford G3 RHMAP | | |
|---|---|---|---|---|---|---|---|---|
| Marker | Map Position | | Marker | Map Position | | Marker | Map Position | |
| | $cR_{3,000}$ | cM | | $cR_{10,000}$ | cM | | $cR_{10,000}$ | cM |
| D3S1266 | 0 | 0.00 | D3S1266 | 0 | 0.00 | D3S1266 | 0 | 0.00 |
| D3S1619 | 18.4 | 4.97 | D3S3727 | 93 | 2.33 | D3S3727 | 33.4 | 0.84 |
| D3S3518 | 18.4 | 4.97 | D3S3518 | 208 | 5.20 | D3S3518 | 96.5 | 2.41 |
| D3S1277 | 19 | 5.13 | D3S1619 | 241 | 6.03 | D3S1619 | 216.1 | 5.40 |
| D3S1768 | 21.5 | 5.81 | D3S1768 | 242 | 6.05 | D3S1277 | 280.7 | 7.02 |
| D3S3623 | 25.4 | 6.86 | D3S1277 | 246 | 6.15 | D3S1768 | 295.5 | 7.39 |
| D3S1298 | 35.6 | 9.61 | D3S3623 | 385 | 9.63 | D3S3623 | 370.7 | 9.27 |
| D3S3527 | 38.8 | 10.48 | D3S1298 | 405 | 10.13 | D3S1298 | 391.8 | 9.80 |
| D3S3685 | 64.3 | 17.36 | D3S3521 | 516 | 12.90 | D3S3521 | 408.1 | 10.20 |
| D3S3559 | 64.3 | 17.36 | D3S3527 | 527 | 13.18 | D3S3527 | 413.4 | 10.34 |
| D3S3521 | 65.3 | 17.63 | D3S3685 | 605 | 15.13 | D3S3685 | 479.8 | 12.00 |
| D3S3678 | 65.9 | 17.79 | D3S3559 | 674 | 16.85 | D3S3559 | 499.4 | 12.49 |
| D3S3647 | 66 | 17.82 | D3S3678 | 697 | 17.43 | D3S3678 | 530.3 | 13.26 |
| D3S1767 | 81.2 | 21.92 | D3S3647 | 702 | 17.55 | D3S3647 | 535.1 | 13.38 |
| D3S2420 | 84.3 | 22.76 | D3S1767 | 831 | 20.78 | D3S1767 | 607.2 | 15.18 |
| D3S3640 | 84.5 | 22.82 | D3S2420 | 876 | 21.90 | D3S3640 | 656.8 | 16.42 |
| D3S1581 | 84.7 | 22.87 | D3S3640 | 877 | 21.93 | D3S2420 | 671.2 | 16.78 |
| D3S2384 | 100.4 | 27.11 | D3S1581 | 927 | 23.18 | D3S1581 | 702.7 | 17.57 |
| D3S3629 | 102.4 | 27.65 | D3S3629 | 948 | 23.70 | D3S2384 | 748.5 | 18.71 |
| D3S1621 | 110 | 29.70 | D3S2384 | 961 | 24.03 | D3S3629 | 775.2 | 19.38 |
| D3S1568 | 110 | 29.70 | D3S1568 | 1015 | 25.38 | D3S1568 | 809 | 20.23 |
| D3S1573 | 111.2 | 30.02 | D3S1621 | 1016 | 25.40 | D3S1621 | 824.8 | 20.62 |
| D3S3660 | 118.9 | 32.10 | D3S1573 | 1017 | 25.43 | D3S1573 | 840.6 | 21.02 |
| D3S1613 | 118.9 | 32.10 | D3S1578 | 1227 | 30.68 | D3S1578 | 918.9 | 22.97 |
| D3S1582 | 118.9 | 32.10 | D3S1588 | 1234 | 30.85 | D3S1588 | 932.1 | 23.30 |
| D3S3672 | 119.4 | 32.24 | D3S1582 | 1254 | 31.35 | D3S3672 | 949.3 | 23.73 |
| D3S1295 | 131.1 | 35.40 | D3S3672 | 1256 | 31.40 | D3S3666 | 961.7 | 24.04 |
| D3S1766 | 133.7 | 36.10 | D3S3666 | 1261 | 31.53 | D3S1582 | 981.6 | 24.54 |
| D3S1547 | 134.7 | 36.37 | D3S3660 | 1262 | 31.55 | D3S3660 | 1010.7 | 25.27 |
| D3S2452 | 135.3 | 36.53 | D3S1613 | 1263 | 31.58 | D3S1613 | 1010.7 | 25.27 |
| D3S1234 | 135.3 | 36.53 | D3S3719 | 1294 | 32.35 | D3S3719 | 1031.2 | 25.78 |
| D3S3635 | 135.4 | 36.56 | D3S2408 | 1301 | 32.53 | D3S2408 | 1038.6 | 25.97 |
| D3S3616 | 135.4 | 36.56 | D3S1606 | 1301 | 32.53 | D3S1606 | 1038.6 | 25.97 |
| D3S3577 | 137.2 | 37.04 | D3S3621 | 1319 | 32.98 | D3S3621 | 1072.1 | 26.80 |
| D3S1285 | 139.1 | 37.56 | D3S3616 | 1383 | 34.58 | D3S3616 | 1104.6 | 27.62 |
| D3S3719 | 145.9 | 39.39 | D3S1295 | 1423 | 35.58 | D3S1295 | 1124.7 | 28.12 |
| D3S1588 | 146.1 | 39.45 | D3S2452 | 1515 | 37.88 | D3S1766 | 1147.4 | 28.69 |
| D3S1578 | 146.1 | 39.45 | D3S1766 | 1524 | 38.10 | D3S2452 | 1155.7 | 28.89 |
| D3S1606 | 146.4 | 39.53 | D3S1547 | 1542 | 38.55 | D3S3577 | 1177.7 | 29.44 |
| D3S3621 | 146.9 | 39.66 | D3S3577 | 1543 | 38.58 | D3S1547 | 1191.1 | 29.78 |
| D3S2408 | 147.6 | 39.85 | D3S1234 | 1553 | 38.83 | D3S1234 | 1195.5 | 29.89 |
| D3S1600 | 148.2 | 40.01 | D3S1600 | 1694 | 42.35 | D3S1600 | 1318.4 | 32.96 |
| D3S3571 | 151 | 40.77 | D3S3571 | 1744 | 43.60 | D3S3571 | 1341.3 | 33.53 |
| D3S3666 | 152.4 | 41.15 | D3S3635 | 1744 | 43.60 | D3S3635 | 1341.3 | 33.53 |
| | | | D3S1285 | 1783 | 44.58 | D3S1285 | 1379 | 34.48 |

**Table 3.4.  RH and genetic maps generated from the RH mapping of 3p STR markers**

The map position of each marker is shown relative to the initial marker.  Pairs of markers shown in blue have identical RH map results and are thus mapped to the same

position. Pairs of markers shown in red are separated by a distance less than the

resolution of the RH panel (G4 = $3.7cR_{3,000}$, G3 = $10cR_{10,000}$) (Carey, 1997).
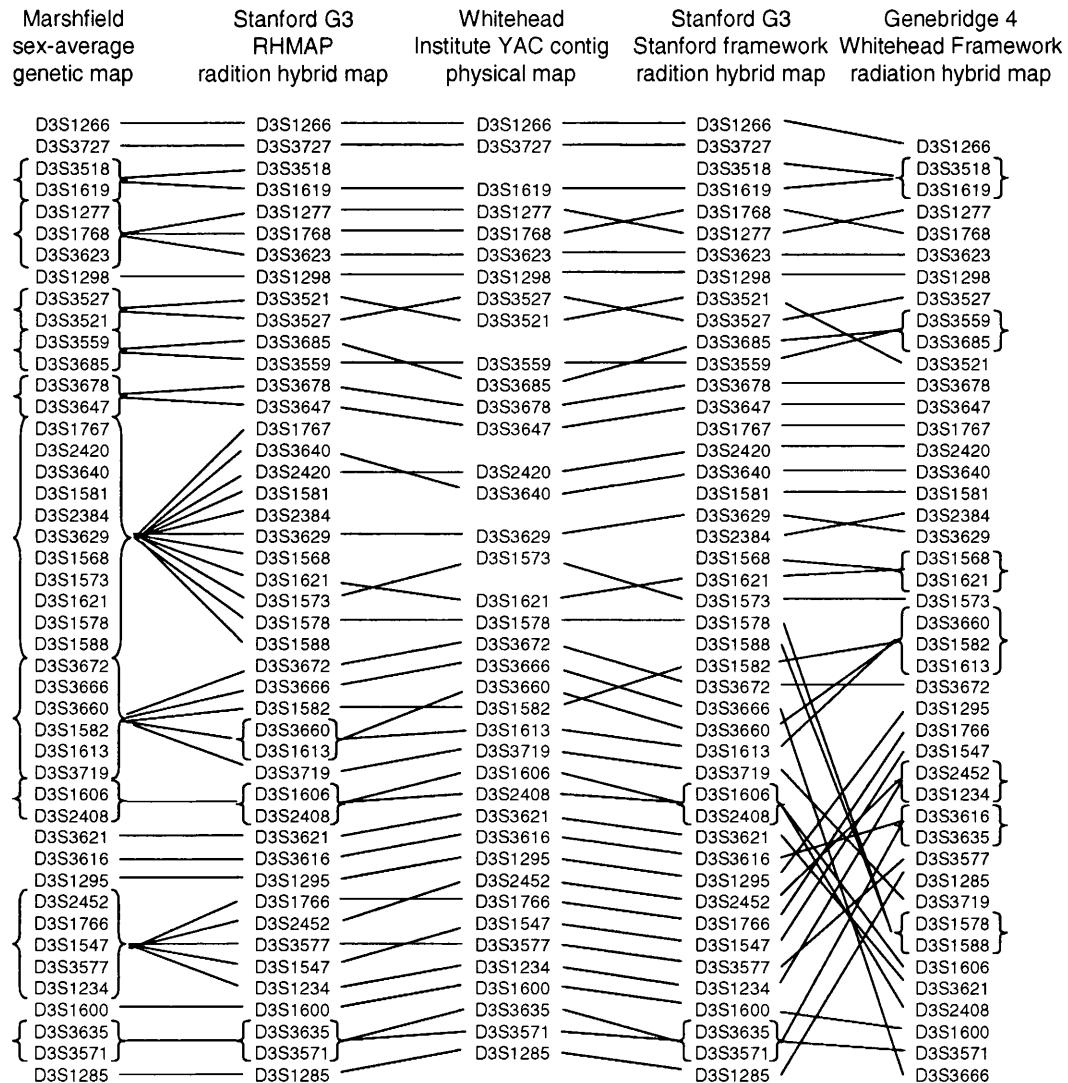


**Figure 3.4. Comparison of marker order between 3p STR marker maps**

Brackets surround markers map to the same position. Lines between maps connect

identical markers. The gaps between markers in the YAC contig physical map separate

different contigs. Starting from the top, the specific YAC contigs are named WC 3.5,

WC 3.6, WC 3.7, WC 3.8, WC 3.9 and WC 3.10.

Simultaneously, because the number of markers to be used for genotyping had decreased from 48 to 43 and because more size information for the markers was generated during RH mapping and acrylamide gel electrophoresis, new panels were designed prior to completing the optimal pooling ratios. However, upon analysing the new size ranges, it was observed that marker D3S3623 could not be placed into a panel without causing an overlap. Thus, D3S3623 was removed from further analysis. The new panels, including the pooling conditions and new PCR parameters for each marker are shown in Table 3.5.

### 3.1.5. Genotyping and quality checking

The 42 STR markers were genotyped (Sections 2.1.5.4.-9.) across the Oxford linkage cohort (Section 2.2.1.) generating approximately 36,000 genotypes (Figure 3.5.). Marker D3S1266 failed to amplify to the level required for robust genotyping, which lead to several missing genotypes. Subsequently, due to the lack of information for D3S1266 it was removed from the linkage study.

The remaining data was analysed with Pedcheck as described in Section 2.1.6.1. Unrelated subjects identified by Pedcheck analysis and DNA that failed to amplify resulted in the alteration of 13 pedigrees and the removal of 13 others (Table 3.6.). The description of the Oxford linkage cohort in Section 2.2.1. includes these changes. Three rounds of repeats were conducted which resulted in an overall genotyping success rate of 99.7% (Table 3.7.).

Recombination analysis using CRI-MAP (Section 2.1.6.2.) identified 48 potential genotyping errors, which were subsequently removed. Multiple recombination events were seen between family 248 consisting of 1 ASP and no parents, suggesting that the

| Panel 1 | | | | | Panel 3 | | | |
|---|---|---|---|---|---|---|---|---|
| Marker | Allele Range | PCR | Pooling | | Marker | Allele Range | PCR | Pooling |
| D3S1547 | 87-101 | 54/30 | 4 | | D3S1606 | 91-113 | 54/30 | 4 |
| D3S1573 | 114-136 | 54/35 | 4 | | D3S3647 | 212-236 | 54/40 | 8 |
| D3S1298 | 184-216 | 54/35 | 4 | | D3S1277 | 257-292 | 54/40 | 8 |
| D3S1767 | 216-246 | 54/30 | 4 | | D3S1621 | 91-125 | 54/30 | 8 |
| D3S1234 | 104-134 | 54/30 | 8 | | D3S1578 | 126-171 | 54/35 | 8 |
| D3S3672 | 138-161 | 54/30 | 8 | | D3S1768 | 200-229 | 54/30 | 8 |
| D3S1766 | 216-254 | 54/35 | 8 | | D3S3629 | 243-268 | 54/40 | 8 |
| D3S3719 | 294-311 | 54/35 | 8 | | D3S2384 | 114-146 | 54/35 | 4 |
| D3S2452 | 135-177 | 54/30 | 4 | | D3S3621 | 214-231 | 54/40 | 4 |
| D3S1588 | 199-226 | 54/30 | 4 | | D3S3571 | 231-288 | 54/35 | 4 |

| Panel 2 | | | | | Panel 4 | | | |
|---|---|---|---|---|---|---|---|---|
| Marker | Allele Range | PCR | Pooling | | Marker | Allele Range | PCR | Pooling |
| D3S3635 | 80-120 | 52/30 | 4 | | D3S3527 | 103-119 | 52/30 | 4 |
| D3S3640 | 129-163 | 52/30 | 4 | | D3S1581 | 133-161 | 54/35 | 4 |
| D3S1285 | 230-260 | 54/30 | 4 | | D3S3559 | 198-220 | 54/30 | 4 |
| D3S3577 | 282-311 | 54/35 | 4 | | D3S3521 | 236-297 | 54/35 | 4 |
| D3S1295 | 98-129 | 54/30 | 8 | | D3S3727 | 124-150 | 50/40 | 8 |
| D3S1600 | 198-227 | 54/35 | 8 | | D3S3518 | 179-190 | 50/40 | 8 |
| D3S3660 | 228-262 | 54/30 | 8 | | D3S3678 | 240-291 | 56/35 | 8 |
| D3S1266 | 303-326 | 52/35 | 8 | | D3S1619 | 134-146 | 52/35 | 4 |
| D3S2420 | 105-136 | 54/30 | 4 | | D3S3685 | 180-227 | 54/35 | 4 |
| D3S1582 | 163-199 | 54/30 | 4 | | D3S1568 | 252-270 | 54/35 | 4 |
| D3S3616 | 204-234 | 54/35 | 4 | | | | | |
| D3S1613 | 245-278 | 54/35 | 4 | | | | | |

**Table 3.5. Final panels for the fluorescent genotyping of 3p STR markers**

The allele range is represented in basepairs, the PCR parameters in $T_A$/cycles and the pooling ratio in microliters. The colour of the background for each marker corresponds to the fluorescent dye label (blue = FAM, yellow = HEX, green = TET) and the colour used to display the respective dye in both ABI Genescan and Genotyper software when using filter set C.
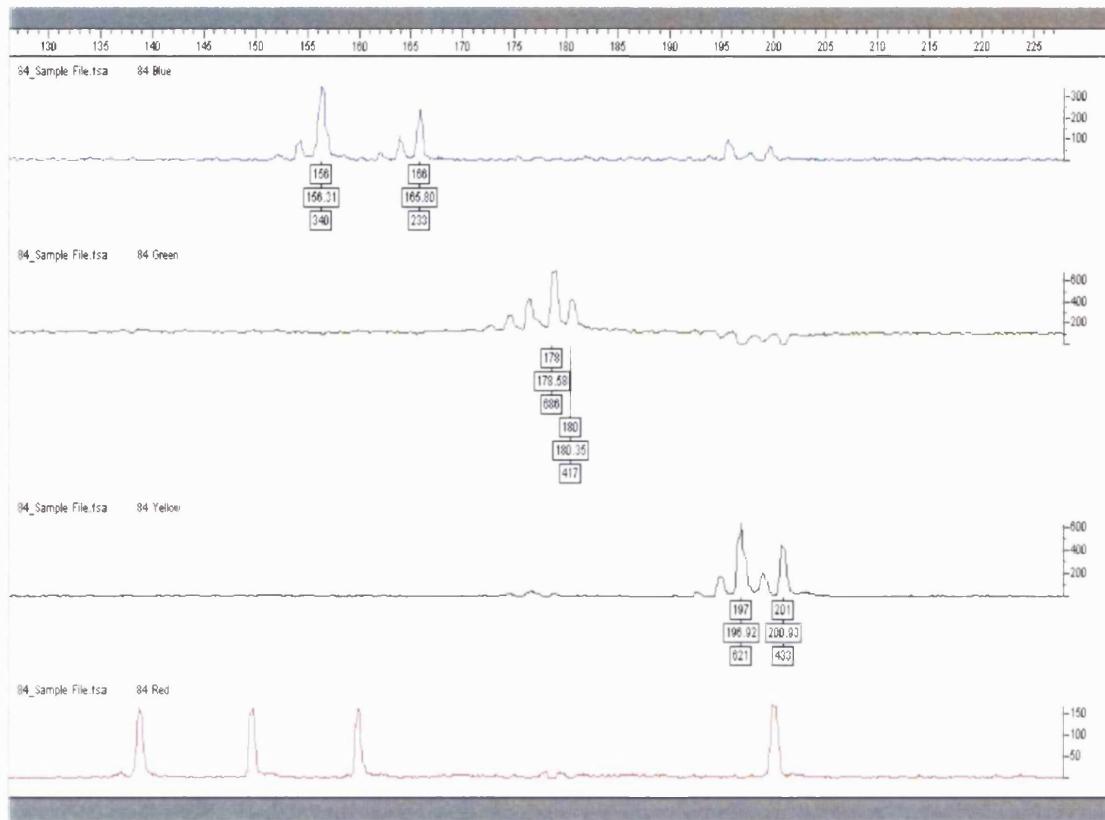
**Figure 3.5. Example of typical STR genotype traces within Genotyper**

For this filter set, the blue, green and yellow lanes contain FAM, TET and HEX dye

labelled markers, respectively. The red lane contains the TAMRA [500] size standard

with the peaks corresponding, left to right, to DNA fragments of sizes 139bp, 150bp,

160bp and 200bp. Each genotype shown is a heterozygote labelled below, from top to

bottom, with allele name, size relative to the size standard and peak intensity. A scale

of intensity is show to the right of each lane. A scale of size relative to the size standard

is shown across the top. Note, due to spectral overlap of the different fluorescent dyes,

peaks of the HEX labelled marker are 'bleeding' into the FAM lane and causing a

negative in the TET lane.

| Family | Problem | Solution |
|--------|---------|----------|
| 8 | DNA of an affected sibling failed to amplify | removed family |
| 14 | wrong DNA - unaffected sibling identical to father | removed unaffected sibling |
| 21 | wrong DNA - affected sibling identical to mother | removed family |
| 44 | affected sibling not related | removed family |
| 45 | DNA of mother failed to amplify | removed mother |
| 49 | DNA of father failed to amplify | removed father |
| 52 | wrong DNA - affected sibling identical to mother | removed family |
| 58 | father not related | removed father |
| 65 | father not related | removed father |
| 66 | wrong DNA - affected sibling identical to unaffected sibling | removed family |
| 80 | unaffacted siblings not related to family | removed unaffected siblings |
| 90 | unaffacted sibling not related to family | removed unaffected sibling |
| 92 | wrong DNA - affected sibling identical to father | removed family |
| 102 | mother not related | removed mother |
| 119 | affected sibling not related | removed family |
| 126 | family not related to each other | removed family |
| 129 | wrong DNA - father identical to mother | removed father |
| 134 | DNA of mother failed to amplify | removed mother |
| 150 | DNA of an affected sibling failed to amplify | removed family |
| 164 | affected sibling not related | removed family |
| 165 | wrong DNA - affected sibling identical to affected sibling | removed family |
| 167 | no affected sibling pair | removed family |
| 172 | unaffacted sibling not related to family | removed unaffected sibling |
| 199 | DNA of an affected sibling failed to amplify | removed family |
| 230 | affected sibling has different father (half-siblings) | made new father |
| 245 | affected sibling has different father (half-siblings) | made new father |

**Table 3.6. Problems with the Oxford linkage cohort pedigrees**

All families removed from analysis exhibited a problem either during amplification or inheritance checking, which resulted in the removal of the only ASP.

|                  | After first pass | After repeat 1 | After repeat 2 | After repeat 3 |
|------------------|------------------|----------------|----------------|----------------|
| PCR fails        | 2107 (6.0%)      | 466 (1.3%)     | 139 (0.4%)     | 38 (0.1%)      |
| Inheritance fails| 372 (1.1%)       | 228 (0.7%)     | 119 (0.3%)     | 66 (0.2%)      |
| Total to repeat  | 2479             | 694            | 258            | 104            |
| Overall success  | 92.89%           | 98.01%         | 99.26%         | 99.70%         |

Table 3.7.  Missing genotypes and success rates for the 3p extension study

siblings were either not related or there were multiple errors in the genotype data for the family. Thus, the family was removed from the study. An affected sibling from family 124, whose DNA had not amplified well, also resulted in multiple recombination events, suggesting multiple false genotypes. This subject was also removed from the study. Again, the description of the Oxford linkage cohort in Section 2.2.1. includes these changes.

## 3.1.6. Genetic map construction and comparison

The 3p STR genotype data from the Oxford linkage cohort was analysed within MultiMap to construct a genetic map (Section 2.1.6.3.). Due to the lack of informative recombinants, caused by the proximity of the markers and the number of families analysed, only 14 of the 41 markers could be placed on the map (Table 3.8.). The map was comparable to both the Marshfield sex-average genetic map and the G3 RHMAP RH map in both overall length and marker order (Figure 3.6.), although the actual map positions were more consistent with the Marshfield map. All markers not mapped to a position were placed into intervals between mapped markers that corresponded to their position on the G3 RHMAP RH map.

## 3.1.7. Linkage analysis

The 3p STR genotype data from the Oxford linkage cohort was analysed for linkage to IBD as described in Section 2.1.6.4. Linkage was not detected across the chromosome 3p region to any IBD phenotype model (Table 3.9.). The highest LOD score was detected in the CD population with a LOD of 1.25 between D3S3640 and D3S1581. Graphs showing the multipoint LOD curves and sharing statistics for all three phenotype models are shown in Figures 3.7. and 3.8., respectively. The average information content across the region was 0.9 (Figure 3.9.). To examine the differences

116

| 3p MultiMap Genetic Map | |
|---|---|
| Marker | Map position |
| D3S3727 | 0 |
| D3S1619 | 3.9 |
| D3S1298 | 7 |
| D3S3527 | 8.2 |
| D3S3685 | 12.1 |
| D3S3647 | 12.5 |
| D3S1568 | 14.5 |
| D3S1578 | 15.9 |
| D3S1582 | 18.3 |
| D3S3719 | 19.4 |
| D3S1766 | 22 |
| D3S3577 | 23.5 |
| D3S3571 | 32.9 |
| D3S1285 | 35.1 |

**Table 3.8. MultiMap generated 3p genetic map**

The genetic map was generated from the Oxford linkage cohort STR genotype data.

The map position of each marker is shown in centiMorgans relative to the initial

marker.
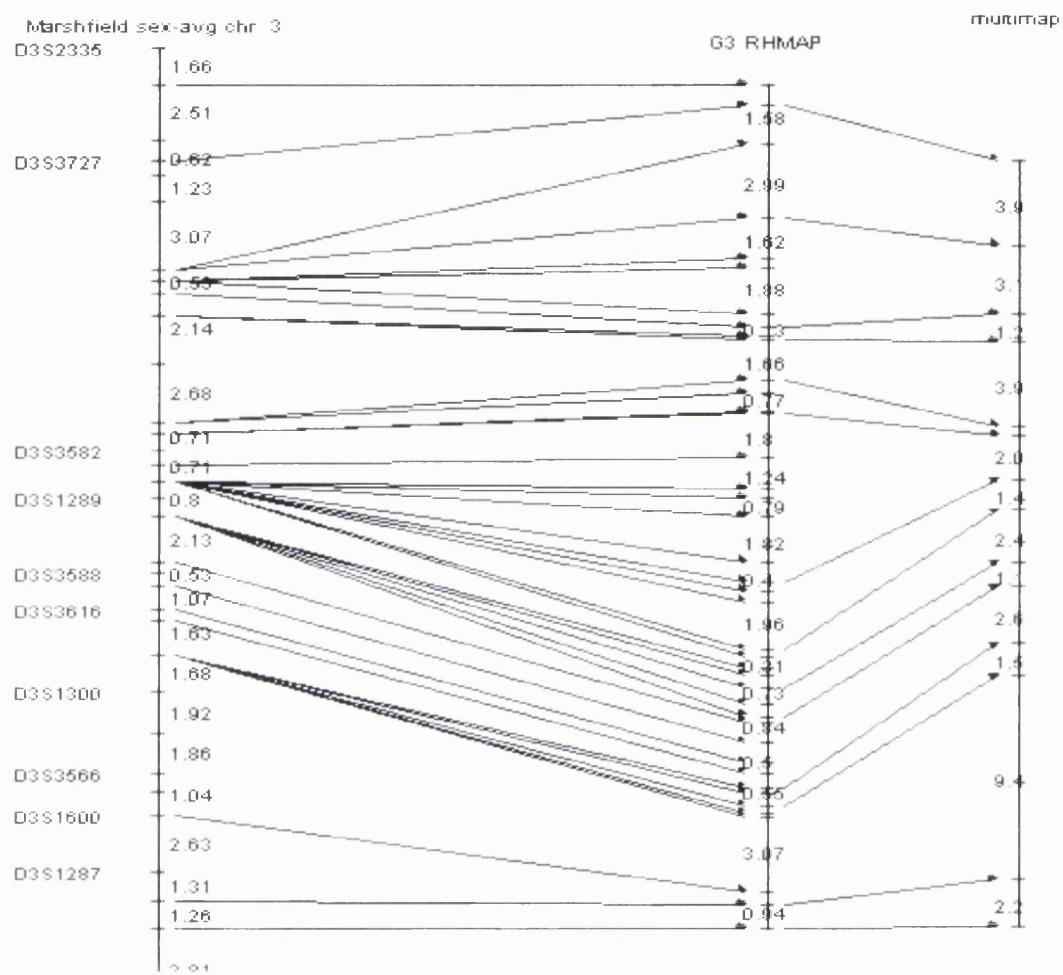
**Figure 3.6. MultiMap generated 3p genetic map comparison**

The maps within the map comparison correspond to the Marshfield sex-average genetic

map shown in Table 3.1., the G3 RHMAP RH map shown in Table 3.4., and the 3p

genetic map shown in the Table 3.8.

| Phenotype Model | Peak Marker Interval | Multipoint LOD | Mean Allele Sharing | Sibling Risk |
|---|---|---|---|---|
| IBD | D3S1578 - D3S3672 | 0.89 | 0.54 | 1.08 |
| CD | D3S3640 - D3S1581 | 1.25 | 0.58 | 1.41 |
| UC | D3S1578 - D3S3672 | 0.20 | 0.53 | 1.07 |

**Table 3.9. The results of the 3p linkage study (G3 RHMAP)**

The results shown are from linkage analysis using the G3 RHMAP RH map shown in

Table 3.4. The mean allele sharing was calculated using the formula $Z_1*0.5+Z_2$. The

sibling risk was calculated using the formula $0.25/Z_0$.
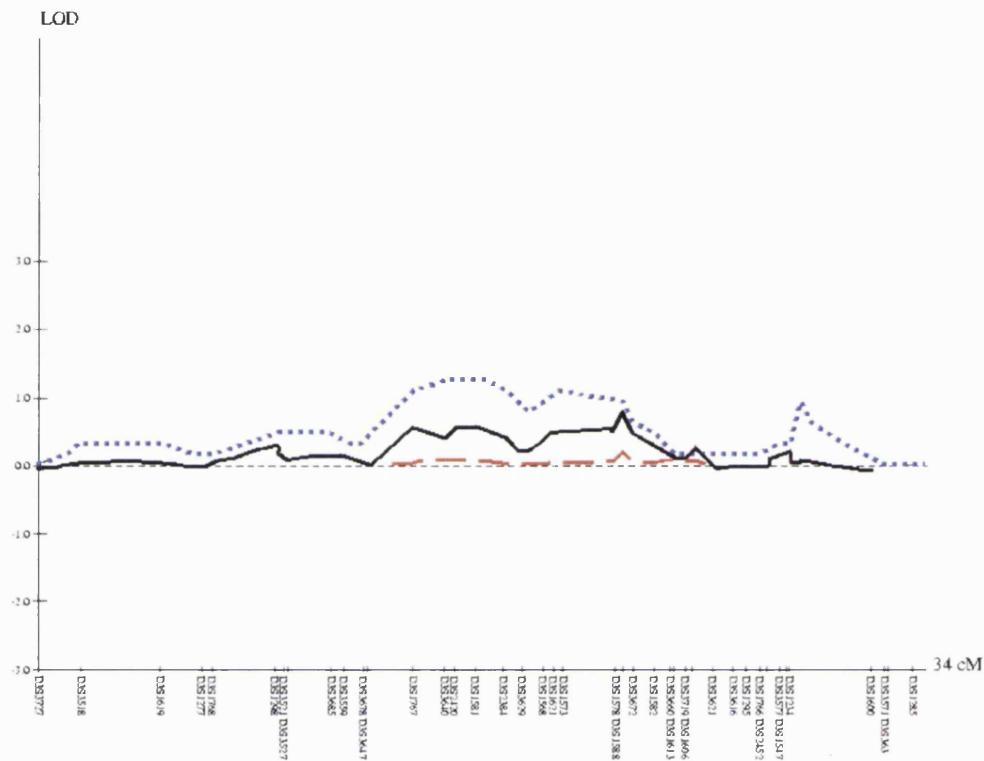
**Figure 3.7. Chromosome 3p non-parametric multipoint linkage curves (G3 RHMAP)**

The G3 RH map shown in Table 3.4. was used for maker order and spacing. Results for the IBD (black, solid line), CD (blue, dotted line) and UC (red, dashed line) phenotype models are shown.
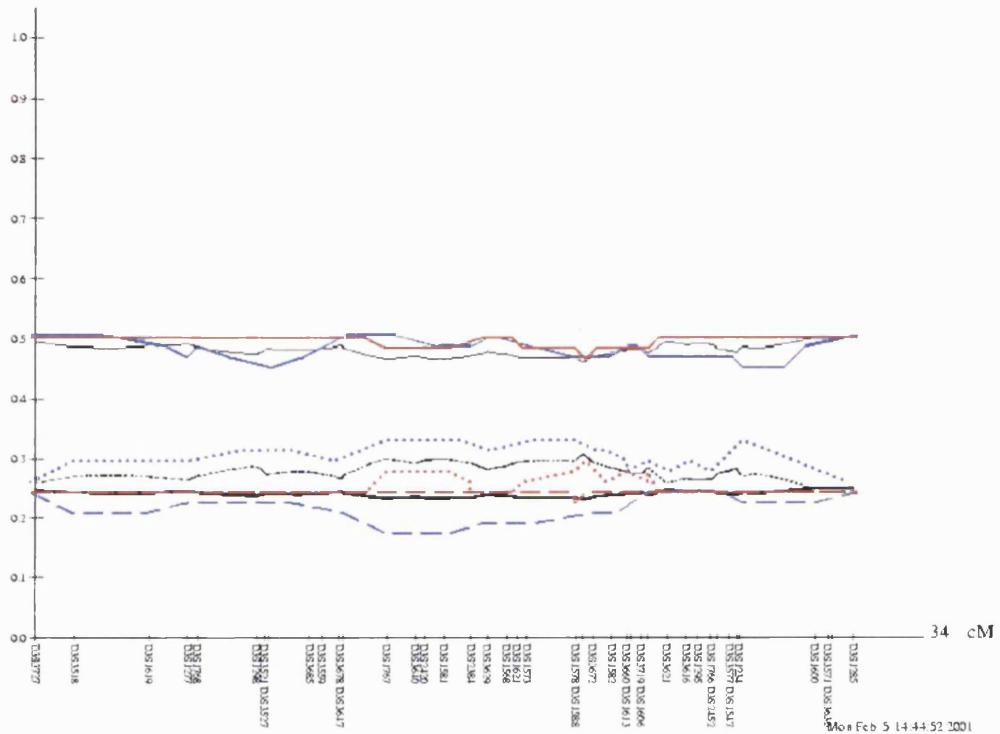
**Figure 3.8. Chromosome 3p sharing proportion curves (G3 RHMAP)**

The G3 RHMAP RH map shown in Table 3.4. was used for maker order and spacing.

Results for the IBD (black lines), CD (blue lines) and UC (red lines) phenotype models

are shown. The sharing statistics for $Z_0$ (dashed lines), $Z_1$ (solid lines) and $Z_2$ (dotted

lines) are shown.
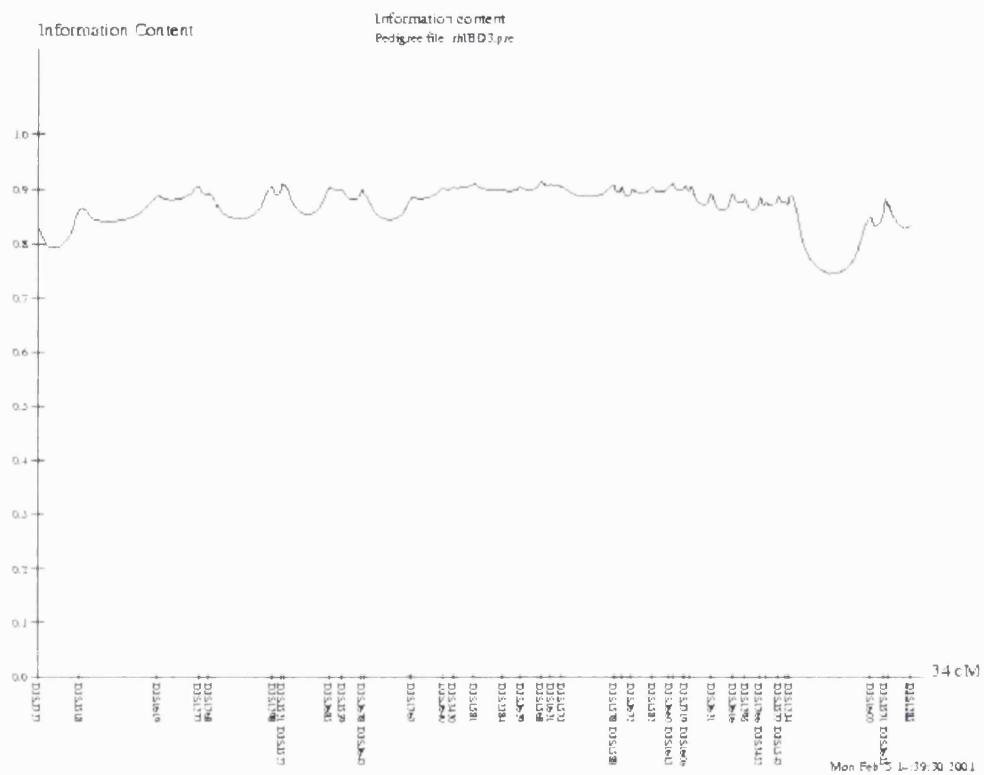
**Figure 3.9. Chromosome 3p average information content (G3 RHMAP)**

The G3 RHMAP RH map shown in Table 3.4. was used for maker order and spacing.

seen between the results from this extension study and those from the previous genome scan (Satsangi *et al.*, 1996b), SIBS1&2 were analysed separately for linkage (Figures 3.10.-3.12.). The result of this analysis is summarised in Table 3.10. To analyse the effect that maker density, marker spacing and allele frequency has on linkage results, the data was also analysed using the Marshfield chromosome 3 sex-average genetic map, the MultiMap genetic map generated from the genotype data, and uniform allele frequencies (Figures 3.13.-3.17.). The results of these analyses are summarised in Table 3.11.

### 3.1.8. TDT association analysis

The 3p STR genotype data from the Oxford linkage cohort was analysed for association to IBD as described in Section 2.1.6.5. Several chromosome 3p STR markers showed positive association with IBD phenotypes prior to correcting for multiple testing (Figure 3.18.). The markers and specific alleles showing the strongest association (p < 0.01) to IBD phenotypes are shown in Table 3.12. However, simulations with 1000 equivalent data sets resulted in 290, 356, and 911 data sets showing greater association to IBD, CD and UC, respectively.

## 3.2.  Chromosome 3p association study with STR markers

### 3.2.1.  Genotyping and quality checking

The 42 chromosome 3p STR markers (Section 3.4.4.) were genotyped (Sections 2.1.5.4.-9.) across the Oxford TDT1 cohort (Section 2.2.2.) generating approximately 16,000 genotypes (Figure 3.5.). The genotype and pedigree data was analysed with Pedcheck as described in Section 2.1.6.1. Unrelated subjects identified by Pedcheck analysis and DNA that failed to amplify resulted in the removal 26 families (Table
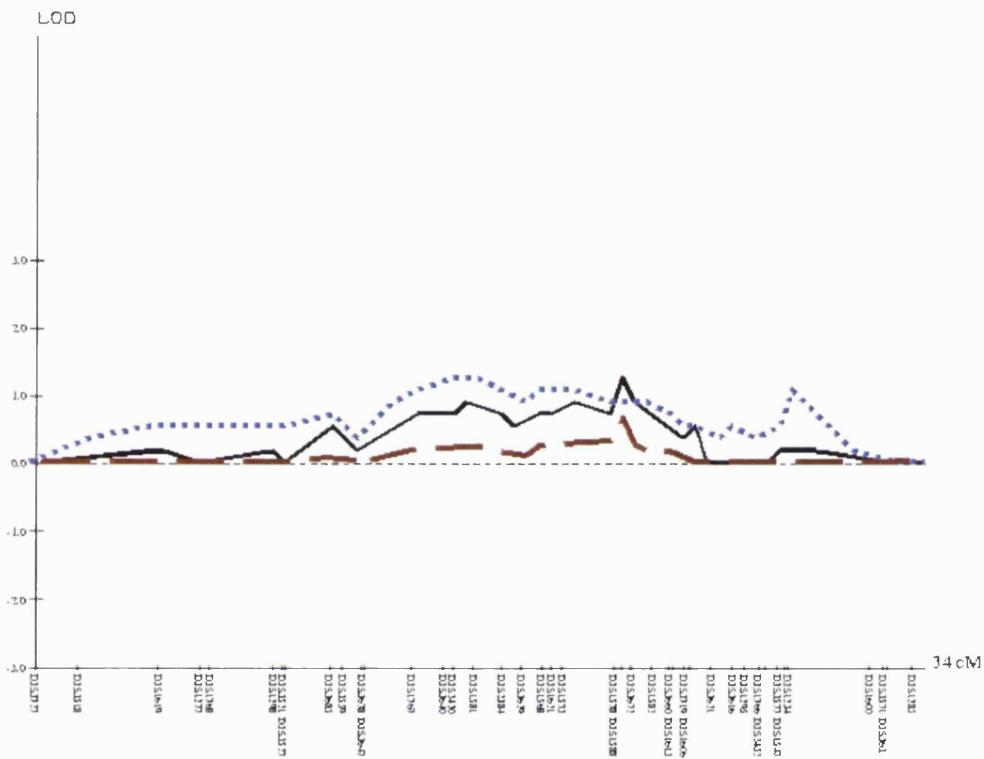
**Figure 3.10. Chromosome 3p non-parametric multipoint linkage curves (SIBS1&2)**

The G3 RHMAP RH map shown in Table 3.4. was used for maker order and spacing.

Results for the IBD (black, solid line), CD (blue, dotted line) and UC (red, dashed line)

phenotype models are shown.
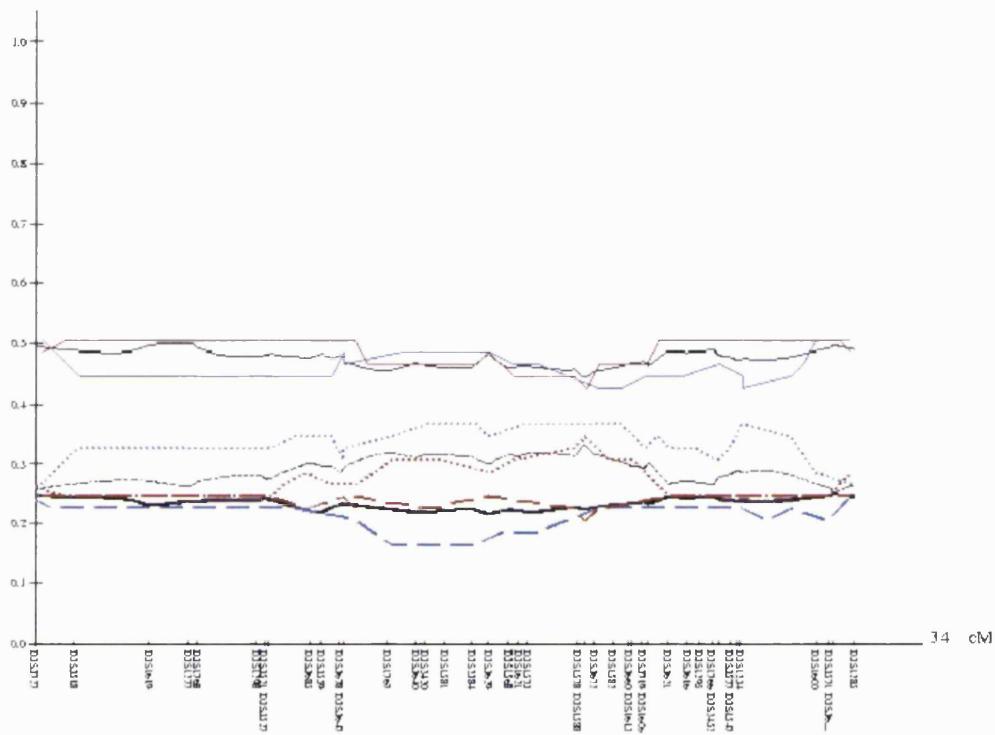
**Figure 3.11. Chromosome 3p sharing proportion curves (SIBS1&2)**

The G3 RH RHMAP map shown in Table 3.4. was used for maker order and spacing.

Results for the IBD (black lines), CD (blue lines) and UC (red lines) phenotype models

are shown. The sharing statistics for $Z_0$ (dashed lines), $Z_1$ (solid lines) and $Z_2$ (dotted

lines) are shown.
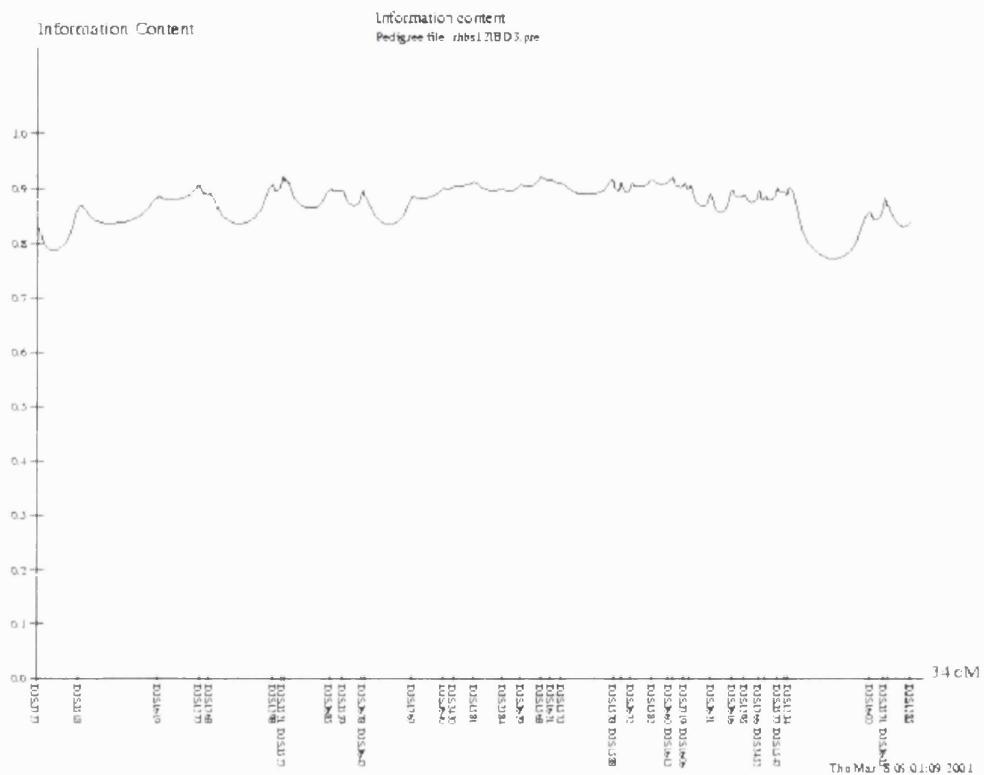
**Figure 3.12. Chromosome 3p average information content (SIBS1&2)**

The G3 RHMAP RH map shown in Table 3.4. was used for maker order and spacing.

| Phenotype Model | Peak Marker Interval | Multipoint LOD | Mean Allele Sharing | Sibling Risk |
|---|---|---|---|---|
| IBD | D3S1588 - D3S3672 | 1.19 | 0.55 | 1.12 |
| CD | D3S1581 - D3S2384 | 1.21 | 0.59 | 1.51 |
| UC | D3S1588 - D3S3672 | 0.74 | 0.57 | 1.16 |

**Table 3.10. The results of the 3p linkage study (SIBS1&2)**

The results shown are from linkage analysis using the G3 RHMAP RH map shown in Table 3.4. The mean allele sharing was calculated using the formula $Z_1*0.5+Z_2$. The sibling risk was calculated using the formula $0.25/Z_0$.
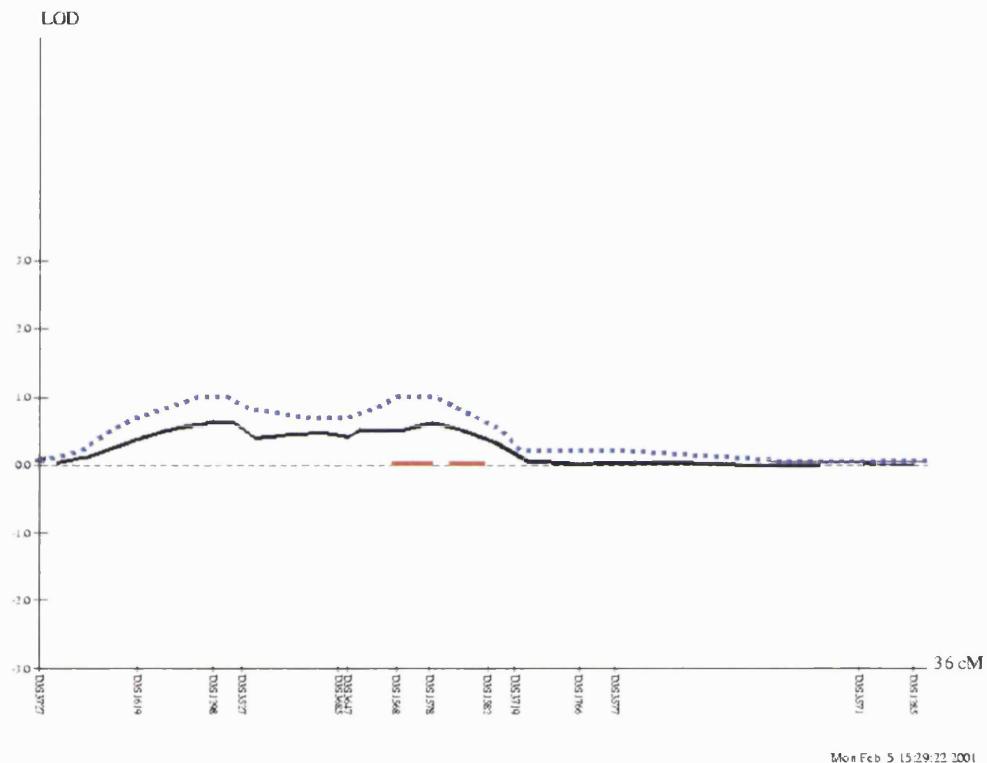
**Figure 3.13. Chromosome 3p non-parametric multipoint linkage curves (MultiMap)**

The MultiMap genetic map shown in Table 3.8. was used for maker order and spacing.

Results for the IBD (black, solid line), CD (blue, dotted line) and UC (red, dashed line)

phenotype models are shown.
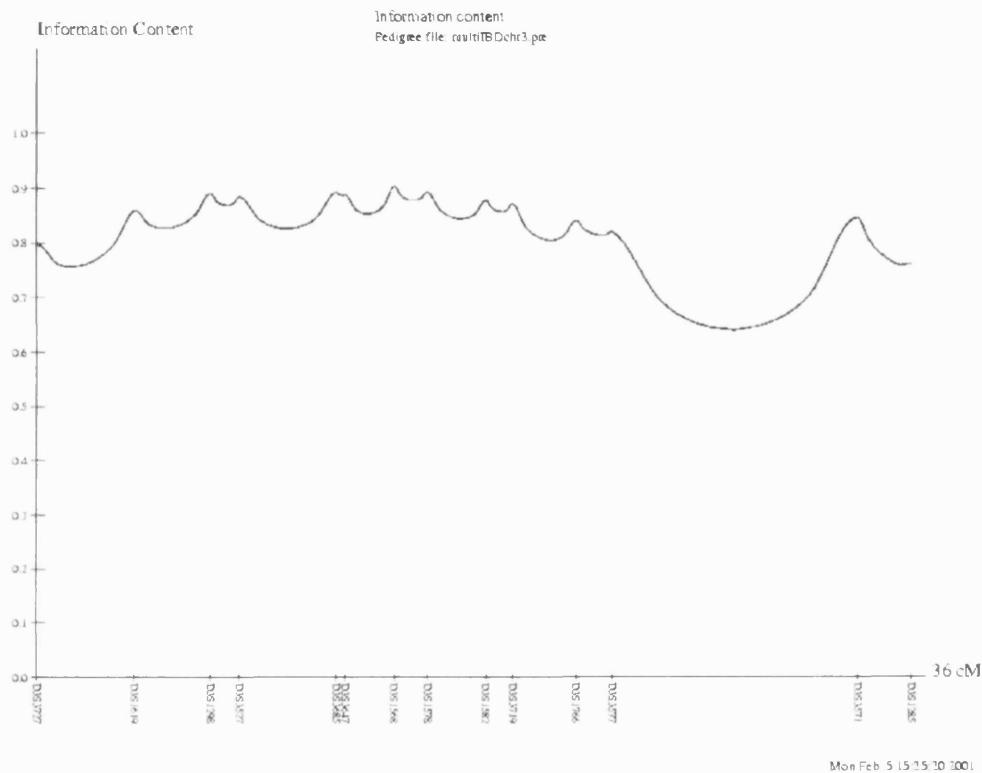
**Figure 3.14. Chromosome 3p average information content (MultiMap)**

The MultiMap genetic map shown in Table 3.8. was used for maker order and spacing.
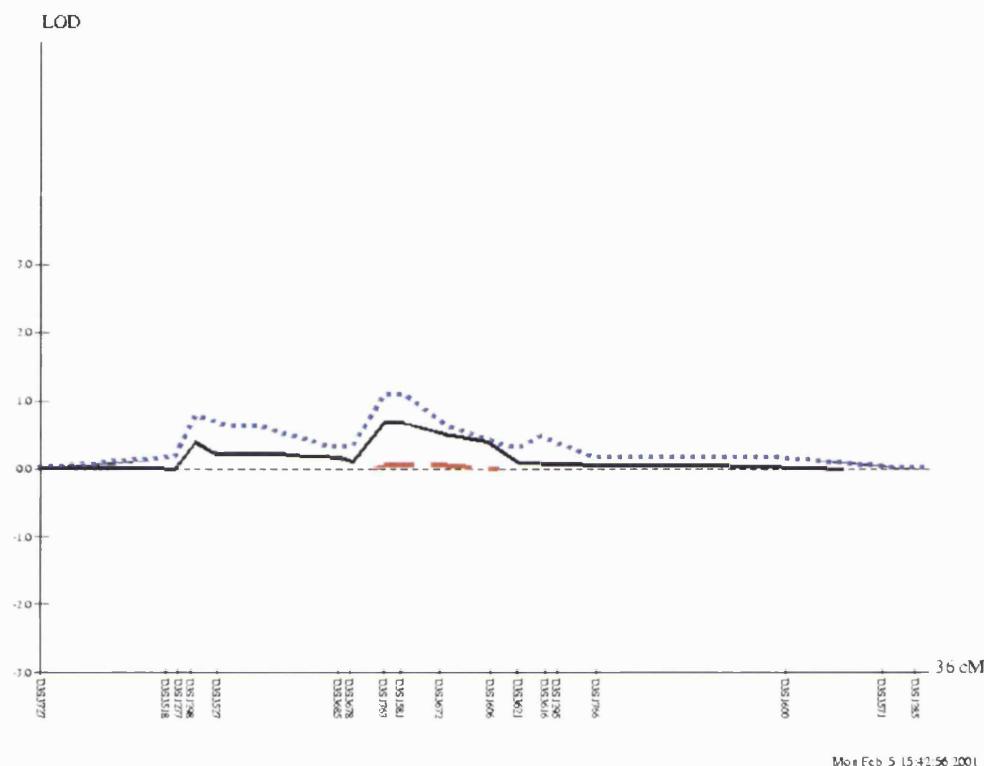
**Figure 3.15. Chromosome 3p non-parametric multipoint linkage curves**

**(Marshfield)**

The Marshfield genetic map shown in Table 3.1. was used for maker order and spacing.

Results for the IBD (black, solid line), CD (blue, dotted line) and UC (red, dashed line)

phenotype models are shown.
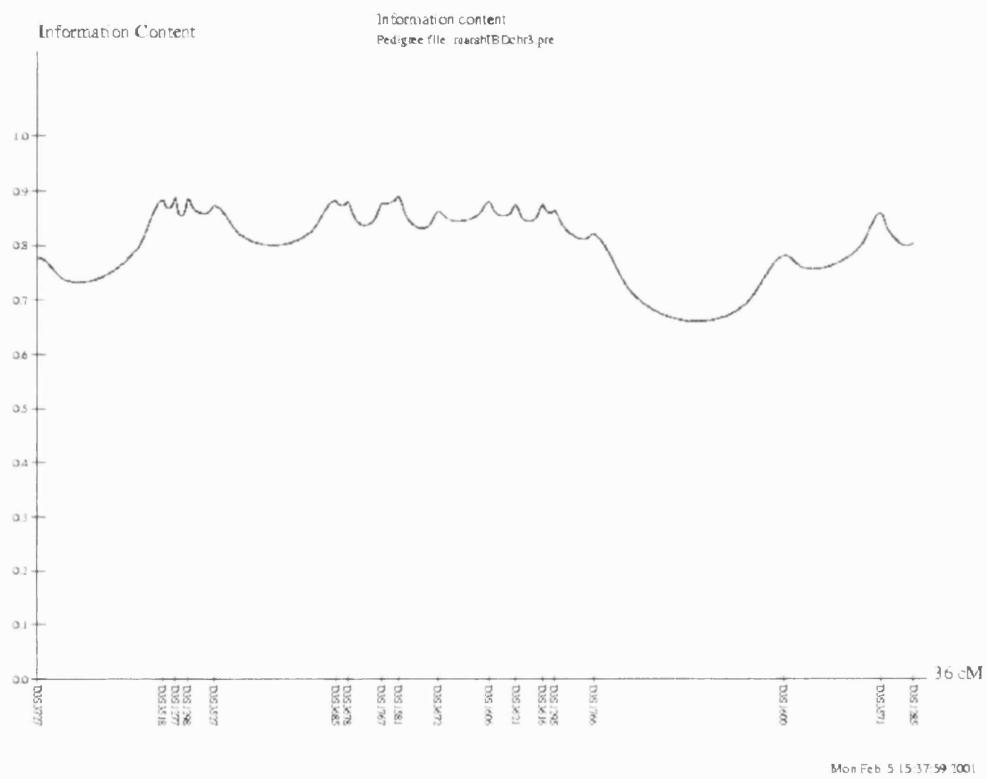
**Figure 3.16. Chromosome 3p average information content (Marshfield)**

The Marshfield genetic map shown in Table 3.1. was used for maker order and spacing.
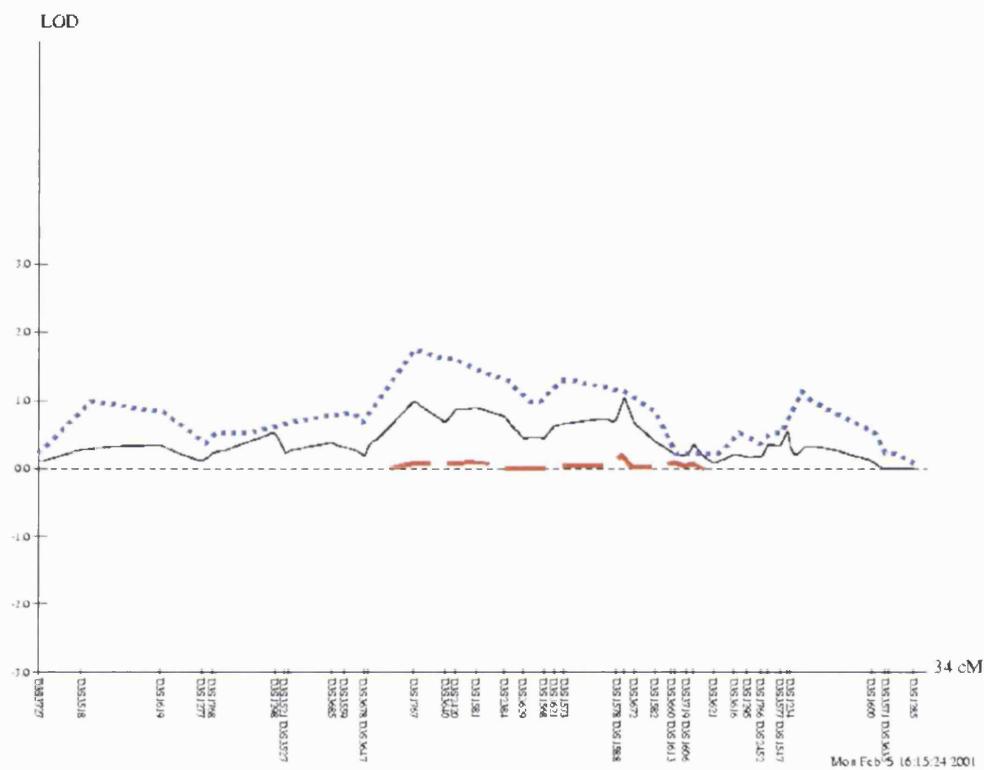
**Figure 3.17. Chromosome 3p non-parametric multipoint linkage curves (Uniform)**

The G3 RH map shown in Table 3.4. was used for maker order and spacing. Uniform allele frequencies were used for linkage analysis rather then those generated from the genotype data. Results for the IBD (black, solid line), CD (blue, dotted line) and UC (red, dashed line) phenotype models are shown.

|  | IBD | CD | UC |
|---|---|---|---|
| MultiMap | 0.61 (D3S1298 - D3S3527) | 1.01 (D3S1578 - D3S1582) | 0.02 (D3S1578 - D3S1582) |
| Marshfield | 0.72 (D3S1767 - D3S3672) | 1.14 (D3S1767 - D3S3672) | 0.08 (D3S1767 - D3S3672) |
| Uniform | 1.05 (D3S1578 - D3S3672) | 1.73 (D3S1767 - D3S3640) | 0.21 (D3S1578 - D3S3672) |

**Table 3.11.  Summary of the results from linkage analysis methods comparison**

Results are given for the IBD, CD and UC phenotype models.  Linkage analysis was conducted using the MultiMap shown in Table 3.8., the Marshfield map shown in Table 3.1., and the RHMAP shown in Table 3.4. using uniform allele frequencies rather than the allele frequencies generated from genotyping.  The marker interval were peak multipoint linkage was seen in shown in brackets next to the LOD score in blue.
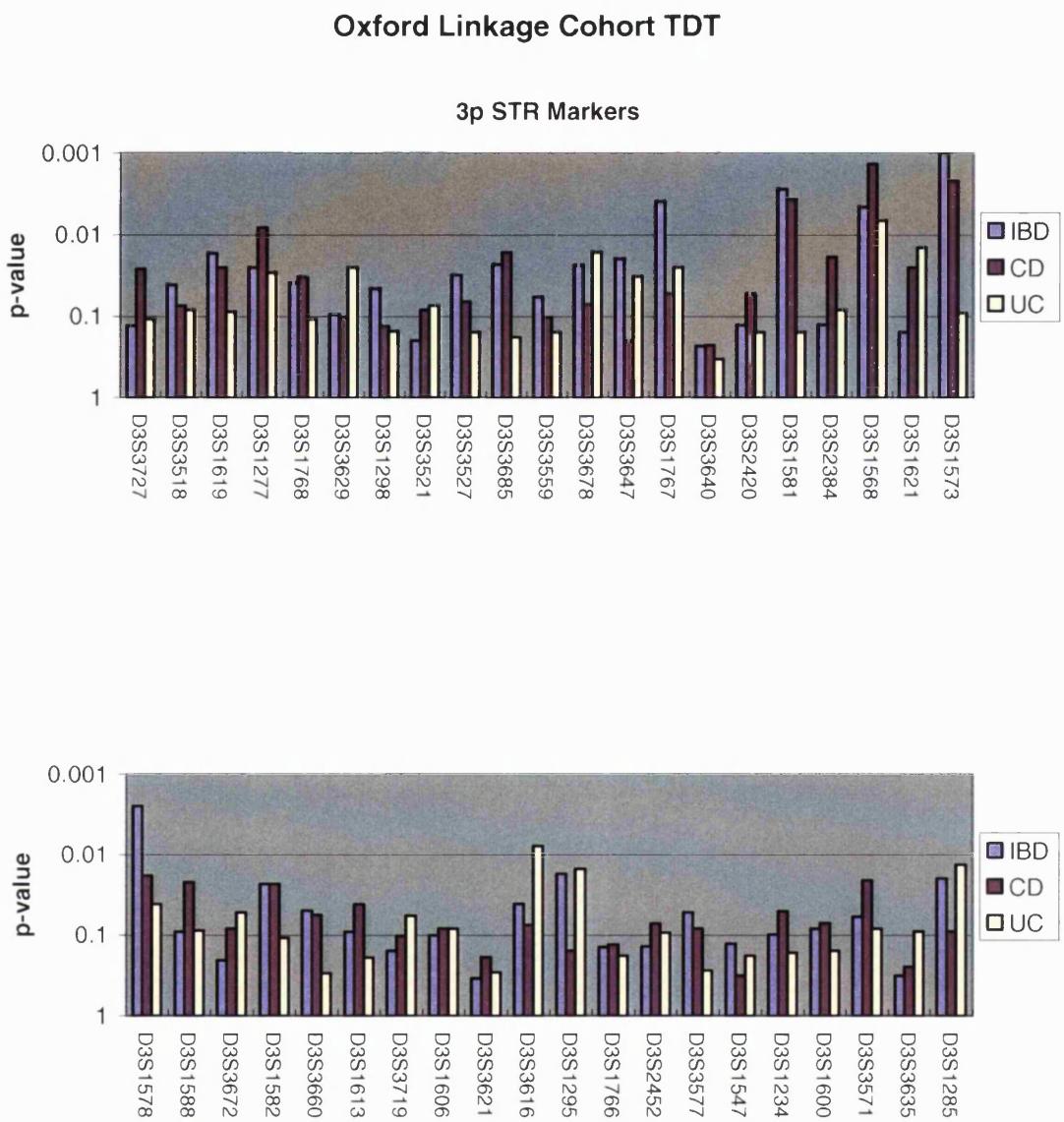
**Figure 3.18. TDT results of 3p STR markers (Oxford linkage cohort)**

The G3 RH map shown in Table 3.4. was used for maker order. The p-values shown have not been corrected for multiple testing.

| Marker | Phenotype | Allele | Transmissions | Non-transmissions | chi-square | p-value |
|---|---|---|---|---|---|---|
| D3S1573 | ibd | 1 | 45 | 82 | 10.78 | 0.001026 |
| D3S1573 | cd | 1 | 23 | 49 | 9.39 | 0.002183 |
| D3S1568 | cd | 17 | 1 | 13 | 10.29 | 0.001341 |
| D3S1568 | ibd | 17 | 4 | 17 | 8.05 | 0.004556 |
| D3S1568 | ibd | 13 | 39 | 67 | 7.4 | 0.006536 |
| D3S1568 | uc | 13 | 13 | 31 | 7.36 | 0.006656 |
| D3S1578 | ibd | 11 | 76 | 43 | 9.15 | 0.002485 |
| D3S1578 | ibd | 13 | 26 | 50 | 7.58 | 0.005905 |
| D3S1581 | ibd | 7 | 73 | 41 | 8.98 | 0.002726 |
| D3S1581 | cd | 7 | 53 | 27 | 8.45 | 0.00365 |
| D3S1767 | ibd | 2 | 25 | 50 | 8.33 | 0.003892 |
| D3S3616 | uc | 3 | 16 | 35 | 7.08 | 0.007802 |
| D3S1277 | cd | 5 | 21 | 42 | 7 | 0.008151 |
| D3S1277 | cd | 9 | 77 | 48 | 6.73 | 0.009491 |

**Table 3.12. 3p STR markers showing strongest association (Oxford linkage cohort)**

3.13.). The description of the Oxford TDT cohort in Section 2.2.2. includes these changes. Three rounds of repeats were conducted which resulted in an overall genotyping success rate of 99.7% (Table 3.14.).

### 3.2.2. TDT association analysis

The 3p STR genotype data from the Oxford TDT1 cohort was analysed for association to IBD as described in Section 2.1.6.5. Several chromosome 3p STR markers showed positive association with IBD phenotypes prior to correcting for multiple testing (Figure 3.19.). The markers and specific alleles showing the strongest association (p < 0.01) to IBD phenotypes are shown in Table 3.15. However, simulations with 1000 equivalent data sets resulted in 698, 286, and 462 data sets showing greater association to IBD, CD and UC, respectively.

## 3.3. Chromosome 3p candidate gene association studies

### 3.3.1. Gene selection

The Genebridge 4 Genemap99 RH map at the NCBI Internet site was analysed to identify genes that mapped into chromosome 3p. All known genes in a 15cM region centered at D3S1573, the marker exhibiting suggestive linkage to IBD in the Satsangi et al. (1996) genome scan (Section 1.4.3.2.), were investigated as potential IBD positional candidate genes (Table 3.16.). After an initial literature analysis, CCR5, CCR2, CCRL2, and GNAI2 were selected for further association studies as good IBD candidate genes.

| Family | Problem | Family | Problem |
|--------|---------|--------|---------|
| 5 | DNA failed to amplify | 214 | contaminated DNA |
| 36 | father not related | 217 | DNA failed to amplify |
| 41 | DNA failed to amplify | 221 | proband not related |
| 45 | DNA failed to amplify | 233 | contaminated DNA |
| 95 | DNA failed to amplify | 245 | mother not related |
| 102 | proband not related | 249 | DNA failed to amplify |
| 133 | proband not related | 255 | DNA failed to amplify |
| 148 | DNA failed to amplify | 257 | DNA failed to amplify |
| 152 | proband not related | 262 | DNA failed to amplify |
| 178 | DNA failed to amplify | 267 | DNA failed to amplify |
| 190 | proband not related | 284 | father not related |
| 191 | proband not related | 288 | contaminated DNA |
| 205 | proband not related | 322 | contaminated DNA |

**Table 3.13. Families removed from the Oxford TDT cohort**

|                  | After first pass | After repeat 1 | After repeat 2 | After repeat 3 |
|------------------|------------------|----------------|----------------|----------------|
| PCR fails        | 503 (3.2%)       | 177 (1.1%)     | 47 (0.3%)      | 10 (0.1%)      |
| Inheritance fails| 111 (0.7%)       | 69 (0.4%)      | 39 (0.2%)      | 36 (0.2%)      |
| Total to repeat  | 614              | 246            | 86             | 46             |
| Overall success  | 96.07%           | 98.43%         | 99.45%         | 99.71%         |

**Table 3.14. Missing genotypes and success rates for the 3p STR association study**
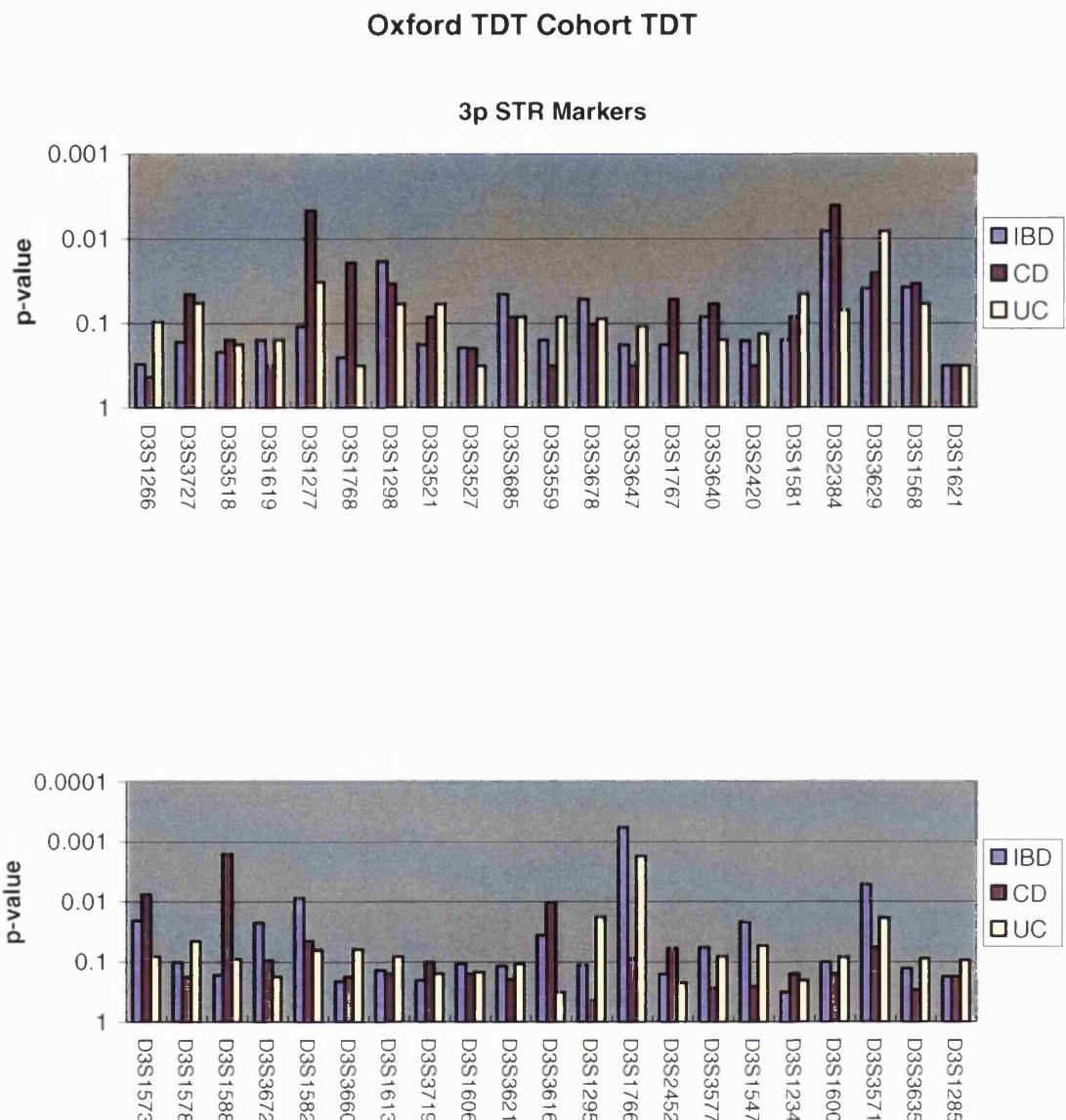
**Figure 3.19. TDT results of 3p STR markers (Oxford TDT cohort)**

The G3 RH map shown in Table 3.4. was used for maker order. The p-values shown

have not been corrected for multiple testing.

| Marker | Phenotype | Allele | Transmissions | Non-transmissions | chi-square | p-value |
|--------|-----------|--------|---------------|-------------------|------------|---------|
| D3S1766 | ibd | 5 | 42 | 80 | 11.84 | 0.000581 |
| D3S1766 | uc | 5 | 26 | 54 | 9.8 | 0.001745 |
| D3S1766 | ibd | 6 | 57 | 32 | 7.02 | 0.008049 |
| D3S1766 | uc | 6 | 42 | 20 | 7.81 | 0.005206 |
| D3S1588 | cd | 5 | 2 | 15 | 9.94 | 0.001616 |
| D3S2384 | cd | 4 | 26 | 9 | 8.26 | 0.004059 |
| D3S2384 | cd | 2 | 5 | 18 | 7.35 | 0.006714 |
| D3S2384 | ibd | 2 | 21 | 42 | 7 | 0.008151 |
| D3S1277 | cd | 5 | 3 | 15 | 8 | 0.004678 |
| D3S3571 | ibd | 6 | 10 | 27 | 7.81 | 0.005193 |
| D3S1573 | cd | 1 | 3 | 14 | 7.12 | 0.007633 |
| D3S3629 | uc | 10 | 7 | 0 | 7 | 0.008151 |
| D3S1582 | ibd | 7 | 50 | 27 | 6.87 | 0.008765 |

**Table 3.15. 3p STR markers showing strongest association (Oxford TDT cohort)**

| Gene abreviation | Gene name/description | Gene abreviation | Gene name/description |
|---|---|---|---|
| D3S3537 | | MAPKAPK3 | mitogen-activated protein kinase-activated protein kinase 3 |
| D3S3521 | | NPR2L | Homologous to yeast nitrogen permease |
| OSR1 | oxidative-stress responsive 1 | GNAI2 | guanine nucleotide binding protein, alpha inhibiting activity, polypeptide 2 |
| ACAA | acetyl-Coenzyme A acyltransferase | SEMA3F | semaphorin III/F |
| GPR13 | G protein-coupled receptor 13 | ARP | Arginine-rich protein |
| PLCD1 | Phospholipase C, delta 1 | GNAT1 | Guanine nucleotide binding protein, alpha transducing activity polypeptide 1 |
| LAMR1 | laminin receptor 1 (67kD); Ribosomal protein SA | HYAL3 | hyaluronoglucosaminidase 3 |
| MYD88 | myeloid differentiation primary response gene (88) | MST1R | macrophage stimulating 1 receptor |
| HYA22 | HYA22 protein | CACNA2D2 | calcium channel, voltage-dependent, alpha 2/delta subunit 2 |
| RPSA | Ribosomal protein SA | DOCK3 | dedicator of cyto-kinesis 3 |
| GC20 | translation factor sui1 homolog | ITIH3 | pre-alpha (globulin) inhibitor, H3 polypeptide |
| CCK | cholecystokinin | KIAA0800 | KIAA0800 gene product |
| NKTR | natural killer-tumor recognition sequence | UBE1L | Ubiquitin-activating enzyme E1, like |
| D3S3685 | | TKT | transketolase |
| D3S3559 | | PRKCD | Protein kinase C, delta |
| D3S3564 | | ALAS1 | aminolevulinate, delta-, synthase 1 |
| HIG1 | hypoxia-inducbile gene 1 | PPP2R5C | Protein phosphatase 2, regulatory subunit B (B56) |
| VIPR1 | vasoactive intestinal peptide receptor 1 | HYAL1 | Hyaluronoglucosaminidase 1 |
| CYP8B1 | Cytochrome P450, subfamily VIIIB, polypeptide 1 | GPX3 | glutathione peroxidase 3 |
| KIAA0096 | KIAA0096 protein | BAP1 | BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase) |
| KIAA0028 | Leucyl-tRNA synthetase, mitochondrial | TNNC1 | troponin C, slow |
| HBA1 | hemoglobin, alpha 1 | IFRD2 | Interferon-related developmental regulator 2 |
| MCM4 | Minichromosome maintenance deficient 4 | GRM2 | glutamate receptor, metabotropic 2 |
| CCR9 | chemokine (C-C motif) receptor 9 | SCYA5 | small inducible cytokine A5 |
| CCR3 | chemokine (C-C motif) receptor 3 | ITIH4 | inter-alpha (globulin) inhibitor H4 |
| ZNF35 | zinc finger protein 35 | STK2 | serine/threonine kinase 2 |
| LOC51099 | CGI-58 protein | ITIH1 | inter-alpha (globulin) inhibitor, H1 polypeptide |
| KIAA0851 | KIAA0851 protein | LTF | lactotransferrin |
| ZNF20 | zinc finger protein, C2H2, rapidly turned over | D3S1578 | |
| TNA | tetranectin (plasminogen-binding protein) | D3S1588 | |
| ZNP | zinc finger protein | D3S1581 | |
| D3S3678 | | D3S3640 | |
| D3S3647 | | D3S1568 | |
| D3S1767 | | D3S1621 | |
| KIAA0116 | KIAA0116 protein | D3S2420 | |
| CCR1 | chemokine (C-C motif) receptor 1 | D3S3629 | |
| CCR5 | chemokine (C-C motif) receptor 5 | D3S1573 | |
| CCR2 | chemokine (C-C motif) receptor 2 | D3S2456 | |
| MYL3 | myosin, light polypeptide 3, alkali; ventricular, skeletal, slow | D3S2384 | |
| CAMP | cathelicidin antimicrobial peptide | CACNA1D | Calcium channel, voltage-dependent, L type, alpha 1D subunit |
| CDC25A | cell division cycle 25A | U3-55K | U3 snoRNP-associated 55-kDa protein |
| MAP4 | microtubule-associated protein 4 | HT017 | HT017 protein |
| PTHR1 | parathyroid hormone receptor 1 | D3S1582 | |
| PFKFB4 | 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4 | D3S3666 | |
| ARI2 or TRAID1 | Ariadne-2 homolog (all-trans retinoic acid inducible RING finger ) | D3S3672 | |
| SMARCC1 | SWI/SNF related, matrix associated, actin dependent regulator of chromatin, C1 | D3S3660 | |
| QARS | glutamine-tRNA synthetase | D3S1613 | |
| PLXN5 | plexin 5 | D3S3719 | |
| PRKAR2A | Protein kinase, cAMP-dependent, regulatory, type II, alpha | KIAA0378 | KIAA0378 protein |
| USP4 | Ubiquitin specific protease 4 | WNT5A | Wingless-type MMTV integration site family, member 5A |
| IMPDH2 | IMP (inosine monophosphate) dehydrogenase 2 | KIAA1105 | KIAA1105 protein |
| COL7A1 | collagen, type VII, alpha 1 | D3S3621 | |
| GPX1 | glutathione peroxidase 1 | LOC57406 | Lipase protein |
| UQCRC1 | ubiquinol-cytochrome c reductase core protein I | FLNB | Filamin B, beta (actin-binding protein-278) |
| DAG1 | dystroglycan 1 (dystrophin-associated glycoprotein 1) | HESX1 | Homeo box (expressed in ES cells) 1 |
| AMT | aminomethyltransferase (glycine cleavage system protein T) | ARF4 | ADP-ribosylation factor 4 |
| ARHA | ras homolog gene family, member A | EEF1D | eukaryotic translation elongation factor 1 delta |
| CSPG5 | Chondroitin sulfate proteoglycan 5 (neuroglycan C) | PDHB | pyruvate dehydrogenase (lipoamide) beta |
| KIAA0263 | KIAA0263 gene product | ACOX2 | acyl-Coenzyme A oxidase 2, branched chain |
| RBM6 | RNA binding motif protein 6 | CADPS | Ca2+-dependent activator protein for secretion |
| APEH | N-acylaminoacyl-peptide hydrolase | SCA7 | Spinocerebellar ataxia 7 |
| RBM5 | RNA binding motif protein 5 | D3S1547 | |
| TRIP | TRAF interacting protein | D3S1766 | |
| HYAL2 | Hyaluronoglucosaminidase 2 | D3S1234 | |
| IFNAR2 | interferon (alpha, beta and omega) receptor 2 | D3S3577 | |
| ACY1 | aminoacylase 1 | D3S2452 | |
| DUSP7 | dual specificity phosphatase 7 | | |

**Table 3.16.  Known genes mapping into a 15cM region around D3S1573**

The genes were retrieved from the Genebridge 4 Genemap99 RH map.  The genes selected for subsequent candidate gene association studies are shown in green.  The 3p STR markers shown in blue were used as framework markers to map the genes between.  The peak linkage marker (D3S1573) from the Satsangi et al (1996) genome scan is indicated in red.

### 3.3.2. Fragment design and amplification

For each candidate gene, a cDNA sequence and an encompassing gDNA sequence was retrieved from the GenBank sequence database. Alternative transcripts were identified and retrieved for both CCR2 and CCRL2. Promoter regions were identified and retrieved for both CCR5 and CCR2. A list of the accession numbers for each sequence retrieved is shown in Table 3.17. Oligonucleotides were designed to amplify 52 polymorphism detection fragments across the non-coding regions of the candidate gene as described in Section 2.1.3.1. (Table 3.18.). An illustration of the fragment positions relative to each candidate gene is shown in Figure 3.20.-3.22. The fragments were amplified using the annealing temperatures (Section 2.1.1.6.) shown in Table 3.18. and 40 cycles (Section 2.1.3.1). GNAI2 fragments 24 and 27 failed to amplify and were not used for polymorphism detection.

### 3.3.3. Polymorphism detection

Melting curve analysis was performed and subsequent selection of WAVE™ polymorphism detection conditions were made for each fragment (Section 2.1.3.2.). Polymorphism detection within each fragment was performed by dHPLC at the temperature(s) and corresponding predicted de-binding percent buffer B concentration shown in Table 3.19. (Section 2.1.3.3.). Putative polymorphisms (Figure 3.23.) were detected in 22 of the 50 remaining fragments (Table 3.19.).

### 3.3.4. Polymorphism verification and characterisation by sequencing

All 22 fragments containing putative polymorphisms were sequenced to verify and characterise the polymorphisms (Sections 2.1.3.4. – 9.). For each fragment, all samples showing polymorphic WAVE™ traces and at least one sample with a normal trace were

| Sequence Description | Accession # |
| --- | --- |
| GNAI2 gDNA | U73169 |
| GNAI2 cDNA | NM_002070 |
| CCR5, 2, and L2 gDNA | U95626 |
| CCR5 cDNA | NM_000579 |
| CCR5 promoter | AF017632 |
| CCR2a cDNA | NM_000648 |
| CCR2b cDNA | NM_000647 |
| CCR2 promoter | AF068265 |
| CCRL2a cDNA | AF015525 |
| CCRL2b cDNA | AF015524 |
| CCRL2c cDNA | NM_003965 |

**Table 3.17. Genbank sequences retrieved for the 3p candidate gene association study**

| Fragment | Forward oligonucleotide | $T_A$ | Reverse oligonucleotide | $T_A$ |
|---|---|---|---|---|
| ccr.f1 | AAGGTGTGCCTGAACCTCTTTAGC | 60 | TACTAGAAGATGACAGCACCAGCC | 60 |
| ccr.f2 | AGTGAGAGGAGTTGATGTGATGGG | 60 | GTGACAGCCTCTTTACCTTCTGC | 60 |
| ccr.f3 | CACGATGATAGGTGTCTGAGGC | 60 | GGAAGAACACACTTAGTGGTACAGC | 60 |
| ccr.f4 | ATGTGAGCAGAAGTGACATCTGCC | 60 | GCAGCAATAAAATCCTGCTGTGGC | 60 |
| ccr.f5 | GGTAAAGAATGGAAGGTGGAGAAGC | 60 | GATGAGGCAGATGTCACTTCTGC | 60 |
| ccr.f6 | CTGAACCAAAATGTGTCCCTTGCC | 60 | TGATAGTACCTGGCTAAAGGAGGC | 60 |
| ccr.f7 | TTATTGGTGCCTCCTTTAGCCAGG | 60 | GTCACTTTCTTGCTTCACACACTCC | 60 |
| ccr.f8 | ACTACAACAAAGCCAGGAGTGAGG | 60 | AAAGCTCTGTCTTCAAGCTCCAGC | 60 |
| ccrf9 | CCCAATGTACAATGTTCCTGACTGG | 60 | TTACTTCTGCCTCTGTGGATCTCC | 60 |
| ccrf10 | CAACTCTAAGGCTTAGGGGAAAGC | 60 | CACCCTTGAACTATTCACCTTCTGC | 60 |
| ccrf11 | GAAGGTGAATAGTTCAAGGGTGGG | 60 | GAGATGTGGACAGCATGTTGTGG | 60 |
| ccrf12 | TCCGGAGTGGTAGTTACATGCATG | 60 | ATATCAGTCAACAGTGAGACGGGC | 60 |
| ccrf13 | TATTGACCAGGAGTCAACTCAGGC | 60 | AGGCTGAAACAAGGAAGCACAGC | 60 |
| ccrf14 | TGGAGTAACACACACTGCAAAGC | 58 | AAGTGTAGAGCCTTTTGGGCACC | 60 |
| ccrf15 | GTCAAATTCTGGTCTCCCACAAAGC | 60 | GTAACCATGTGTCCTGTGTAAGACC | 60 |
| ccrf16 | GCTCATAAGTTCTTGGAATGGACCC | 60 | AAAGAAGCAACAGGGATGGTGAGG | 60 |
| ccrf17 | TCCAGGCTGTCTTTCACTGAATGC | 60 | TCATTTCAAAGGGAGGGAGAGACG | 60 |
| ccrf18 | CATCATAGTACAGGTAAGGTGAGGG | 60 | CAAGTCAATTGCTCAGCAACCAGC | 60 |
| ccrf19 | GTACCTCTAAGAGTAAGGTTGCTGG | 60 | TGCCCTTCCTTGGTTCATTCTACC | 60 |
| ccrf20 | CACCATTCCAAATCTGGGCTGTTC | 60 | TGGCATGTGACACTCTGAGTAGG | 60 |
| ccrf21 | TGACAGTAGCTGGTACCTGTTACC | 60 | GACAGCTGAGACGTTTTTCTCTCG | 60 |
| ccrf22 | CTTATGGCCATGTTGATGCTGAGC | 60 | AAACACAGCAGAGCAGAGTGATGG | 60 |
| ccrf23 | GTAGTAACACCCCACTTCAACCC | 60 | CTCTTTTTCCCACAGCTTCCAGC | 60 |
| ccrf24 | CGGTGATGAGAAGAACTAGCTAGC | 60 | TTTAGTGCAAGCAAGATGCTGGGC | 60 |
| ccrf25 | CTTGCTTGCACTAAAGAGAGCTGG | 60 | GCACAAAGCCAAACCCTTACTTAGC | 60 |
| g2.f1 | GGAACAGGATAGACAGGCTAATAGC | 60 | ATCAAGGCTTAAACCTGACTGGG | 58 |
| g2.f2 | CAGTTTTGGGAGTGAATACAGGCC | 60 | CAATGAAATCTGGGTTCCTCAGGG | 60 |
| g2.f3 | TAAGGGAACTGGTAATGAGGAGCC | 60 | CATTTCTCAAGCACAGCCCTTGC | 60 |
| g2.f4 | AATCCCCAGACAAGGACTTTGACC | 60 | TAAGAGGACTTAGGGGAGGATCAGC | 61 |
| g2.f5 | ATCCTTCTCTGCAGGCATCACTCC | 61 | TTTGCAAAAGGCCAGCTGGGATGC | 61 |
| g2.f6 | ACTAGCTGAATGAGGAGTTGAGGG | 60 | ATGGTCCAGTCTTGCATTTCCTCG | 60 |
| g2.f7 | TGAAAGCATGTTCCTTAGGGCTGC | 60 | GAGACCATGCTAAGAATCACGGC | 60 |
| g2.f8 | TGATTCTTAGCATGGTCTCCTGGC | 60 | AACAGACCAACAGACTCCACACC | 60 |
| g2.f9 | ATCTCTGACTTGGCAGTAGGAAGC | 60 | CTTCTCTTATCCCAGGAAGGTGC | 60 |
| g2.f10 | TTACTGTGGTGGCCGAGGAAATGC | 61 | TGGACAAAGTCTCCCGCTACATG | 60 |
| g2.f11 | TGTGGTAAGTTCAGAGTTCTGGGC | 60 | CAGATGTTTCACCAGCCTTGGTC | 60 |
| g2.f12 | AACTTTACTCTGGGATCACCTCCC | 60 | TACCAGGACTCTCACAGCATTGC | 60 |
| g2.f13 | ATCTGAGAACCACAGGCTGAGCTG | 61 | CTGCTGTGGTCATCTCATTTGAAGG | 60 |
| g2.f14 | CTGTCCATTGCTCTTCATCTGCTC | 60 | CCACAGGTCTTGTCACTTGATGC | 60 |
| g2f15 | TGAGTGTGTCTGCGTGTTTACACC | 60 | CCTTACAACAGCTCCCTTAGAAGC | 60 |
| g2f16 | TGGACTAGGCAGCATACATGTGG | 60 | ACAAAGCTGCCTTTGTGTCCAGC | 60 |
| g2f17 | ACACAAAGGCAGCTTTGTGAGGC | 60 | TAGCTGTCTGCACTCAGGGAAGC | 61 |
| g2f18 | GAGAGGATCTGAAAATGGGAGGC | 60 | CCAGATTTCTGCAAAGATGCTGGG | 60 |
| g2f19 | CTGATTCATGCCACCTTCTATTGCC | 60 | TAGTAGCTTCACTCTCCCCATCTC | 60 |
| g2f20 | AGAGATGGGGAGAGTGAAGCTAC | 60 | CCCACTCTTAGTAGGTGATGTGG | 60 |
| g2f21 | ACCAGAGCAATGCTGTGAGAGTC | 60 | TGGAACTGGGATTCGAACTCAGG | 60 |
| g2f22 | TCATCCAGACCTGAGTTCGAATCC | 60 | ACTTGCTGTGAACAGCTGCCTGAG | 61 |
| g2f23 | AGAGTTTGGTGGGAAGGAGTCTC | 60 | CCTGATTCACCTAGTCTCTCAGC | 60 |
| g2f24 | AAATGACAAAGACACGCCTCCTCC | 60 | GTTCTTGTCGATCATCTTAGAGCGC | 60 |
| g2f25 | AAGAATGCCTCCCGGTAGATTGG | 60 | TTCTGCGAGTTGGGCGAGTTAGG | 61 |
| g2f26 | AAGGGGTTTATGCTAAACACCGGG | 60 | TTAGCTGGGTGGCTTCTCATATCC | 60 |
| g2f27 | TTCCCTGAGTGCAGACAGCTAGG | 61 | AAAGCTTGCCTCTCTCTTCCTAGC | 60 |

Table 3.18. Oligonucleotide sequences and corresponding annealing temperature

for each polymorphism detection fragment across the 3p candidate genes

144
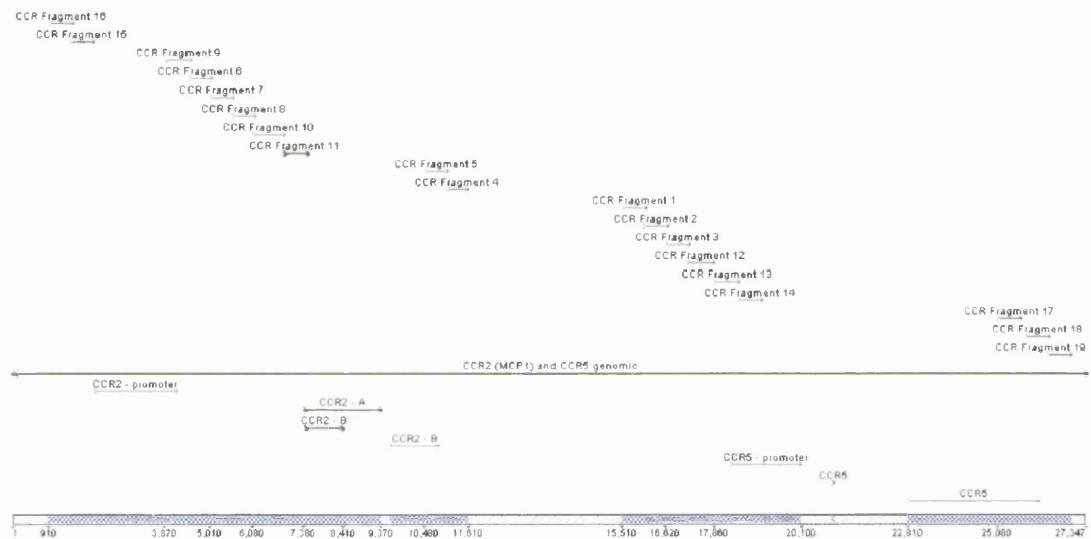
**Figure 3.20. Polymorphism detection fragments designed across CCR5 and CCR2**
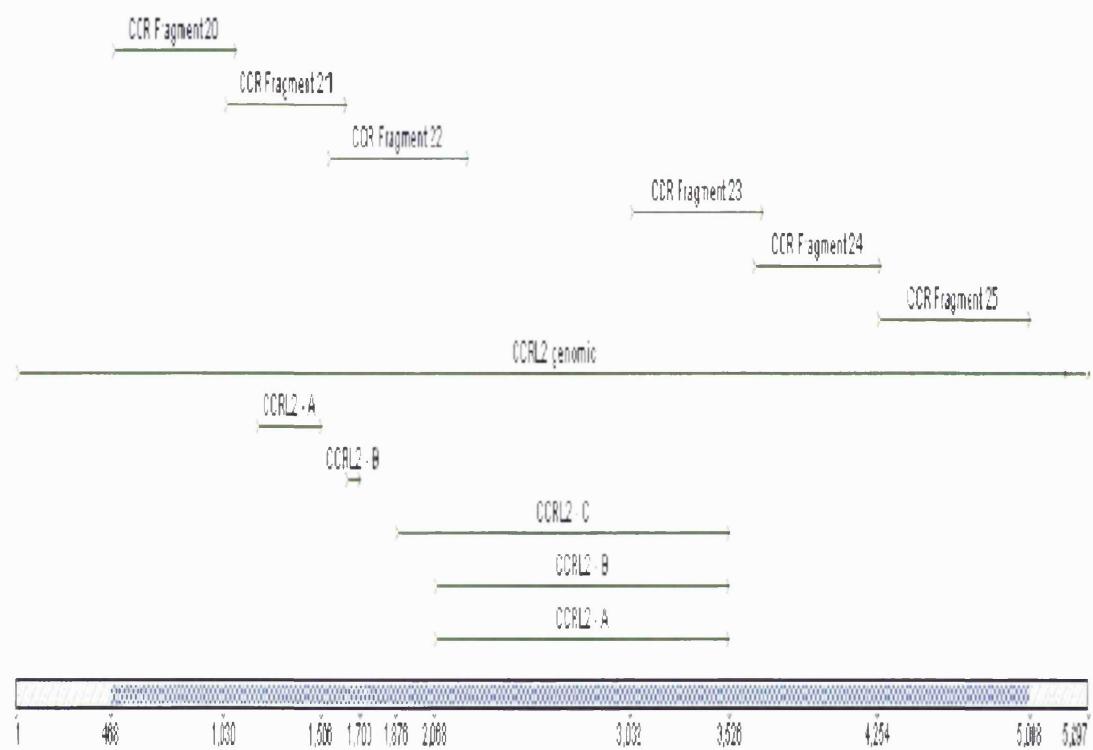
**Figure 3.21. Polymorphism detection fragments designed across CCRL2**
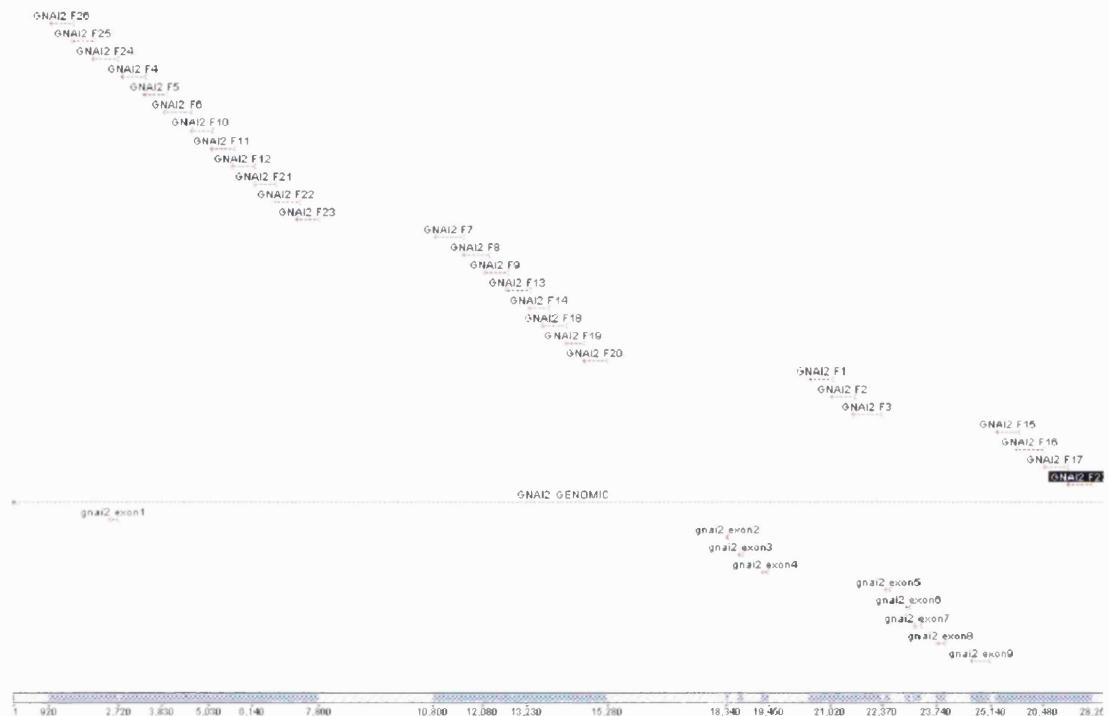
**Figure 3.22. Polymorphism detection fragments designed across GNAI2**

| Fragment | Temperature (C) | Buffer B (%) | Unique Traces | Fragment | Temperature (C) | Buffer B (%) | Unique Traces |
|---|---|---|---|---|---|---|---|
| ccr.f1 | 62 | 64 | 0 | g2.f1 | 62 | 63 | 2 |
| ccr.f2 | 61 | 64 | 0 | g2.f2 | 63 | 64 | 2 |
| ccr.f3 | 56 | 64 | 2 | g2.f3 | 64 | 64 | 1 |
| ccr.f3 | 61 | 64 | 0 | g2.f4 | 63 | 64 | 0 |
| ccr.f4 | 57 | 64 | 0 | g2.f5 | 64 | 65 | 0 |
| ccr.f4 | 61 | 64 | 4 | g2.f6 | 64 | 66 | 0 |
| ccr.f5 | 56 | 64 | 1 | g2.f7 | 59 | 65 | 0 |
| ccr.f5 | 60 | 64 | 0 | g2.f7 | 63 | 64 | 0 |
| ccr.f6 | 58 | 64 | 0 | g2.f8 | 63 | 65 | 1 |
| ccr.f7 | 57 | 65 | 1 | g2.f9 | 65 | 65 | 0 |
| ccr.f7 | 60 | 64 | 0 | g2.f10 | 62 | 64 | 0 |
| ccr.f8 | 60 | 64 | 0 | g2.f11 | 60 | 65 | 0 |
| ccr.f9 | 59 | 65 | 1 | g2.f11 | 62 | 64 | 0 |
| ccr.f9 | 63 | 61 | 1 | g2.f11 | 65 | 57 | 0 |
| ccr.f10 | 58 | 65 | 1 | g2.f12 | 61 | 65 | 0 |
| ccr.f10 | 61 | 61 | 1 | g2.f12 | 63 | 64 | 0 |
| ccr.f11 | 60 | 65 | 0 | g2.f13 | 63 | 64 | 0 |
| ccr.f12 | 53 | 65 | 0 | g2.f13 | 65 | 61 | 0 |
| ccr.f12 | 56 | 65 | 0 | g2.f14 | 58 | 65 | 0 |
| ccr.f12 | 61 | 61 | 1 | g2.f14 | 61 | 64 | 0 |
| ccr.f13 | 59 | 65 | 1 | g2.f15 | 57 | 65 | 0 |
| ccr.f14 | 53 | 65 | 2 | g2.f15 | 64 | 61 | 0 |
| ccr.f14 | 58 | 65 | 1 | g2.f16 | 64 | 65 | 0 |
| ccr.f14 | 61 | 61 | 0 | g2.f16 | 69 | 57 | 0 |
| ccr.f14 | 64 | 57 | 0 | g2.f17 | 65 | 65 | 0 |
| ccr.f15 | 60 | 65 | 1 | g2.f17 | 68 | 65 | 0 |
| ccr.f16 | 56 | 65 | 0 | g2.f18 | 62 | 65 | 0 |
| ccr.f16 | 59 | 61 | 1 | g2.f18 | 64 | 61 | 0 |
| ccr.f17 | 59 | 65 | 0 | g2.f19 | 62 | 65 | 1 |
| ccr.f18 | 55 | 65 | 0 | g2.f20 | 60 | 65 | 0 |
| ccr.f18 | 59 | 61 | 0 | g2.f20 | 63 | 61 | 0 |
| ccr.f19 | 56 | 65 | 0 | g2.f20 | 65 | 57 | 1 |
| ccr.f19 | 59 | 61 | 1 | g2.f21 | 60 | 65 | 0 |
| ccr.f19 | 62 | 57 | 1 | g2.f21 | 63 | 61 | 0 |
| ccr.f20 | 62 | 65 | 0 | g2.f22 | 62 | 65 | 0 |
| ccr.f21 | 58 | 65 | 0 | g2.f23 | 59 | 65 | 0 |
| ccr.f21 | 63 | 61 | 0 | g2.f23 | 64 | 65 | 1 |
| ccr.f22 | 57 | 65 | 0 | g2.f25 | 59 | 65 | 0 |
| ccr.f22 | 60 | 61 | 1 | g2.f25 | 62 | 61 | 1 |
| ccr.f23 | 53 | 65 | 0 | g2.f25 | 65 | 57 | 1 |
| ccr.f23 | 57 | 65 | 0 | g2.f26 | 63 | 65 | 0 |
| ccr.f23 | 60 | 61 | 0 | | | | |
| ccr.f24 | 57 | 65 | 1 | | | | |
| ccr.f24 | 60 | 61 | 1 | | | | |
| ccr.f24 | 64 | 57 | 3 | | | | |
| ccr.f25 | 58 | 65 | 0 | | | | |
| ccr.f25 | 61 | 61 | 0 | | | | |
| ccr.f25 | 65 | 57 | 0 | | | | |

**Table 3.19. Results of polymorphism detection within 3p IBD candidate gene fragments**

The table contains the temperature at which dHPLC was conducted on each fragment and the corresponding percentage buffer B at which the fragment should be released from the capillary. Some fragments were analysed at more than one melting temperature.
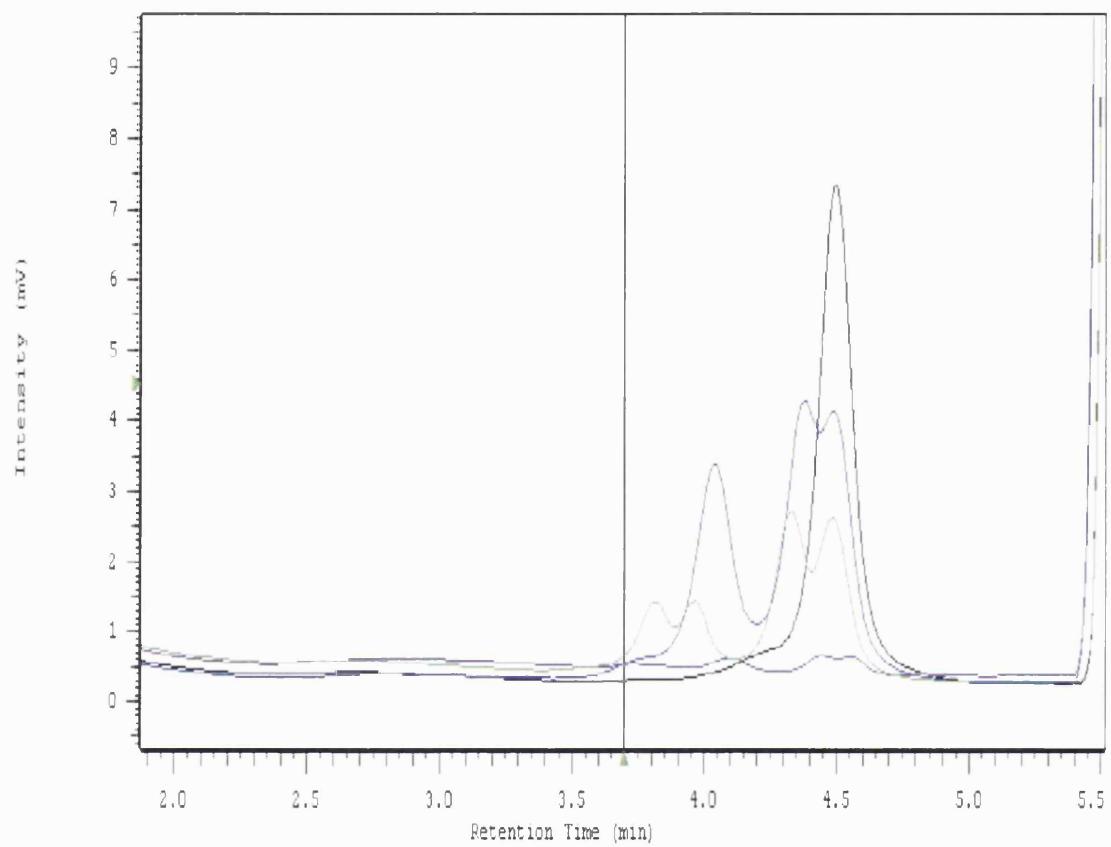
**Figure 3.23. Typical WAVE™ dHPLC trace**

The black trace with a single homoduplex peak was created by the amplification of a fragment without a polymorphism. The blue and green traces with extra heteroduplex peaks were created by the amplification of the same fragment with one and two polymorphisms, respectively.

sequenced. From the sequence data (Figure 3.24.), 36 SNP and 2 deletion/insertion polymorphisms were identified (Table 3.20.).

### 3.3.5. Selection of polymorphisms and subsequent assay design for genotyping

Of the 38 identified polymorphisms, 9 were chosen for subsequent genotyping and association analysis on the basis of firstly their position and secondly their predicted allele frequency which was estimated from the number of heterozygotes seen in the 12 control samples during polymorphism detection. The highest frequency polymorphisms every 2kb to 6kb across each candidate gene were chosen. In addition to these 9 polymorphisms, 2 from a previous study (Mummidi *et al.*, 1998), CCR5Δ32 and CCR5-627, and one found during the initial alignment of the CCRL2 alternative spliced cDNA sequences, CCRL2-524, were also selected. Assays were designed to genotype the 12 polymorphisms as described in Sections 2.1.4.1. and 2.1.4.2. The oligonucleotides and a description of each assay are shown in Tables 3.21. and 3.22., respectively.

### 3.3.6. GNAI2 initial association study

The 4 selected polymorphisms spanning GNAI2 were genotyped (Sections 2.1.4.3. – 5.) across the Oxford TDT2 cohort (Section 2.2.2.) generating approximately 2,300 genotypes (Figure 3.25.). The genotype and pedigree data were analysed with Pedcheck as described in Section 2.1.6.1. For each family and marker showing misinheritance, the proband genotype was removed. The genotype data was analysed for association with IBD phenotypes as described in Section 2.1.6.5. Markers G2F25-298, G2F23-313, and G2F19-359 were in almost complete inter-marker LD and showed positive two-point and haplotype association to CD prior to correcting for multiple testing (Figure 3.26.). Marker G2F2-103 did not show two-point association with CD. Yet, because G2F2-103 exhibited partial inter-marker LD with the other markers, it showed positive

150
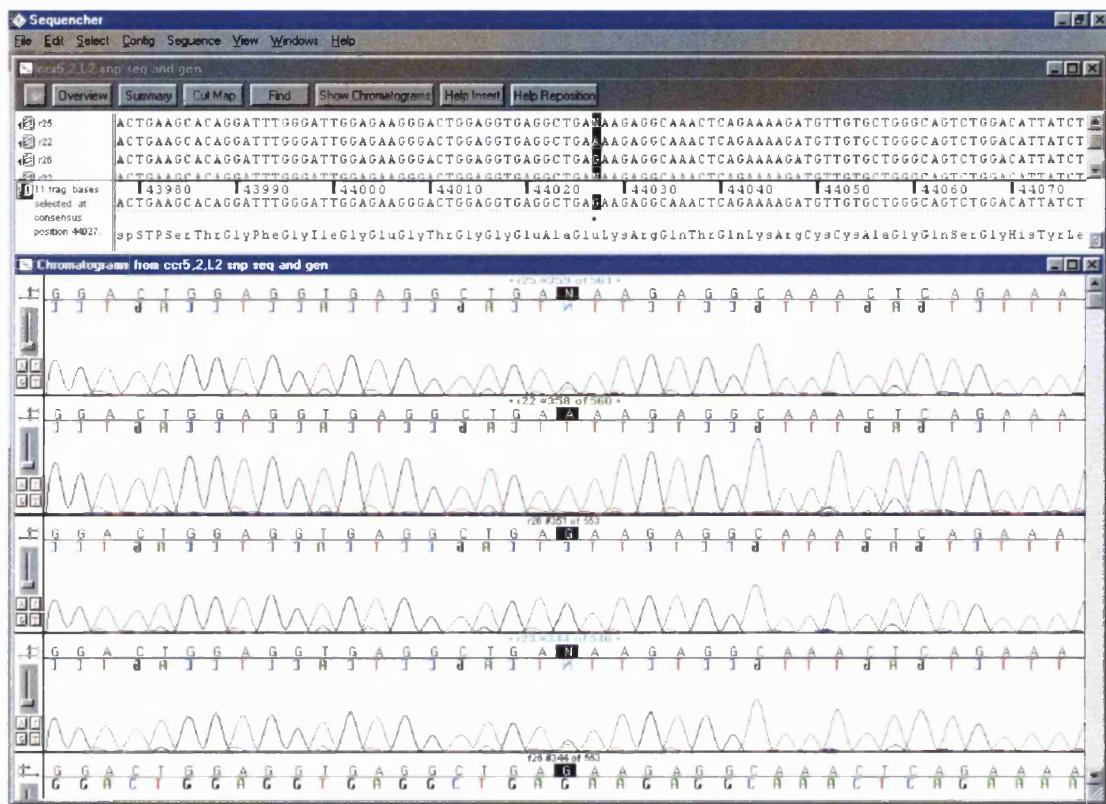
**Figure 3.24. An example of typical heterozygous sequencing results**

The base highlighted is an A/G SNP with all three possible genotypes presented (A = A/A, G = G/G, and N = A/G).

| Fragment | Polymorphism ID | Polymorphism |
|---|---|---|
| ccr.f3 | ccrf3-509 | del/ATC |
| ccr.f3 | ccrf3-541 | A/G |
| ccr.f4 | ccrf4-35 | A/G |
| ccr.f4 | ccrf4-53 | A/C |
| ccr.f4 | ccrf4-96 | G/C |
| ccr.f4 | ccrf4-142 | A/G |
| ccr.f5 | ccrf5-381 | T/A |
| ccr.f7 | ccrf7-370 | A/G |
| ccr.f7 | ccrf7-471 | A/G |
| ccr.f9 | ccr9-335 | A/G |
| ccr.f9 | ccrf9-479 | T/A |
| ccr.f10 | ccrf10-492 | C/T |
| ccr.f10 | ccrf10-624 | A/G |
| ccr.f10 | ccrf10-697 | G/A |
| ccr.f12 | ccrf12-407 | G/A |
| ccr.f13 | ccrf13- | A/G |
| ccr.f14 | ccrf14-111 | A/C |
| ccr.f14 | ccrf14-468 | A/C |
| ccr.f14 | ccrf14-482 | C/T |
| ccr.f14 | ccrf14-486 | del/CTAT |
| ccr.f15 | ccrf15-444 | A/G |
| ccr.f16 | ccrf16-475 | T/C |
| ccr.f19 | ccrf19-421 | A/G |
| ccr.f19 | ccrf19-496 | T/C |
| ccr.f22 | ccrf22-363 | A/G |
| ccr.f24 | ccrf24-232 | G/T |
| ccr.f24 | ccrf24-376 | A/G |
| ccr.f24 | ccrf24-521 | T/C |
| ccr.f24 | ccrf24-554 | A/G |
| g2.f1 | g2f1-127 | A/G |
| g2.f1 | g2f1-535 | T/G |
| g2.f2 | g2f2-103 | A/G |
| g2.f3 | g2f3-455 | T/C |
| g2.f8 | g2f8-140 | T/C |
| g2.f19 | g2f19-359 | T/C |
| g2.f20 | g2f20-104 | T/C |
| g2.f23 | g2f23-313 | G/C |
| g2.f25 | g2f25-298 | A/T |

**Table 3.20. Polymorphisms identified within 3p IBD candidate gene fragments**

The Polymorphism ID is combination of the fragment name and the position of the polymorphism in basepairs from the start of the corresponding forward fragment primer shown in Table 3.17.

| SNP Name | Forward Primer | $T_A$ | Reverse Primer | $T_A$ |
|---|---|---|---|---|
| ccrf16-475 | AGGGAAAAGAGCTTTCCTTCATCTGCTTCTTAATGGC | 64 | ccr.f16 reverse | 60 |
| ccrf10-697 | TGCCTGGGGAATCTTAAGGTGTGG | 61 | ccr.f10 reverse | 60 |
| ccrf4-35 | ccr.f4 forward | 60 | CTCAGTGGCTTATCACAGCAAACACAGACTTCGTG | 66 |
| ccrf3-541 | ccr.f3 forward | 60 | ccr.f3 reverse | 60 |
| ccr5-627 | GCCTTACTGTTGAAAAGCCCTGTG | 60 | GAGCTTAAAATAAGCTGGAGAATAGATCTCTGGTCTGCA | 64 |
| ccr5-32del | TCACTTGGGTGGTGGCTGTGTTTGCGTCTC | 66 | GTAGCAGATGACCATGACAAGCAGCGGCAG | 66 |
| ccrf19-421 | ccr.f19 forward | 60 | ccr.f19 reverse | 60 |
| ccrl2-524 | ATTTCTCCACAGGGCAGTCTGAAG | 60 | AAAAATGCAATATTGTAGGGCGCCCACATCAGAAGGAAGACCACCATTA | 67 |
| g2f25-298 | g2f25 forward | 60 | g2r25 reverse | 60 |
| g2f23-313 | g2f23 forward | 60 | g2r23 reverse | 60 |
| g2f19-359 | g2f19 forward | 60 | g2r19 reverse | 60 |
| g2f2-103 | g2.f2 forward | 60 | g2.r2 reverse | 60 |

**Table 3.21. Oligonucleotide sequences and corresponding annealing temperature for the 3p candidate gene polymorphisms selected for genotyping**

The oligonucleotide names in place of actual sequence correspond to the polymorphism detection fragment oligonucleotide sequences shown in Table 3.18.

| Fragment | SNP Name | Assay type | Endonuclease | Digest fragment lengths | Alleles | Frequency |
|---|---|---|---|---|---|---|
| ccr.f16 | ccrf16-475 | MM PCR-RFLP | Hae III | 147(36,111) | T(C*) | 83/17 |
| ccr.f10 | ccrf10-697 | PCR-RFLP | Nla III | 55,343,45,181(94,87) | G(A*) | 83/17 |
| ccr.f4 | ccrf4-35 | PCR-RFLP | Aci I | 177(33,144) | A(G*) | 60/40 |
| ccr.f3 | ccrf3-541 | PCR-RFLP | Msl I | 586(536,50) | A*(G) | 55/45 |
| Previous study | ccr5-627 | MM PCR-RFLP | Pst I | 255(220,35) | T(C*) | 60/40 |
| Previous study | ccr5-32del | PCR | | 191(159) | (del) | 85/15 |
| ccr.f19 | ccrf19-421 | PCR-RFLP | Hha I | 594(421,173) | A(G*) | 54/46 |
| cDNA alignment | ccrl2-524 | PCR-RFLP | EcoR I | 800(520,280) | A(T*) | 79/21 |
| g2.f25 | g2f25-298 | PCR-RFLP | Nla III | 584(299,285) | A(T*) | 90/10 |
| g2.f23 | g2f23-313 | PCR-RFLP | Bsm I | 602(314,288) | G*(C) | 90/10 |
| g2.f19 | g2f19-359 | PCR-RFLP | Bsa I | 141,392(207,185) | T*(C) | 90/10 |
| g2.f2 | g2f2-103 | PCR-RFLP | Ban II | 551(105,446),60 | A(G*) | 65/35 |

**Table 3.22. Genotyping assays for the 3p candidate gene polymorphisms selected for association analysis**

MM is an abbreviation for mismatch. The digested fragment lengths are in basepairs. The digest fragment lengths in brackets correspond to the fragments produced when restriction endonuclease digestion caused by an allele of the polymorphism occurs. The allele in brackets is the allele that creates the RFLP site. The allele with an asterisk is the more frequent allele. The allele frequencies shown GNAI2 and CCR SNPs are those generated from the genotyping of the Oxford TDT2 and TDT3 cohorts, respectively.
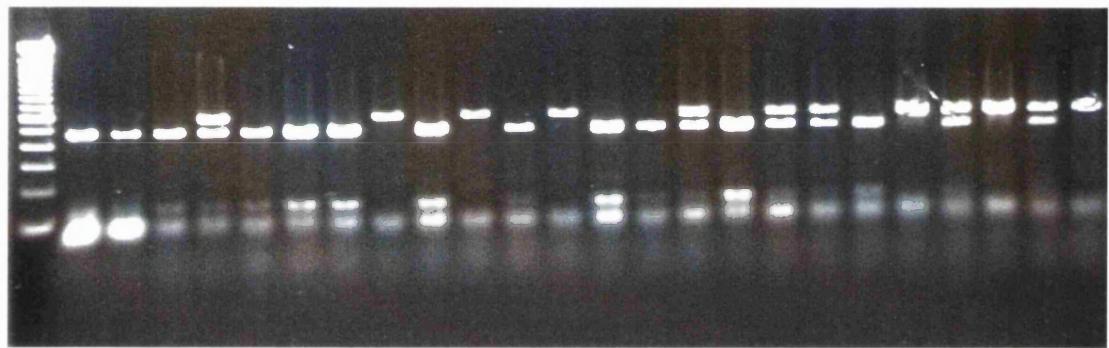
**Figure 3.25. A typical PCR-RFLP genotyping agarose gel electrophoresis result**

The assay shown here is G2F2-103 (Table 3.22.). Rows A and C of a 96-well assay are

shown above, loaded A1, C1, A2, C2, etc. 100bp size standard was loaded in the first

well to size digested and undigested fragments.

GNAI 2

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | |

Markers: G2F25-298, G2F23-313, G2F8-140, G2F20-104 / G2F19-359, G2F2-103 / G2F1-535 / G2F1-127 / G2F1-53, G2F3-455

-(~6kb)-  -(~6kb)-  -(~6kb)-

10%    10%    10%    35%

IBD = 0.2      IBD = 0.3          IBD = 0.3      IBD = 0.5
CD = 0.004     CD = 0.02          CD = 0.007     CD = 0.3
UC = 0.6       UC = 0.6           UC = 0.6       UC = 0.6     TDT2 association
        IBD = 0.2      IBD = 0.2          IBD = 0.4              analysis
        CD = 0.007     CD = 0.01          CD = 0.007          with 185 ped.
        UC = 0.6       UC = 0.5           UC = 0.2
        LD = 1.9e$^{-126}$    LD = 9.2e$^{-131}$    LD = 3.6e$^{-8}$

----------------------------------------------------------------------

IBD = 0.2                                    IBD = 0.6
CD = 0.08                                    CD = 0.8
UC = 1.0                                     UC = 0.6     TDT3 association
                    IBD = 0.4                                 analysis
                    CD = 0.1                                with 333 ped.
                    UC = 0.8
                    LD = 1.5e$^{-12}$

**Figure 3.26. GNAI2 association study results**

The p-values resulting from two-point association analysis are shown directly under the
marker analysed. The p-values resulting from haplotype association analysis are shown
between each pair of markers tested. Significant p-values are displayed with black and
non-significant p-values are shown in red. The inter-marker LD for each adjacent
marker pairs is also shown between the markers tested. The allele frequency of the rare
allele for each genotyped assay is shown in blue. The distance between each genotyped
assay is shown in brackets between adjacent markers. The other polymorphisms
discovered by polymorphism detected during the project are shown in smaller font
relative to the polymorphisms genotyped.

association to CD in haplotype association analysis. Simulations with 1000 equivalent data sets resulted in 11 and 7 showing greater two-point and haplotype association to CD, respectively. No association was detected between GNAI2 polymorphisms and either UC or IBD.

### 3.3.7. GNAI2 extension study

Only G2F25-298 and G2F2-103 were used for the GNAI2 extension study. With G2F25-298, G2F23-313 and G2F19-359 showing almost complete inter-marker LD, only one marker needed to be genotyped and therefore G2F25-298 was selected because it was furthest away from G2F2-103 and at the start of the gene. The 2 polymorphisms were genotyped (Sections 2.1.4.3. – 5.) across the Oxford TDT3 families not included in Oxford TDT2 cohort (Section 2.2.2.). The genotyping generated approximately 900 additional genotypes (Figure 3.25.). The genotype and pedigree data was analysed with Pedcheck as described in Section 2.1.6.1. For each family and marker showing misinheritance, the proband genotype was removed. The genotype data from the initial GNAI2 association study (Section 3.3.6.) and this study were combined and analysed for association with IBD phenotypes (Section 2.1.6.5.). No association was seen between both polymorphisms and any IBD phenotype (Figure 3.26.).

### 3.3.8. CCR2, CCR5 and CCRL2 association study

The 8 chosen polymorphism spanning CCR2, CCR5 and CCRL2 were genotyped (Sections 2.1.4.3. – 5.) across the Oxford TDT3 cohort (Section 2.2.2.) generating approximately 8,000 genotypes (Figure 3.25.). The genotype and pedigree data was analysed with Pedcheck as described in Section 2.1.6.1. The proband genotype was removed for each family and marker showing misinheritance. The genotype data was analysed for association with IBD phenotypes as described in Section 2.1.6.5. All

157

markers showed some form of inter-marker LD. Marker CCR5Δ32 in CCR5 and

CCRF10-697 in CCR2 showed positive two-point association to UC prior to correcting

for multiple testing (Figure 3.27. and 3.28.). The association with CCR5Δ32 decreased

in significance when conducting haplotype analysis with adjacent markers, whilst the

association with CCRF10-697 increased. Simulations with 1000 equivalent data sets

resulted in 33 and 93 sets showing greater two-point and haplotype association to UC,

respectively. Positive haplotype association, yet no two-point association, was observed

between CCR5-627 and CCR5Δ32 and CD. However, simulations with 1000

equivalent data sets resulted in 401 data sets showing greater haplotype association to

CD. No association was detected between CCR5, CCR2 and CCRL2 polymorphisms

and IBD.


### 3.3.9. CCR5 and GNAI2 RH mapping

CCR5-627 and G2F2-103 were amplified by PCR across the G3 RH panel using a 60C

$T_A$ and 35 cycles (Section 2.1.2.1.). The results were visualised and scored as described

in Sections 2.1.2.2. and 2.1.2.3. (Figure 3.29.) Both markers exhibited high-fidelity

amplification across the RH panel (Figure 3.3.). The G3 data was sent to the Stanford

RH server for analysis against the existing framework map (Section 2.1.2.3.1.). Marker

CCR5-627 mapped between markers D3S3647 and D3S1767 and marker G2F2-103

mapped between markers D3S1573 and D3S1578 (Table 3.4.).


## 3.4. Chromosome 6p replication study


### 3.4.1. Markers and oligonucleotides

The 13 STR markers used in the Hampe et al. (1999) extension study, that detected

significant linkage to chromosome 6p (Section 1.4.3.5.), were selected for analysis in
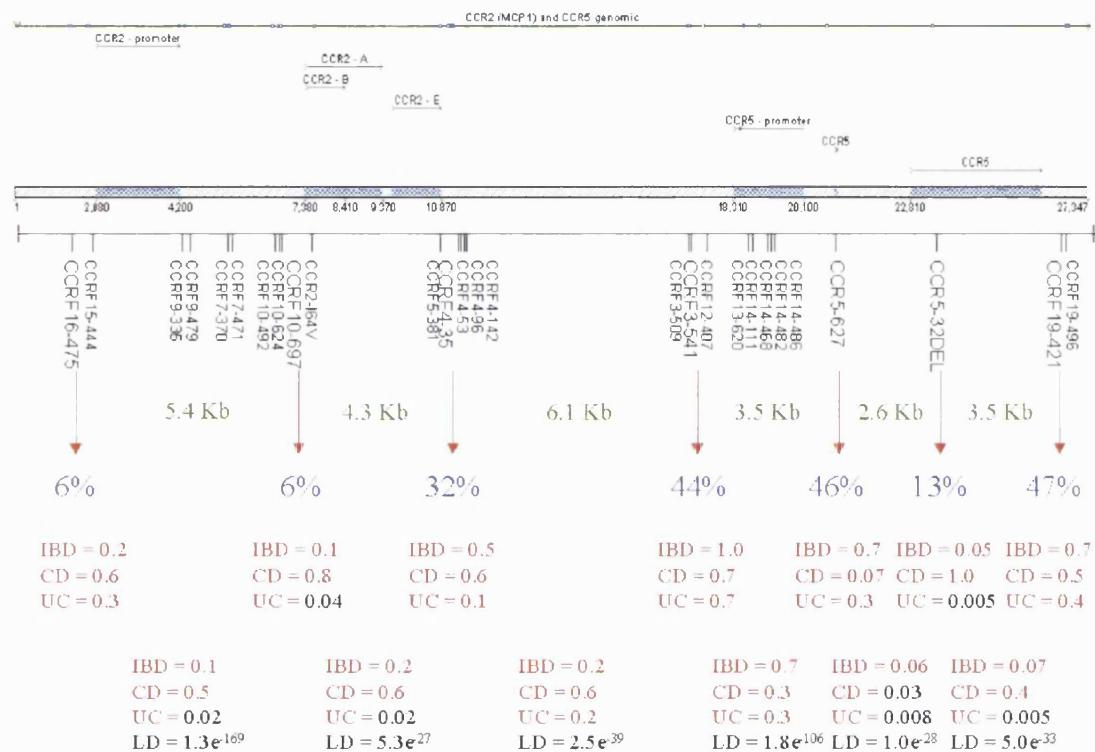
**Figure 3.27. CCR2 and CCR5 association study results**

The p-values resulting from two-point association analysis are shown directly under the marker analysed. The p-values resulting from haplotype association analysis are shown between each pair of markers tested. Significant p-values are displayed with black and non-significant p-values are shown in red. The inter-marker LD for each adjacent marker pair is also shown between the markers tested. The allele frequency of the rare allele for each genotyped assay is shown in blue. The distance between each genotyped assay is shown in green. The other polymorphisms discovered by polymorphism detected during the project are shown in smaller font relative to the polymorphisms genotyped.
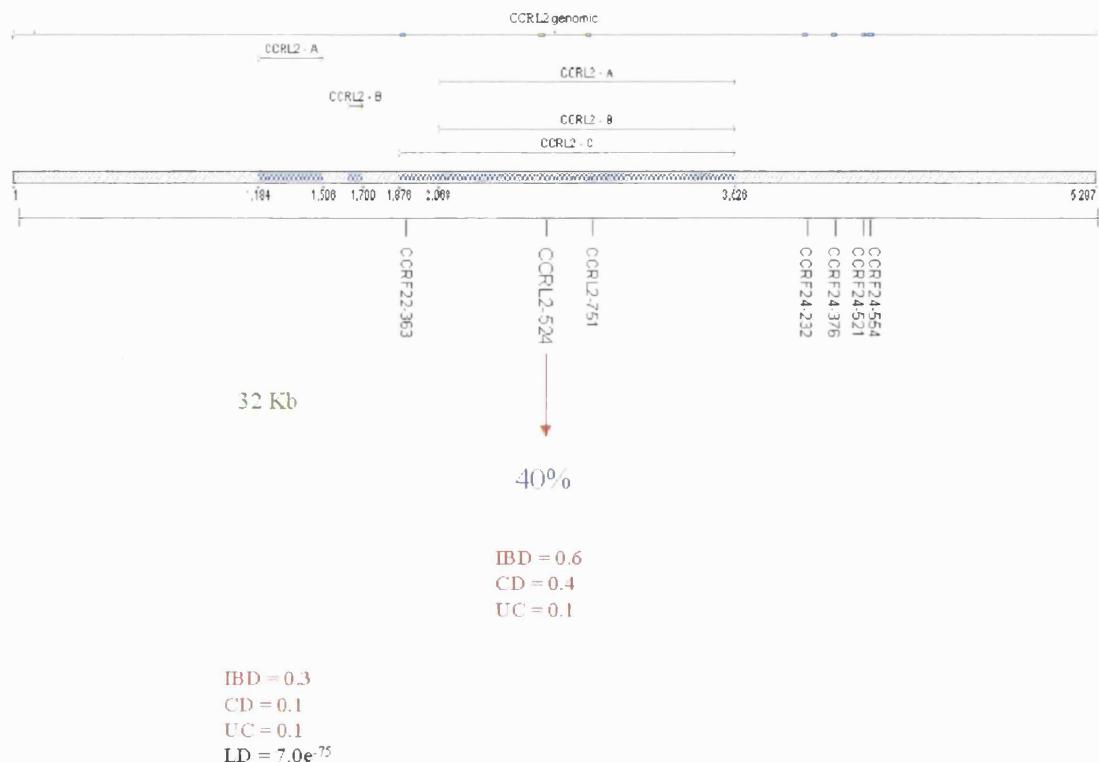
**Figure 3.28. CCRL2 association study results**

The p-values resulting from two-point association analysis are shown directly under the marker analysed. The p-values resulting from haplotype association analysis, the inter-marker LD, and the distance between CCRL2-524 and CCRF19-421 (Figure 3.27.) are shown to the left of the marker. Significant p-values are displayed with black and non-significant p-values are shown in red. The allele frequency of the rare allele is shown in blue. The other polymorphisms discovered by polymorphism detected during the project are shown in smaller font relative to the polymorphisms genotyped.

G2F2-103  00000001000000000000000000000000000100000010000001000000000000001100100000000001000001
CCR5-672 00000000000001000001000000000000000000000001000000000000000000010000101001000000001000010

## Figure 3.29.  G3 RH panel results for CCR5 and GNAI2

For each data vector, the 83 human-hamster hybrids are ordered in the official G3 order.

The results are coded 0 for negative, 1 for positive and R for ambiguous.

the replication study (Table 3.23.). A pair of oligonucleotides was designed (Section 2.1.5.2.) to amplify across the tandem repeat of each marker (Table 3.24.). In order to choose an appropriate fluorescent label for each marker, the markers were placed into panels (Section 2.1.5.3.) based upon the length of each amplicon, determined from the associated Genebank entry. The $T_A$ of each oligonucleotide for each marker was calculated (Section 2.1.1.6.). The panel layout, including the fluorescent label attached to each marker and the $T_A$ of each oligonucleotide is shown in Table 3.24.

### 3.4.2. PCR parameter optimisation and amplification across RH panels

The 13 STR markers were amplified by PCR across the G3 RH panel using a 54C $T_A$ and 30 cycles (Section 2.1.2.1.). The results were visualised and scored as described in Sections 2.1.2.2. and 2.1.2.3. (Figure 3.30.).

All markers exhibited high-fidelity amplification across the RH panels (Figure 3.3.) at the predicted PCR conditions except D6S461, D6S257, D6S301, and D6S105. The amplification of D6S301 and D6S105 resulted in spurious fragments ranging from 100bp to 200bp. Increasing the $T_A$ for the amplification of these markers failed to remove the non-specific bands and thus the markers were removed from the study. The initial amplification of the other two markers was of weak intensity. These markers were subsequently amplified with increased number of PCR cycles in increments of 5 additional cycles, up to 40 cycles. After optimisation, D6S461 and D6S257 amplified well with 35 cycles and 40 cycles of PCR, respectively.

### 3.4.3. RH map construction and comparison

The G3 data from the 11 remaining 6p STR markers was sent to the Stanford RH server for analysis against the existing framework map (Section 2.1.2.3.1.). Based upon the

| Marker | Map position | Heterozygousity |
|--------|--------------|-----------------|
| D6S344 | 0.00 | 0.75 |
| D6S309 | 12.68 | 0.84 |
| D6S470 | 16.82 | 0.80 |
| D6S289 | 28.54 | 0.76 |
| D6S461 | 38.74 | 0.72 |
| D6S105 | 40.87 | 0.81 |
| D6S276 | 43.00 | 0.83 |
| D6S265 | 43.00 | 0.79 |
| D6S291 | 48.10 | 0.70 |
| D6S426 | 59.04 | 0.85 |
| D6S271 | 64.96 | 0.86 |
| D6S257 | 78.48 | 0.87 |
| D6S301 | 109.71 | 0.76 |

**Table 3.23. STR markers selected for the 6p replication study**

The marker order, distances and heterozygosity were retrieved from the chromosome 6

Marshfield sex-average genetic map. The map position of each marker is relative to its

distance from the most p-terminal marker chosen for the study.

| Marker | Forward Oligonucleotide | $T_A$ (C) | Reverse Oligonucleotide | $T_A$ (C) | Est. Length |
|---|---|---|---|---|---|
| D6S470 | GAGGGTATCTCTTAAGCGATC | 56 | ACATCCATCACCACACTTGC | 55 | 106 |
| D6S291 | ATATCCATCCGGCATTCAGG | 55 | TGTGGTGATGGTTTCACAGG | 55 | 150 |
| D6S276 | ATCATCCCCAGAAGGAACAC | 55 | GTTTTGCTACAGCACAGTGAG | 56 | 243 |
| D6S461 | CTCCTACGCATTTCTGTAGG | 55 | GGGTGTCTACTGCAAAGAAAC | 56 | 298 |
| D6S289 | CAAGAATGGTCTGAACTCGG | 55 | GCTGATAGGCTAGTGTACTG | 55 | 349 |
| D6S105 | AAAGTGCTGGGATTACAGGC | 55 | TCAAGGAAGAGAGACCATGC | 55 | 143 |
| D6S265 | ATTATAGTCACCCTACTGTGC | 54 | GATCTAATCGAGGTAAACAGC | 54 | 190 |
| D6S426 | CCAAATCTGTTGAGTTCTCCC | 56 | TGGAACCAGAAACTGTGGC | 55 | 265 |
| D6S257 | GACTCAATTGAGAACTCGTCC | 56 | AGCAGATAGGCAACACCATC | 55 | 318 |
| D6S301 | CTCAGCCTCCACAATCATATG | 56 | CAGGATTCTCCAGAGAAACAG | 56 | 126 |
| D6S344 | AGAAGCAAAAGGATCCCTCC | 55 | GCTAATGCATGACAATAATATTTCC | 53 | 169 |
| D6S271 | AGATTGCACCACTGTACACC | 55 | TTGTGGTGGGTGTTACAGC | 55 | 258 |
| D6S309 | CTAAAACCTGAAAGGCAGAGG | 56 | GGGCAATTTCATCCTTGTGC | 55 | 300 |

**Table 3.24. Initial panel for the fluorescent genotyping of 6p STR markers**

The estimated length of each marker was calculated using the Genebank sequence starting from the start of the 'forward' oligonucleotide and ending at the end of the 'reverse' oligonucleotide. The colour of the background for each marker corresponds to the fluorescent dye label (blue = FAM, yellow = HEX, green = TET) and the colour used to display the respective dye in both ABI Genescan and Genotyper software when using filter set C.

164

```
D6S257   00010101000010R011001000000000000100110000001100010100110000000010101000100010001100
D6S265   0000010100000000011100000000011000000100011000000011110001001000010100101011001000000
D6S271   00000000001000001000000000101000000100100000001001100101000100001010000001000000011
D6S276   0010000100000000110000000010000000010000000000010001000001000000101001001000000000010
D6S289   000000000011000010000R0001000001000000000010000001000000000000000000000100011000000
D6S291   0000000110000001000000000010100010001010000000001000000000000101010111110101010000000
D6S309   0010001000110010000000000000010000001000000000100000100000R00001000000000000000000000010
D6S344   1010000000010000100000000010R0000001000000001000001000000010100011000100000000000000
D6S426   10000100000000001100010000010100000R0010R0000000000001000000110000101000000001010010110
D6S461   0010000100000000110000000010000000010000000000010000000000001010010010000000000010
D6S470   000100R0001000001000R000011001000000000000000000100R0000000100R010000000R00100000000R10
```

## Figure 3.30.  G3 RH panel results for 6p STR markers

For each data vector, the 83 human-hamster hybrids are ordered in the official G3 order.

The results are coded 0 for negative, 1 for positive and R for ambiguous.

resulting positions and distances of the submitted markers amongst the framework

markers, a G3 RH map including only the submitted markers was generated. The G3

RH data was also imported and analysed within the RHMAP program as described in

Section 2.1.2.3.3. However, a reliable marker order and map length was unable to be

generated due to the large distance between the 11 markers. The centiRay distances of

the G3 framework RH map were converted into centiMorgans (Section 2.1.2.4.) for

subsequent map comparison and linkage analysis. The resulting G3 framework RH

map is shown in Table 3.25. The map was compared to the Marshfield sex average

genetic map for overall length and map position (Figure 3.31.).

### 3.4.4. Pooling strategy

Based upon a comparison of the intensity of the DNA bands on the RH mapping PCR

agarose gel photos, the cycle parameter of the PCR protocol was increased to 35 for

D6S265, D6S344, and D6S289, 40 for D6S426 and D6S461, and 45 for D6S257. With

these new PCR conditions, the 11 remaining STR markers were analysed by acrylamide

gel electrophoresis to design an optimal pooling strategy as described in Section

2.4.5.10. The final genotyping panel, including the pooling conditions and new PCR

parameters for each marker are shown in Table 3.26.

### 3.4.5. Genotyping and quality checking

The 11 STR markers were genotyped (Sections 2.1.5.4.-9.) across the Oxford linkage

cohort (Section 2.2.1.) generating approximately 10,000 genotypes (Figure 3.5.).

Marker D6S461 exhibited severe preferential amplification making accurate genotyping

impossible and was thus removed from further analysis. The remaining data was

analysed with Pedcheck as described in Section 2.1.6.1. The DNA that repeatedly

failed to amplify was consistent with those listed in Table 3.6. One round of repeats

| Stanford G3 RH Server | | |
| --- | --- | --- |
| Marker | Map Position | |
| | $cR_{10,000}$ | cM |
| D6S344 | 0 | 0 |
| D6S309 | 172 | 4.3 |
| D6S470 | 320 | 8 |
| D6S289 | 776 | 19.4 |
| D6S276 | 1072 | 26.8 |
| D6S461 | 1084 | 27.1 |
| D6S265 | 1384 | 34.6 |
| D6S291 | 1932 | 48.3 |
| D6S426 | 2184 | 54.6 |
| D6S271 | 2320 | 58 |
| D6S257 | 2860 | 71.5 |

**Table 3.25. RH and Genetic maps generated from the RH mapping of 6p STR markers**

The map position of each marker is shown relative to the initial marker.
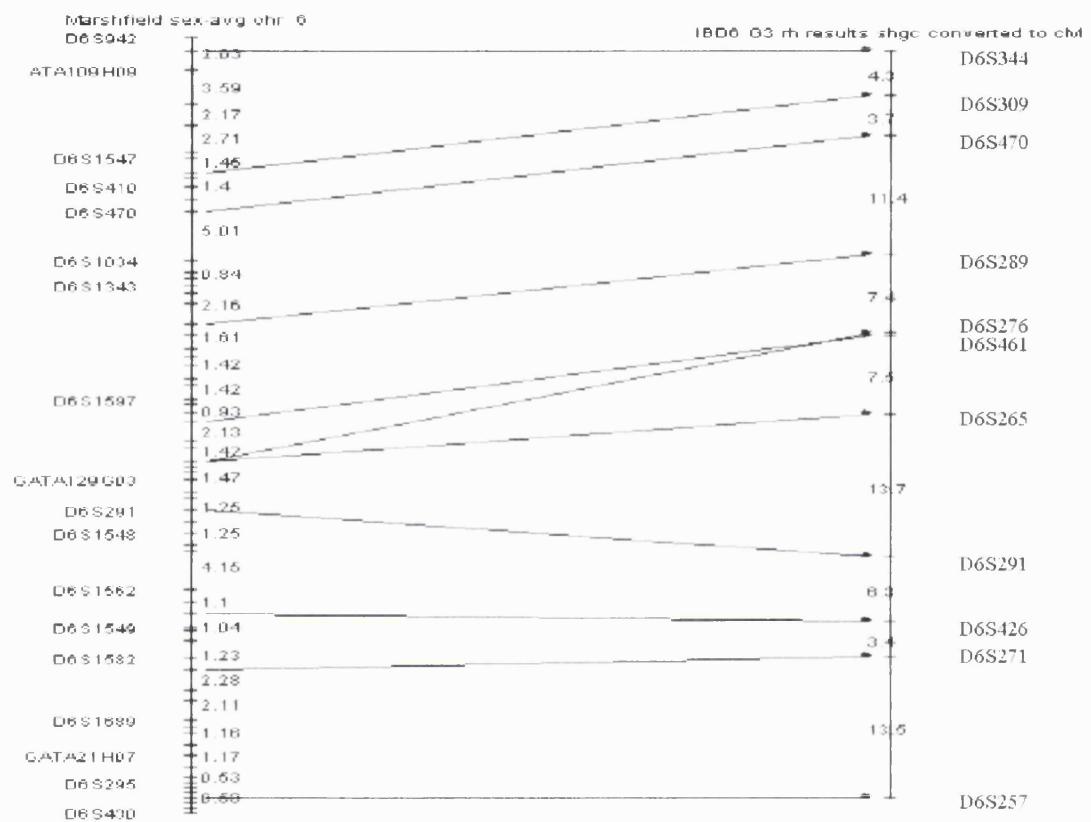
**Figure 3.31. Comparison of marker order between 6p STR marker maps**

Lines between maps connect identical markers.

| Marker | Allele Range | PCR | Pooling |
|--------|--------------|-------|---------|
| D6S470 | 102-124 | 54/30 | 2 |
| D6S291 | 156-168 | 54/30 | 3 |
| D6S276 | 241-271 | 54/30 | 3 |
| D6S461 | 299-325 | 54/35 | 3 |
| D6S289 | 342-360 | 54/35 | 3 |
| D6S265 | 193-215 | 54/35 | 3 |
| D6S426 | 241-267 | 54/40 | 3 |
| D6S257 | 307-333 | 54/45 | 6 |
| D6S344 | 172-190 | 54/35 | 3 |
| D6S271 | 243-279 | 54/30 | 3 |
| D6S309 | 294-314 | 54/30 | 3 |

**Table 3.26. Final panel for the fluorescent genotyping of 6p STR markers**

The allele range is represented in basepairs, the PCR parameters in $T_A$/cycles and the pooling ratio in microliters. The colour of the background for each marker corresponds to the fluorescent dye label (blue = FAM, yellow = HEX, green = TET) and the colour used to display the respective dye in both ABI Genescan and Genotyper software when using filter set C.

was conducted resulting in an overall success rate of 99.3% (Table 3.27.).

Recombination analysis using CRI-MAP (Section 2.1.6.2.) identified 15 potential

genotyping errors, which were subsequently removed.


### 3.4.6. Genetic map construction and comparison

The 6p STR genotype data from the Oxford linkage cohort was analysed within

MultiMap (Section 2.1.6.3.) to construct a genetic map (Table 3.28.). The map was

compared to both the Marshfield sex-average genetic map and the G3 framework RH

map (Figure 3.32.) Overall, the 6p genetic map generated from MultiMap was more

consistent in length and marker order with the G3 framework RH map and in map

position with the Marshfield sex-average genetic map.


### 3.4.7. Linkage analysis

The 6p STR genotype data from the Oxford linkage cohort was analysed for linkage to

IBD as described in Section 2.1.6.4. Positive linkage was detected in the 6p region with

a non-parametric peak multipoint LOD score of 3.04 near D6S291 for IBD. The mean

allele sharing (MAS) at the point of peak linkage was 0.58. There was almost equal

contribution from CD and UC ASPs to the linkage, as shown by the similar $\lambda_s$ values

$(0.25/z_0)$ of 1.2, 1.2, and 1.3, calculated for each of the phenotype models, IBD, CD,

and UC respectively (Table 3.29.). Graphs showing the multipoint LOD curves and

sharing statistics for all three phenotype models are shown in Figures 3.33. and 3.34.,

respectively. The average information content across the region was 0.76 (Figure

3.35.).

|                  | After first pass | After repeat 1 |
|------------------|------------------|----------------|
| PCR fails        | 393 (4.6%)       | 51 (0.6%)      |
| Inheritance fails| 21 (0.2%)        | 13 (0.2%)      |
| Total to repeat  | 414              | 64             |
| Overall success  | 95.13%           | 99.25%         |

**Table 3.27. Missing genotypes and success rates for the 6p replication study**

| 6p MultiMap Genetic Map | |
|---|---|
| Marker | Map position |
| D6S344 | 0 |
| D6S309 | 13 |
| D6S470 | 16.4 |
| D6S289 | 25.4 |
| D6S276 | 33.2 |
| D6S265 | 35.1 |
| D6S291 | 37.1 |
| D6S426 | 41.1 |
| D6S271 | 45.2 |
| D6S257 | 57.8 |

**Table 3.28. MultiMap generated 6p genetic map**

The genetic map was generated from the Oxford linkage cohort STR genotype data.

The map position of each marker is shown in centiMorgans relative to the initial

marker.
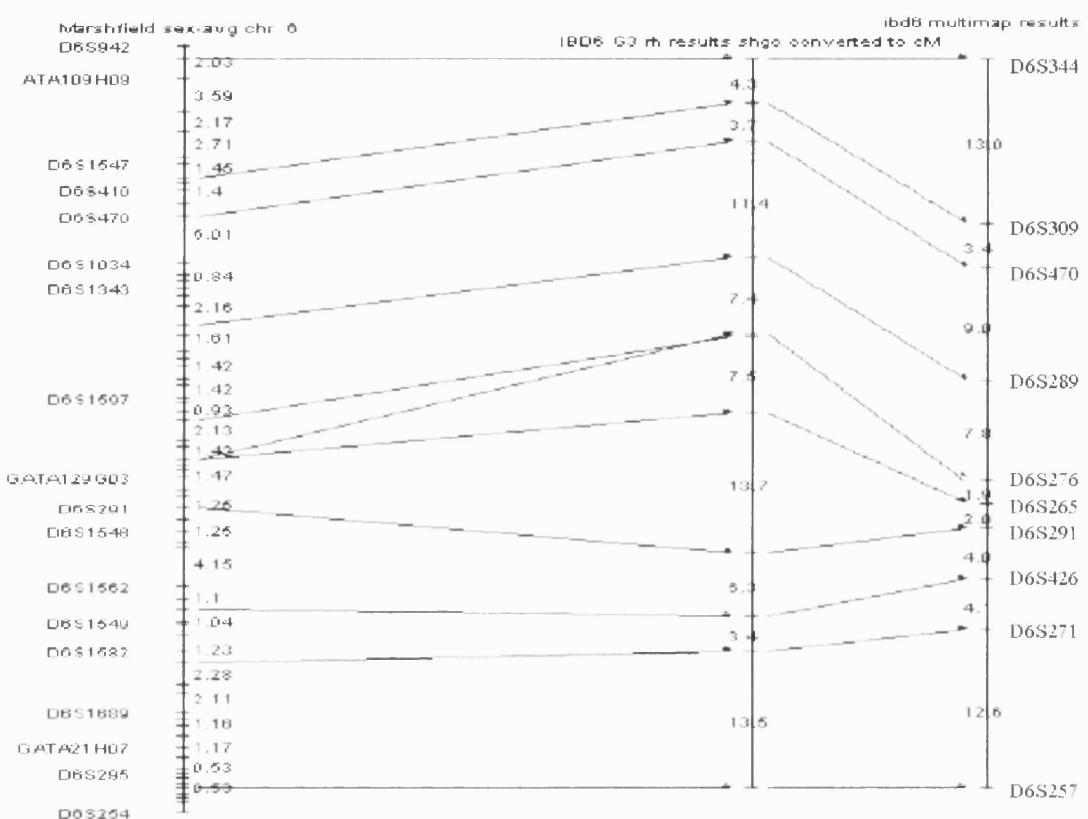
**Figure 3.32. MultiMap generated 6p genetic map comparison**

The maps within the map comparison correspond to the Marshfield sex-average genetic map shown in Table 3.23., the G3 framework RH map shown in Table 3.25., and the 3p genetic map shown in the Table 3.28.

| Phenotype Model | Peak Marker Interval | Multipoint LOD | Mean Allele Sharing | Sibling Risk |
|:---:|:---:|:---:|:---:|:---:|
| IBD | D6S291 - D6S426 | 3.04 | 0.58 | 1.19 |
| CD | D6S291 - D6S426 | 1.23 | 0.57 | 1.16 |
| UC | D6S426 - D6S271 | 0.89 | 0.58 | 1.27 |

**Table 3.29. The results of the 6p linkage study**

The results shown are from linkage analysis using the G3 RH map shown in Table 3.25. The mean allele sharing was calculated using the formula $Z_1*0.5+Z_2$. The sibling risk was calculated using the formula $0.25/Z_0$.
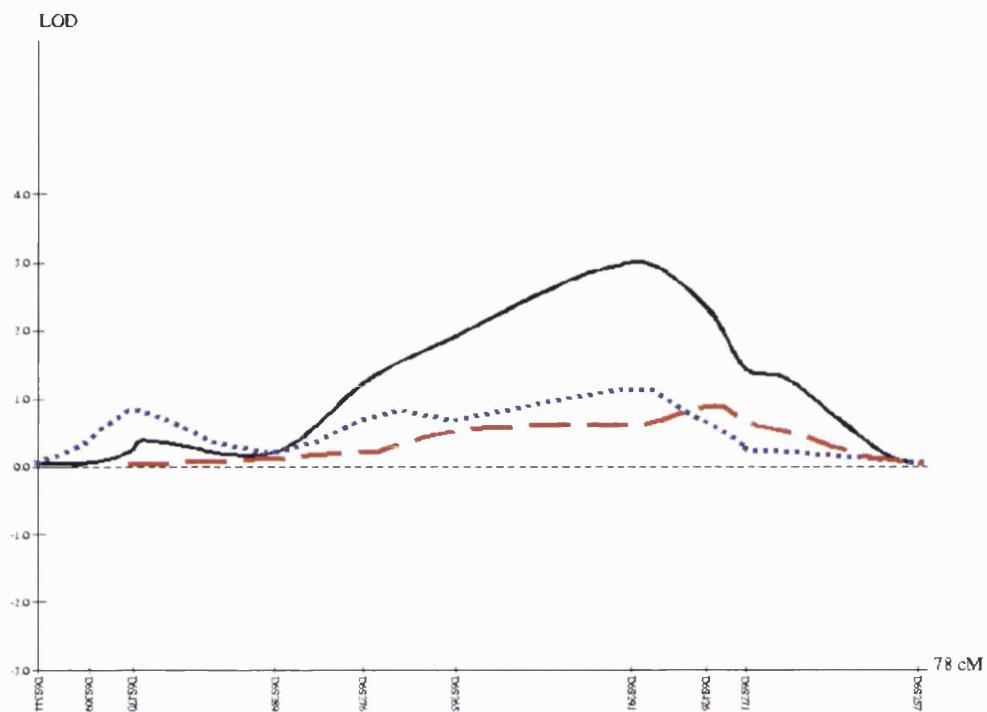
**Figure 3.33. Chromosome 6p non-parametric multipoint linkage curves**

The G3 RH map shown in Table 3.25. was used for maker order and spacing. Results

for the IBD (black, solid line), CD (blue, dotted line) and UC (red, dashed line)

phenotype models are shown.
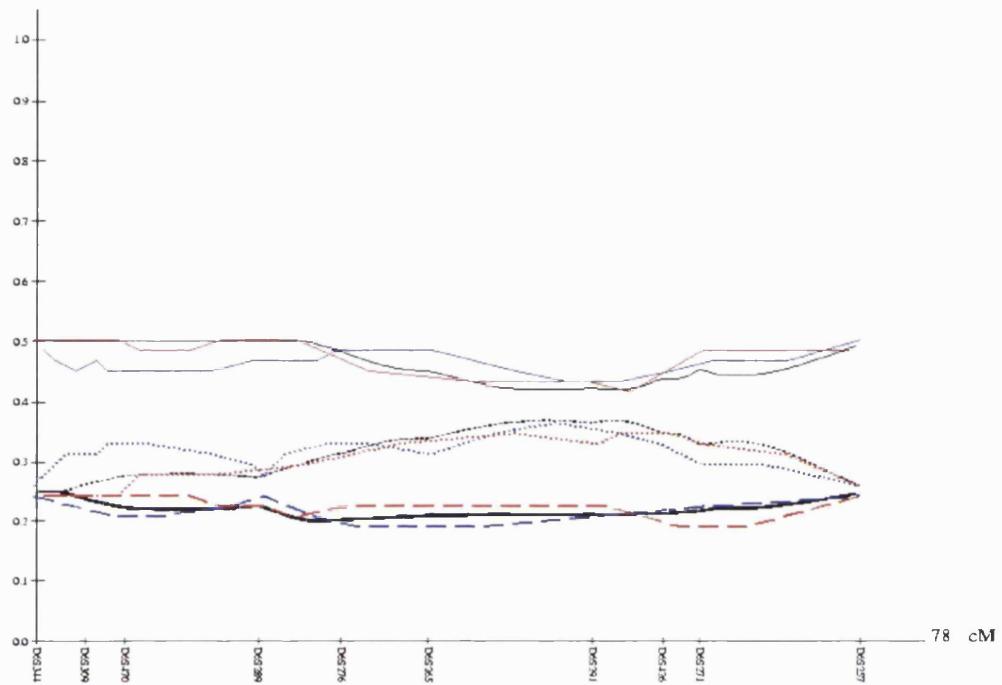
**Figure 3.34. Chromosome 6p sharing proportion curves**

The G3 RH map shown in Table 3.25. was used for maker order and spacing. Results

for the IBD (black lines), CD (blue lines) and UC (red lines) phenotype models are

shown. The sharing statistics for $Z_0$ (dashed lines), $Z_1$ (solid lines) and $Z_2$ (dotted lines)

are shown.
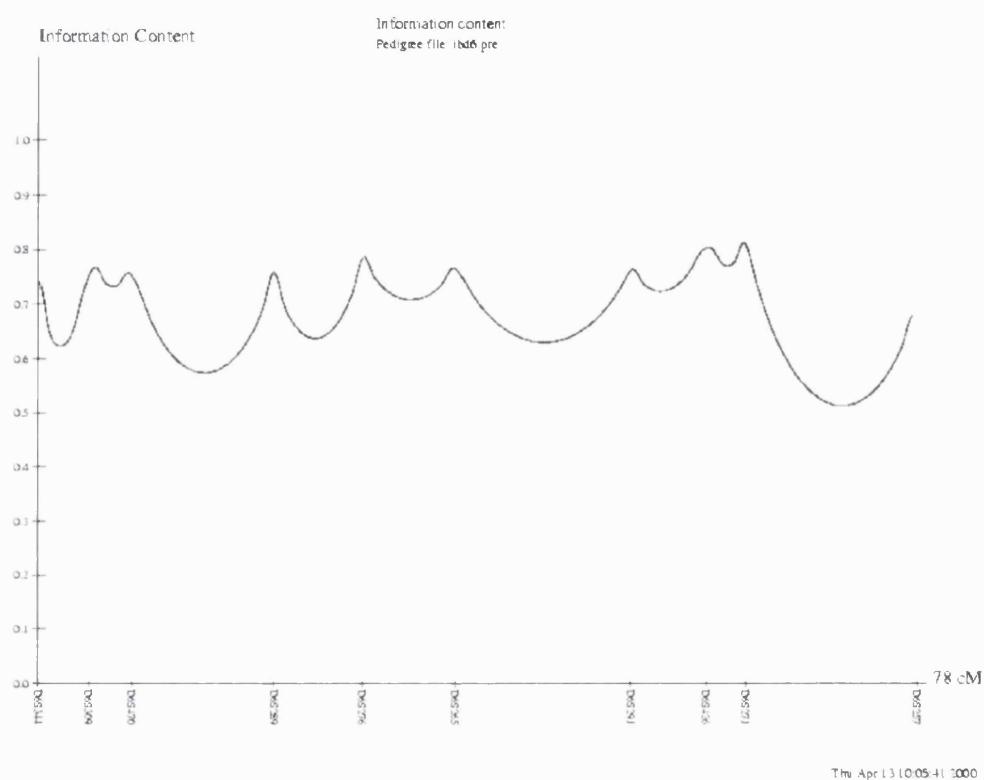
**Figure 3.35. Chromosome 6p average information content**

The G3 RH map shown in Table 3.25. was used for maker order and spacing.

# Chapter 4

# 4. Discussion

## 4.1. Linkage

### 4.1.1. Chromosome 3p extension study

#### 4.1.1.1. Failure to extend the 3p linkage region

To better stratify the numerous IBD linkage regions by their effect on the disease

population, it is important that extension studies are conducted on genome scan linkage

regions when the size of the ASP cohort increases. With the exception of the IBD

linkage regions on 14q and 19p, all other significant (LOD > 3.59) IBD linkages have

resulted from studies across suggestive loci in expanded cohorts (Duerr *et al.*, 2000;

Hampe *et al.*, 1999b; Hugot *et al.*, 1996; Rioux *et al.*, 2000; Satsangi *et al.*, 1996b). In

the present study, a suggestive IBD loci on 3p from the Satsangi et al. (1996) genome

scan was re-analysed by conducting a study in the region with an expanded population

of 284 IBD ASPs (Table 2.1.). It was hoped that, like the loci mentioned above, the

extension study would result in significant linkage. However, linkage was not detected

in the region for any phenotype model (Table 3.9.).

#### 4.1.1.2. Genetic heterogeneity and differences in methodology

Due to the difference in results from this extension study and genome scan by Satsangi

et al. (1996), the 151 families included in the previous study were analysed separately

for linkage (Table 3.10.). Since the additional families used in the present study were ascertained using the same criteria as used previously (re: clinical diagnosis, family size, ethnicity), the difference in results may be caused by both genetic heterogeneity and differences in the methodology of analysis. The analysis showed that for SIBS1&2 the observed increase in sharing within the 3p region (IBD MAS of 0.55), was more consistent with the sharing observed in the present extension study rather than the previous genome scan (IBD MAS of 0.54 and 0.63, respectively). If the difference in results between the present extension study and the previous genome scan were caused by genetic heterogeneity, then sharing at the level observed in the previous genome scan should have been detected when SIBS1&2 were analysed separately. Since SIBS1&2 data was not consistent with the previous genome scan data, the differences between the two studies are more likely to be caused by differences in methodology of linkage analysis. In the previous genome scan, two-point linkage analysis rather than mulitpoint analysis was used. In two-point analysis, because sharing status is not inferred by using data of surrounding markers, increased bias caused by the proportion of informative families is observed (Section 1.4.1.3.). If a marker by chance is more informative in families sharing the region, increased sharing and thus linkage is observed for the marker despite the fact that if all families were informative no increase in sharing would have been detected. It is this bias caused by using two-point analysis that may account for most of the difference in the results seen between the two studies.

4.1.1.3. Relative significance of the 3p putative IBD susceptibility region

Since the publication of the Satsangi et al. (1996) IBD genome scan, other independent region-specific studies in the 3p IBD linkage region have been conducted (Annese *et al.*, 1999; Brant *et al.*, 1998; Rioux *et al.*, 1998; Vermeire *et al.*, 2000). These studies also failed to see linkage to IBD. To date, 5 genome scans and 4 replication studies

have failed to detect linkage to chromosome 3p. It has been noted that suggestive

linkage will occur by chance once per genome scan and thus the linkage to 3p

previously reported by Satsangi et al. (1996) may be a false positive. However, a recent

genome scan by Rioux et al. (2000) has reported linkage to the chromosome 3 region,

with a LOD of 2.4, which suggests that within the 3p region the possibility a gene of

small effect on IBD aetiology remains.


## 4.1.2. Chromosome 6p replication study


4.1.2.1. Replication of the *IBD3* susceptibility locus

IBD genome scans have resulted in the identification of several regions of linkage.

However, only two regions have been subsequently replicated by independent region-

specific studies. Recently, Hampe et al. (1999b) reported linkage to 6p in an extension

study using the largest IBD population to date (428 ASPs). In the present study, the 6p

linkage region was replicated in the Oxford IBD cohort containing 284 ASPs (Table

2.1.). The nonparametric peak multipoint LOD score of 3.04 near D6S291 (Table 3.29.)

exceeds the proposed criteria for independent replication of significant linkage (LOD =

1.5) (Lander and Kruglyak, 1995). However, while this represents the first region-

specific replication study of the *IBD3* locus, both a recent genome scan by Rioux et al.

(2000) and an earlier candidate gene study by Yang et al. (1999) also reported

suggestive linkage to 6p between D6S1281 and D6S1019 in a cohort of 183 IBD ASPs

(LOD = 2.3) and around TNF in a cohort of 70 CD ASPs (P = 0.002), respectively. All

three IBD studies showed linkage contributed by both the UC and the CD populations.

However, both the linkage peak from the Rioux et al. (2000) study and the peak from

this study were approximately 15 cM proximal to the initial reports of linkage (Hampe

*et al.*, 1999b; Yang *et al.*, 1999b). This discrepancy is not surprising since differences

between actual gene location and peak linkage are not unexpected in complex traits (Kruglyak and Lander, 1996; Roberts *et al.*, 1999). Taken together the results support the existence of an IBD susceptibility locus at 6p (*IBD3*).

### 4.1.2.2. Relative significance of *IBD3*

*IBD3* is the only locus where positive linkage has been reported in each of the three largest independent IBD studies (Table 1.1.). Both *IBD1* and *IBD2* have been explored in independent linkage studies with greater power than the original studies reporting linkage to the regions. However, these later studies of greater power have all failed to show evidence of linkage to the regions (Cho *et al.*, 1998; Duerr *et al.*, 2000; Hampe *et al.*, 1999a; Lesage *et al.*, 2000; Rioux *et al.*, 2000). One study even excluded the possibility of a gene in the *IBD1* and *IBD2* intervals with a $\lambda_s$ of 2.0 or greater (Rioux *et al.*, 1998). The differences between these studies may be because of genetic heterogeneity or the difficulty in detecting genes of modest effect in complex traits (Risch and Merikangas, 1996). In the first instance, since linkage cohorts are frequently ascertained differently and consist of different ethnicity, a linkage region detected in one study may contain a gene that only affects people with the same ethnicity and severity of disease and would thus be undetected in a population of different composition. In the second instance, if a cohort analysed for linkage were biased in favour of families containing a particular disease gene, increased linkage would be detected compared to a non-biased population. If the disease gene in question also had a low displacement (t) in the overall population, than the actual increase in sharing (Y or MAS) caused by the gene could be insufficient to detect linkage above a LOD of 2.19 (Risch, 2000). Therefore, as a result of bias in the initial study, subsequent studies containing larger populations with less bias could fail to detect linkage in the same region. Nonetheless,

| Gene | Description |
| --- | --- |
| D6S276 | |
| ZNF204 | zinc finger protein 204 |
| D6S265 | |
| H1F5 | H1 histone family, member 5 |
| PRSS# | thymus specific serine peptidase |
| tob family | tob family |
| zf47 | zinc finger protien |
| COX11 | cytochrome C oxidase subunit 11 |
| H2BFR | H2B histone family, member R |
| SRE-ZBP | zinc finger protein (SRE-ZBP) |
| H2AFP | H2A histone family, member P |
| H3FK | H3 histone family, member K |
| HLA-A | major histocompatibility complex, class I, A |
| ZNF173 | zinc finger protein 173 |
| HLA-F | gene for human leukocyte antigen F |
| KIAA0036 | KIAA0036 gene product |
| HCGV | hemochromatosis candidate gene V |
| CAK | cell adhesion kinase |
| hcgVIII | hcgVIII |
| PPP1R10 | protein phosphatase 1, regulatory subunit 10 |
| ZNF165 | zinc finger protein 165 |
| P5-1 | P5-1 |
| ZNF184 | kruppel-related zinc finger protein |
| CDSN | corneodesmosin |
| KIAA0170 | KIAA0170 gene product |
| ABC50 | ATP-binding cassette 50 (TNF-alpha stimulating |
| DIF2 | Differentiation-dependent gene 2 |
| DDX16 | DEAD/H (Asp-Glu-Ala-Asp/His) box polpeptide 16 |
| HLA-Bw72 antigen | HLA-Bw72 antigen |
| HLA-C | Major histocompatibility complex, class I, C |
| MICA | MHC class I polypeptide-related sequence A |
| HLA-E | Major histocompatibility complex, class I, E |
| GNL1 | guanine nucleotide binding protein-like 1 |
| MOG | myelin oligodendrocyte glycoprotein |
| FLOT1 | flotilin 1 |
| SC1 | putative trans-acting factor involved in cell cycle control |
| RXR-beta | retinoid X receptor beta |
| HLA-G | HLA-G histocompatibility antigen, class I, G |
| RPS19 | ribosomal protein S19 |
| zinc finger protein | zinc finger protein |
| MICB | MHC class I polypeptide-related sequence B |
| HLA-J | Major histocompatibility complex, class I, J |
| C2 | complement component 2 |
| beta-tubulin | clone 24703 beta-tubulin |
| ARA24 | androgen receptor associated protein 24 |
| H2BFN | H2B histone family, member N |
| hsp 70 | heat shock protein |
| HLA-DRB1 | Major histocompatibility complex, class II, DR beta 1 |
| NOTCH4 | Notch (drosophila) homolog 4 |
| HLA-DR alpha chain | HLA-DR alpha chain |
| TAPBP | TAP binding protein (tapasin) |
| TAP1 | transporter 1, ABC (ATP-binding cassette) |
| HLA-DQB1 | major histocompatibility complex, class II, DQ beta 1 |
| BING4 | BING4 |
| Lst-1 gene | Lst-1 gene |
| HSPA1L | heat shock 70kD protein-like 1 |
| AIF1 | alograft inflammatory factor 1 |
| B3GALT3 | UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 3 |
| RING1 | ring finger protein 1 |
| HLA-DMB | major histocompatibility complex, class II, DM beta |
| lymphocyte antigen | lymphocyte antigen |
| C4A | complement component 4A |
| 1C7 precursor | 1C7 precursor |
| LTA | lymphtoxin alpha (TNF superfamily, member 1) |
| CLIC1 | chloride intracellular channel 1 |
| TNF | tumor necrosis factor (TNF superfamily, member 2) |
| G9 | G9 gene encoding sialidase |
| STK19 | serine/threonine kinase 19 |
| D6S51E | HLA-B associated transcipt-2 |
| D6S52E | HLA-B associated transcript-3 |
| HLA-DPB1 | major histocompatibility complex, class II, DP beta 1 |
| NFKBIL1 | nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor-like 1 |
| PPT2 | palmitoyl-protein thioesterase 2 |
| ITPR3 | inositol 1,4,5-triphosphate receptor, type 3 |
| AGER | advanced glycosylation end product-specific |
| HLA-DOB | major histocompatibility complex, class II, DO beta |
| COL11A2 | collagen, type XI, alpha 2 |
| SACM2L | suppressor of actin mutations 2, yeast, homolog-like |
| PHF1 | PHD finger protein |
| CREBL1 | cAMP responsive element binding protein-like 1 |
| PSMB8 | proteasome subunit, beta type 8 |
| HLA-DMA | major histocompatibility complex, class II, DM alpha |
| BF | B-factor, properdin |
| RD | Radin blood group |
| RNF3 or RING3 | female sterile homeotic-related gene 1 or ring finger protein 3 |
| RPS10 | ribosomal protein S10 |
| TAF2I | TATA box binding protein (TBP)-associated factor, RNA polymerase II, I, 28kD |
| SRPK1 | SFRS protein kinase 1 |
| FKBP5 | FK506-binding protein 5 |
| PPARD | peroxisome proliferator activated receptor, delta |
| KNSL2 | Kinesin-like 2 |
| CLPS | colipase, pancreatic |
| HLA-DNA | major histocompatibility complex, class II, DN alpha |
| D6S291 | |
| CSBP1 | cytokine suppressive anti-inflammatory binding protein 1 (p38 MAP kinase) |
| KIAA0229 | KIAA0229 gene product |
| GLP1R | glucagon-like peptide 1 receptor |
| KIAA0082 | KIAA0082 gene product |
| Ndr protein kinase | Ndr protein kinase |
| COX6A1 | cytochrome c oxidase subunit VIa polpeptide 1 |
| KIAA0646 | C3HC4-type zinc finger protein |
| CFLAR | CASP8 and FADD-like apoptosis regulator |
| ORP150 | 150 kDa oxygen-regulated protein |
| CDKN1A | cyclin-dependent kinase inhibitor 1A (p21, Cip1) |
| SFRS3 | splicing factor arginine/serine-rich 3 |
| pBS13 | pBS13 precursor polypeptide, testis-specific |
| ZNF76 | zinc finger protein 76 (expressed in testis) |
| PIM1 | pim-1 oncogene |
| SNRPC | small nuclear ribonucleoprotein polypeptide C |
| apolipoprotein M | apolipoprotein M |
| TAP2 | transporter 2, ABC (ATP binding cassette) |
| CYP21 | cytochrome P450, subfamily XXI |
| TSBP | tetis specific basic protein |
| GSN | gelsolin (amyloidosis, Finnish type) |
| MOCS1 | molybdenum cofactor synthesis 1 |
| KIAA0349 | KIAA0349 gene product |
| KIAA0240 | KIAA0240 gene product |
| PPP2R5D | protein phosphatase 2, regulatory subunit B (B56), delta isoform |
| KIAA0708 | KIAA0708 gene product |
| ATP6C | ATPase, H+ transporting, lysosomal (vacuolar proton pump) 16kD |
| KIAA0076 | KIAA0076 gene product |
| BYSL | bystin-like |
| GUCA1A | guanylate cyclase activator 1A (retina) |
| TBCC | tubulin-specific chaperone c |
| D6S426 | |
| RDS | retinal degeneration, slow (retinitis pigmentosa 7) |
| SRF | serum response factor (c-fos serum response element-binding transcription factor) |
| RCL | RCL |
| NFYA | nuclear transcription factor Y, alpha |
| PGC | progastricsin (pepsinogen C) |
| RPL24 | ribosomal protein L24 |
| VEGF | vascular endothelial growth factor |
| KIAA0381 | KIAA0381 gene product |
| MIDF1 | MyoD family inhibitor |
| CCND3 | cyclin D3 |
| pre TCR alpha | pre TCR alpha |
| Pxaaa 1p | Pxaaa 1p |
| CTG4a | CTG4a |
| PTK7 | protein tyrosine kinase 7 |
| D6S271 | |

**Table 4.1.  Known genes mapping into the 1 LOD support interval around D6S291**

The genes were retrieved from the Genebridge 4 Genemap99 RH map.  The 6p STR markers shown in blue were used as framework markers to map the genes between.  The peak linkage marker (D6S291) is indicated in red.

whatever the cause of the discrepancies between IBD studies for both *IBD1* and *IBD2*, the consistency of positive results to IBD3 between large IBD linkage studies suggests that the *IBD3* locus may make the largest contribution to IBD susceptibility overall.

### 4.1.2.3. Candidate genes in the *IBD3* region

The confirmation of the IBD linkage to 6p now warrants high-density association studies across the region, starting with positional candidate genes. The *IBD3* locus contains the major histocompatibility complex (MHC) region that has been implicated in other linkage studies of autoimmune diseases, including inflammatory diseases such as asthma, arthritis and psoriasis (Becker *et al.*, 1998; Jawaheer *et al.*, 2001). This overlap of susceptibility regions suggests that a MHC autoimmune susceptibility gene may be responsible for the positive linkage results. As previously noted in Sections 1.5.2.2.1. and 1.5.2.2.2., DRB1, TNFα, and other MHC genes have already shown positive association with IBD. However, another hypothesis is that the overlap of disease susceptibility loci could indicate an inflammation specific disease gene in the region. One potential candidate gene involved in inflammation, as suggested by it name, is cytokine suppressive anti-inflammatory binding protein 1 (CSBP1). CSBP1 RH maps directly adjacent to D6S291, the peak linkage marker in this study (Table 4.2). Still, even with multiple overlapping disease linkage regions, the existence of a potential IBD specific disease gene in the region should not be ignored. Recently, peroxisome proliferator activated receptor gamma (PPARγ) has been implicated in IBD (Su *et al.*, 1999), which suggests that a family member like PPARδ that again maps adjacent to D6S291 could also be implicated (Table 4.1). Since the region contains several IBD candidate genes and could also harbour more than one IBD susceptibility gene, susceptibility to IBD conferred by the *IBD3* locus may not be completely understood

until all autoimmune, inflammation and IBD specific candidate genes within the region have been thoroughly tested for IBD association.

### 4.1.3. Another genome scan?

Both linkage studies reported here were conducted with an extended linkage cohort that included the families used for the Satsangi et al. (1996) genome scan (Section 2.2.1.). However, because of the increase in power using the extended cohort, a positive linkage region originally reported from the genome scan was not detected (3p) while a new positive linkage region was detected (6p). These differences, seen between the first two regions analysed and the original genome scan, suggest that another genome scan using the extended cohort may be warranted. Due to the increase in power, significant linkage regions, like *IBD3*, which result from this second genome scan may be a better indicator as to which putative IBD susceptibility regions contain the genes affecting the greatest proportion of European Caucasian IBD patients.

### 4.1.4. Saturation genotyping

Previously, it was standard procedure to conduct a saturation linkage study with as many STR markers as available from the public databases after an initial linkage region was identified from a genome scan. Recently however the benefits of this type of study have been questioned. Saturation studies were supposed to provide two benefits. Firstly, the increase in marker density should increase the information content across the region and therefore the linkage analysis programs would make fewer assumptions about sharing and sites of recombination, providing a more resolved and accurate representation of the linkage in the region. Secondly, association analysis could be conducted on the dense set of STR markers, further refining the disease gene interval.

184

Although these benefits sound attractive, the actual gain in information compared to the amount of work is small.

Due to the lack of high-resolution genetic maps RH maps or physical yeast artificial chromosome (YAC) or bacterial artificial chromosome (BAC) maps have to be used to order makers. Although both provide a higher resolution than genetic maps markers are frequently positioned in the wrong order because of remaining insufficient resolution, chimeras and gaps. Incorrect marker order causes increased false recombination across the region, which decreases the appearance of sharing and therefore the linkage in the region. Even so, most false recombination should be removed by analysis with statistical packages like CRI-MAP (Lander and Green, 1987). Yet, if the extra recombination caused by incorrect marker order is removed by removing the marker genotypes, then a portion of the gain of increased marker density is also removed.

Another reason for the diminishing returns from saturation studies is that because of the high heterozygosities of STR markers and success in genotyping, the actual increase in information content across the region after saturation is small. In this study, when the information content was compared in the same 35cM region using 14, 18 and 41 markers, less than a 5% difference was noted in average information content (Figures 3.9., 3.14., and 3.16.) and less than a LOD difference of 0.25 in linkage results (Table 3.9.and 3.11.), increasing with marker density. Similarly, in another study, when the marker density was increased in the region of interest from 10cm to 2cM again only a small increase in LOD score detected (Wjst *et al.*, 2000).

In spite of the lack of increased linkage information, people continue to conduct saturation studies to look for association across regions. However, there are again

several reasons why this type of association study will result in insignificant data. Firstly, as seen in this study (Sections 3.1.8. and 3.2.2.), results of greater significance are frequently achieved in simulations with random equivalent data sets compared to the actual data sets themselves because of the number of alleles in STR markers and the number of markers tested in saturation studies. In other words, after correcting for multiple testing, most significant results would be discounted. Secondly, STR markers have a relatively high mutation rate ($1 \times 10^{-3}$) and contract as well as expand (Xu *et al.*, 2000). Therefore, in a common disease like IBD where the disease mutations are hypothesised to have originated prior to the expansion of the human population, LD between the mutation and a neighbouring STR marker will not only decrease over time due to recombination but will also decrease because of recurring mutation. This added decrease in LD makes it again less likely to detect association to a common disease mutation using STR markers, especially since the direction of mutation is unpredictable. Lastly, STR markers are not found in sufficient density for association analysis in common diseases. Recent estimations of LD in common diseases, suggest that markers spaced between 6kb and 100kb across a region would be required to detect association (Collins *et al.*, 1999; Kruglyak, 1999). Thus, even an extensive saturation study, like the present 3p saturation study where one STR marker was genotyped every megabase, would be at best one tenth the density required for accurate testing of association.

### 4.1.5. Choice of marker map for linkage analysis

Typically, genetic maps are used for linkage analysis so that results depict the actual sharing across a chromosome as accurately as possible. Recently, a comparison of the physical map resulting from the sequencing of chromosome 22 and the corresponding Genethon sex-average genetic map showed that inter-marker distance varied between the maps, indicating that recombination 'jungles' and 'deserts' existed across the

chromosome (Dunham *et al.*, 1999). This existence of regional differences in recombination rate invalidated the assumption of a constant rate of recombination used in both genetic map construction and linkage analysis. Ironically though, because both linkage analysis and genetic map construction employ this same assumption about recombination, when a genetic map is used for linkage analysis, regional differences in recombination rates are corrected for and the depiction of linkage is accurate. However, when using RH and physical maps for linkage analysis, the false assumption of a constant rate of recombination is not corrected for and therefore linkage results are less accurate.

If, for accuracy, it is decided to use a genetic map for analysis, the next question is whether to use a publicly available genetic map or construct one based upon the data being analysed. It has been suggested that bias may be generated if the map used for analysis is constructed with the same subjects as those being analysed for linkage. Decreases in recombination in areas of linkage may be expected when subjects are selected to potentially share a region of DNA harbouring a disease gene. The decrease in recombination would make the region appear smaller than in other genetic maps and therefore decrease the amount of assumed peak sharing and the resulting LOD score. Because of this possible bias, most studies have tended to use the publicly available genetic maps. However, these maps also pose some potential problems. All markers chosen for a study may not exist on one map, especially when trying to use markers reported in a previous study. Also, because the public genetic maps were constructed using only a few families, for example Marshfield used eight, the inter-marker distances may not be as accurate as a map constructed with several pedigrees of identical ethnicity to the population to be analysed for linkage. In this study, to analyse whether the choice of genetic map has a significant impact on linkage results, the 3p linkage data was

analysed using both the Marshfield sex-average genetic map and a genetic map

constructed in MultiMap. No significant differences were seen between the two maps

in marker order, map position and overall length (Figure 3.6.) and thus, not surprisingly,

only a difference of about 0.1 was observed between the resulting LOD scores (Table

3.11.). These results suggest that the decision of which genetic map to use may not be

critical.


In this project, RH maps were used for linkage analysis instead of genetic maps. This

may create inaccuracies due to the assumption of a constant recombination rate.

However, no public genetic or physical maps were available that contained all the

markers used in the study in both a specified order and with an individual map position,

so the use of an RH map was required. To analyse the effect on linkage results of using

this type of marker map, a comparison between the linkage results generated using the

Marshifield sex-average map and the G3 RHMAP RH map was made (Table 3.9. and

3.11.). Again, only a difference of about 0.1 between LOD scores was observed.

However, it should be noted that the increase in information content with increase in

marker density was not taken into account in this comparison (Section 4.1.4.) and

therefore any correlation made by this comparison may be inaccurate. Nonetheless,

combined with the results of the previous comparison, it suggests that overall the choice

of map for linkage analysis may not have a large impact on results.


## 4.1.6. Impact of allele frequency source

When conducting linkage analysis, marker allele frequencies are used to calculate the

probability of each possible genotype for any subjects required for analysis whose

genotypes are not available. For example, if two siblings both had the same allele from

their mother but the mother's genotype was not available, then it would not be known

188

whether the mother was homozygous or heterozygous for the allele and thus it would be

impossible to tell whether the siblings were sharing the same maternal chromosome.

However, if it were known that the allele that the siblings shared from their mother was

rare then it would be more likely that the mother was heterozygous for the allele and

thus more likely that the siblings were sharing the maternal chromosome. Due to the

inferences made from allele frequencies during linkage analysis, differences in the allele

frequencies used for analysis can effect the results (Thiel *et al.*, 2000). In the study

reported here, the impact that altered allele frequencies had on linkage results was

tested. Linkage analysis was conducted using both allele frequencies generated from

the population analysed and uniform allele frequencies where all alleles were

considered to have the same prevalence. The LOD scores generated when using the

uniform allele frequencies were significantly higher than those resulting from analysis

using population allele frequencies (Table 3.9. and 3.11.), especially in CD where the

LOD score increased from 1.25 to 1.73. Thus, if allele frequencies have this large an

effect on linkage results where 44% of the parental genotypes are missing (Section

2.2.1.), then the impact on linkage studies of late onset disease like osteoporosis and

Alzheimer's disease, where the majority of parental genotypes are missing, could be

dramatic.


Due to the impact of allele frequencies on linkage results, it was important in this study

to be confident in the accuracy of the allele frequencies used. Therefore, the allele

frequencies of the 3p STR markers were compared between those observed in the

Oxford linkage and Oxford TDT cohorts. No significant difference was seen between

the allele frequencies even though the family number, size and composition were

different between the two populations (Section 2.2.1. and 2.2.2.). Therefore, it was

believed that the frequencies used for the 3p linkage study reflected those found in the

European Caucasian population. Also, because there was no difference between the genotyping methods used between the 3p and 6p linkage studies, it was inferred that the allele frequencies used for linkage analysis in the 6p replication study also reflected those found in the European Caucasian population.

## 4.2. Association

### 4.2.1. STR IBD association study on 3p

At the start of 1998, it was standard procedure to try and narrow linkage regions by looking for association using a dense set of STR markers. Genetic maps containing most of the human polymorphic STR markers with linked heterozygosities and allele frequencies were well established. However there were few SNP markers that were publicly available and those SNPs that were available were not mapped. Concurrently, because of the increased interest in the use of linkage studies for gene discovery, high-throughput STR genotyping methods were developed and improving rapidly. At the same time, no equivalent high-throughput SNP genotyping system existed. This made it relatively easy to conduct association analysis with STR markers but almost impossible and extremely time consuming to conduct the same analysis using SNP markers. Therefore, when the present study commenced, the first goal was to refine the 3p linkage region from the Satsangi et al. genome scan (1996) by conducting association analysis with a dense set of STR markers. A total of 42 STR markers were genotyped across a TDT population and an ASP population, the latter including the families used in the Satsangi et al. genome scan (1996). Significant association (p < 0.01) was detected with several alleles from markers in both populations and all phenotype models (Table 3.12. and 3.15.). However, simulations of 1000 random equivalent data sets resulted in several data sets of greater significance (Section 3.1.8.

and 3.2.2.), suggesting that if the results were corrected for multiple testing, no significant associations to IBD phenotypes would remain.

Currently, with the knowledge gained from recent studies and simulations, these negative STR association results are not surprising. As mentioned in Section 4.1.3., the loss of power due to the number of alleles in STR markers, the loss of significance due to multiple testing, the decrease in marker-mutation LD due to recurring mutation of STR markers, and the inadequate number of STR markers compared to the recent estimations of marker density required to detect association to common disease genes, make it unlikely that true association to a common disease variant would be detected using STR markers.

Even so, whilst recent data suggest that the insignificant STR association results are not surprising, correlation made between the results of the analysis of the two independent populations should not be ignored. By comparing the results from the two cohorts shown in Table 3.12 and 3.15, it was noticed that allele 1 from D3S1573 and allele 5 from D3S1277 showed positive non-transmission association to CD in both populations. The two markers are approximately 15cM apart which suggests that if real association between a disease gene and a marker was being detected, either each marker was detecting a different mutation of a different gene or only one of the replicated associations was real. In the latter instance, the replicated association seen with D3S1573 would have more credence. D3S1573 was the 3p marker that showed the largest two-point LOD score to IBD in the Satsangi et al. genome scan. While in the present 3p linkage study, although linkage greater than a LOD of 2.2 was not detected, D3S1573 was within the region that showed linkage to IBD greater than a LOD of 1 (Figure 3.7.). Conversely, at D3S1277 almost no increase in sharing was detected.

Whether, neither, one or both of the replicated associations are real, further analysis with larger populations and higher density biallelic markers will be required to verify putative associations detected in the region.

## 4.2.2. GNAI2 candidate gene association study on 3p

4.2.2.1. No association between GNAI2 and any IBD phenotype

GNAI2 is the α-subunit and the guanine nucleotide-binding peptide of a G protein. G proteins are signal transducers that couple receptors for extracellular signals to effectors like ion channels (Spiegel, 1995). Recently, mice deficient in GNAI2 were shown to present a UC-like phenotype (Rudolph *et al.*, 1995). Following this discovery, Satsangi et al. (1996) published a genome scan detecting linkage between IBD and a marker adjacent to GNAI2, D3S1573. In the present study, to investigate the interaction of GNAI2 and IBD further, polymorphisms spanning GNAI2 were tested for association in a large IBD cohort. In the initial study, significant positive association was seen between a haplotype of polymorphisms spanning GNAI2 and CD rather than UC (Figure 3.26.). However, in a subsequent extension study when the initial CD cohort was expanded 3-fold, no association was detected.

4.2.2.2. Are the GNAI2 discordant results due to genetic heterogeneity or a false positive?

To date, no other association studies have been published between IBD and GNAI2. This is surprising taking into account its location and potential functional significance to UC (Section 4.2.2.1.). In the present study, 9 polymorphisms were identified in approximately 16kb of genomic sequence spanning GNAI2 (Table 3.20.) yet only two had rare allele frequencies greater than 10%. Potentially, it was this difficulty in finding

common polymorphisms to genotype that deterred scientists from analysing GNAI2 for association.

If the present study had ended after the initial detection of positive association the results, like many other positive IBD associations (Section 1.5.2.2.), would have been published and most likely tested in many other independent IBD cohorts with varying outcomes. However, with the present analysis of the expanded TDT cohort failing to detect positive association, time was saved in reaching the state of confusion caused by discordant results. It is at this stage of discordance where words like genetic heterogeneity and ascertainment bias start to be bantered about. In the present study, the later can be ruled out because the patients included were all ascertained using the same criteria (re: clinical diagnosis, family size, and ethnicity). Thus, genetic heterogeneity is the only possible cause of the discrepancy. However, in a common disease most likely affected by common polymorphisms that arose prior to the expansion of the human population, why should it be expected that any population or subset of the population would have different polymorphisms causing disease susceptibility? Recent simulations of common disease in human populations agree that this would not be the case (Kruglyak, 1999). Instead, but rarely suggested, the inconsistencies may actually be due to false positives arising from multiple testing, including the number of studies investigating the same hypothesis. Even though simulations of equivalent data sets suggest that the positive association detected here between Gnai2 and CD would remain positive after correcting for multiple testing, it would seem that the association is more likely a false positive than an effect of genetic heterogeneity because of the inconsistency seen between the association (CD) and the mouse knockout (UC) in addition to the subsequent lack of association detected in the extension study.

### 4.2.3. CCR5, CCR2, and CCRL2 candidate gene association study on 3p

#### 4.2.3.1. Positive association between CCR5 and UC

CCRs are receptors for chemokines, which are proinflammatory cytokines. CCRs have

a direct involvement in the inflammation pathway and thus it has be hypothesised that

mutations within them could cause chronic inflammation, such as is seen in IBD

(MacDermott *et al.*, 1998; Papadakis and Targan, 2000). Based upon this hypothesis,

CCR2, CCR5 and CCRL2 were analysed for association to IBD. The three genes were

contained in a single, fully sequenced BAC that mapped near the peak region of linkage

on chromosome 3 reported in the Satsangi et al. (1996) genome scan. Two markers,

CCRF10-697 near CCR2 and CCR5Δ32 in CCR5, resulted in positive association to UC

(Figure 3.27.). However, the results from 1000 simulations of equivalent data sets

suggest that only the CCR5 association would remain significant after correcting for

multiple testing (Section 3.3.8.).

#### 4.2.3.2. Relative significance of the CCR5 association

To date, no other study has reported positive association between CCR5Δ32 and an IBD

phenotype. However, this is not the first study to look for association with CCR5Δ32.

Three previous smaller studies tested IBD patients for association to CCR5Δ32 but no

association was detected (Hampe *et al.*, 2001; Martin *et al.*, 2001; Rector *et al.*, 2001).

This discordant result could suggest that the association detected in this study was either

a false positive or that the previous studies lacked sufficient power to detect the

association. Evidence which suggests that the former may be true is that CD shows the

greatest linkage to the region and UC almost shows no linkage at all (Table 3.9.). Still,

it is known that association has a greater power than linkage to detect gene effects and

that gene effects conferring an increase in sharing less than 0.54 would only be detectable by association (Risch, 2000). Therefore, it is conceivable that true association could be detected in the absence of linkage. Until CCR5Δ32 is analysed for association in a UC cohort of greater power the question of its significance in IBD will remain unsolved.

### 4.2.3.3. Is CCR5Δ32 a functional UC mutation?

It is interesting that CCR5Δ32 was the only polymorphism tested for association that showed significant results. CCR5Δ32 was one of two polymorphisms tested known to be a protein altering polymorphism, with CCRL2-524, Tyr to Phe, being the other. Yet, only CCR5Δ32 was known to be functional. It is well established the homozygotes for CCR5Δ32 are resistant to HIV-1 infection (Martinson *et al.*, 1997; Mummidi *et al.*, 1997). With a mutation known to affect viral entry, it can be hypothesised that a different environmental element could cause a different phenotype, like UC, via the same mutant pathway. Results from the present study suggest that if the association to UC is true, then CCR5Δ32 is more likely to be the disease causing mutation rather then a polymorphism associated by LD with the disease causing mutation. Analysis of haplotypes including CCR5Δ32 and surrounding polymorphisms showed that the association was strongest at CCR5Δ32 and decreased with the addition of neighbouring markers even though strong LD existed across the region (Figure 3.27.). Combining the present results with the information previously know about CCR5Δ32, studies evaluating the existing CCR5Δ32 cell lines and knockout mice and their response to known IBD associated pathogens and other environmental stimuli are warranted. One such study, looking at the response of CCR5 knockout mice when chemically induced to present colitis, has recently been published (Andres *et al.*, 2000). Discordant to the present association results and similar to HIV, the study found that mice deficient of

CCR5 showed resistance to colitis. In the present study, CCR5Δ32 showed increased

transmission to UC patients. If it is true that CCR5Δ32 is protective against

susceptibility to colitis then positive association with decreased transmission of

CCR5Δ32 to UC patients is expected. More studies are needed to bring clarity to the

situation. However, with the first functional study between CCR5Δ32 and colitis

resulting in a resistant phenotype, consistent with other known CCR5Δ32-disease

interactions, first indications suggest that the present association results may be a false

positive.

## 4.2.4. Concurrent developments in association analysis for common complex disease

### 4.2.4.1. Extent of LD

Recent estimations of the extent to which detectable LD extends away from a common

disease susceptibility locus vary. A population simulation of a common disease variant

arising prior to the expansion of the human population suggested that detectable LD

would extend only 3kb from the variant (Kruglyak, 1999). On the other hand, a genetic

epidemiological study comparing the results from several common disease association

studies suggested that detectable LD extends beyond 100kb from common disease loci,

possibly because of subsequent bottlenecks following population expansions (Collins *et al.*, 1999). In the present candidate gene association studies, the denser estimation of

required marker density was used to thoroughly test each gene for association to IBD.

In each study, inter-marker LD extended across the entire region (Figures 3.26. – 3.27.),

with the largest region being approximately 60kb. These results support the later theory

of a larger extent of LD. Similarly, other recent studies have also observed larger than

expected LD intervals. One study examined inter-marker LD with 11 biallelic markers

in a European American population (McMahon *et al.*, 2000). The extent of LD varied, with the largest region of detectable LD being approximately 450kb.

The existence of large regions of inter-marker LD does not, however, suggest that positive association can be detected with a marker set of lower density. One study showed that even within large regions of inter-marker LD, the extent of LD required to detect positive association declined at a rate such that beyond 10kb from the disease variant, positive association would not be detected (Durocher *et al.*, 2000). Instead, studies have shown that LD between polymorphisms varies and is not solely a factor of distance. Two polymorphisms separated by a few bases can be in almost equilibrium while two other polymorphisms separated by 1cM can be in strong LD (Mohike *et al.*, 2000; Pakstis *et al.*, 2000). Therefore, it would seem that a better association analysis strategy would be to genotype all polymorphisms showing incomplete disequilibrium within a candidate gene region rather than just one polymorphism every n^{th}kb. In agreement, Risch et al. (2000) have shown that negative association does not exclude a significant gene effect in the region and that the genome-wide random polymorphism approach, even at high density, would miss many disease-causing genes.

4.2.4.2. Population choice

Previously, it was thought that isolated and founder-effect populations were the best populations to choose for genetic disease association studies. It was believed that large regions of LD were recently created by novel disease mutation within the isolated populations and bottlenecks experienced by the founder-effect populations. However, it has recently been shown that common disease variants probably existed before the isolation or founder event and that more than one founder was a carrier of the variant.

Models of these populations show that the marker-variant LD within these populations would be the same as within the general population (Kruglyak, 1999).

It has now been suggested that European American populations, of constant size or inbred which have experienced recent admixture, may be better suited for common disease association studies (Wright *et al.*, 1999). This is for a number of reasons. Firstly, LD has repeatedly been shown to increase from Africa and India through Europe and East Asia to America due to the migration out of Africa and the accumulation of random genetic drift at the front of the expansion (Jorde *et al.*, 2000; Pakstis *et al.*, 2000). Secondly, recent populations undergoing admixture after long periods of separation (i.e. African American, Mexican American, etc) may still have sufficient LD remaining for detecting common disease association. Thirdly, populations that have been of a constant size rather than expanding, like the Saami in northern Fenno-Scandinavia, show greater extent of LD (Laan and Paabo, 1997). Finally, inbred populations show extended regions of LD because of increased homozygosity. Overall, because common disease variants probably arose prior to the expansion of the human population, all populations have had the same amount of time to decrease LD. Yet, because of events within certain populations that either decreased the breakdown of LD or increased LD, certain populations may require fewer markers to detect the same common disease association.

Still, it should be noted that the gains made by finding such a disease population might easily be removed by lack of power due to the small size of such populations. Thus, studies using large random American European populations may ultimately prove to be the best population choice for common disease association studies.

In the present study it was a small European Caucasian population that resulted in positive association to CCR5 (Section 4.2.1.1.). If instead, as commented throughout this chapter, a population of greater power had been used for linkage and association analysis, results of higher significance and greater certainty may have been detected.

## 4.2.4.3. Biallelic marker allele frequency choice

In has recently been shown that for association analysis it is optimal for the allelic frequencies of the biallelic polymorphisms chosen for genotyping to match the disease variant frequency (Muller-Myhsok and Abel, 1997). However, because for a common disease it is not known how many disease loci are present and how frequent the respective variants are, it is impossible to know what frequency biallelic polymorphisms to use for association analysis. Instead, it has been suggested that biallelic markers with a range of allele frequencies from 0.25-0.75 should be used for analysis within a disease region or across a candidate gene (Kruglyak and Lander, 1996). This would allow for the frequent and most common variants to be detected with two-point association analysis, while the less frequent variants could still be detected by haplotype association analysis.

In the candidate gene association studies reported here across GNAI2, CCR2, CCR5, and CCRL2, allele frequencies between 0.10-0.90 were used. Both two-point association analysis and haplotype association analysis were conducted. Thus, if the associations detected with both CCR2 and CCR5 are real, the CCR5 variant with stronger two-point association may be a more common disease variant and the CCR2 variant with stronger haplotype association may be a less common disease variant (Figure 3.27.).

## 4.2.4.4. Choice of biallelic marker location

It has been assumed that polymorphisms within coding regions, especially protein altering polymorphisms, are the polymorphisms most likely to be disease causing variants. Therefore, it would seem that screening only the coding regions for polymorphisms to be tested for association would be the best candidate gene association study strategy. Yet, as mentioned in the previous section, for common diseases a dense set of polymorphisms with allele frequencies between 0.25-0.75 may be best for association analysis. Recent studies have shown that both the presence of biallelic polymorphisms and their minor allele frequency decreases as you move from non-coding to synonymous coding to non-synonymous coding regions (Cargill *et al.*, 1999; Halushka *et al.*, 1999). Therefore, it is easier to find the suggested marker density and minor allele frequency required for common disease association analysis by screening the non-coding regions (i.e. introns, UTRs, and promoters).

In the present candidate gene association studies, the approach of only screening non-coding regions for polymorphism detection was employed. A total of 38 biallelic polymorphisms (Table 3.20.) with minor allele frequencies of 0.06-0.47 were detected in approximately 35 kb of genomic sequence giving a high-frequency (> 0.05), non-coding region polymorphism rate of about 1/900 bp. However, whilst this provided sufficient high-frequency polymorphisms to screen every 6 kb, after reviewing the recent information on extent of LD (Section 4.2.4.1.), it would appear that a strategy which included the screening for polymorphisms in both the coding region as well as the non-coding region combined with the analysis of every resulting high-frequency biallelic polymorphism would offer a better chance of detecting association.

## 4.3. Improvements in methods

### 4.3.1. Polymorphic marker maps

#### 4.3.1.1. RH maps

In the chromosome 3p saturation and extension linkage study, RH maps were constructed to order the selected STR markers and calculate inter-marker distance. Both the G4 and G3 RH panels were analysed. Using the G4 panel, most markers could not be ordered with certainty (Table 3.4.) since the average inter-marker physical distance was smaller than the resolution of the panel (1 Mb). Thus, the resulting G4 RH map was extremely discrepant when compared to the YAC contig physical maps (Figure 3.4.). Better correlation was seen between the YAC maps and the G3 maps generated, because of the higher resolution of the G3 RH panel (250 kb) (Figure 3.4.). Still, some markers could not be ordered with certainty (Table 3.4.). To help resolve this problem a new RH mapping panel, TNG, is now available from Research Genetics. Using the TNG panel, markers as close as 30kb can be ordered and positioned with certainty. If the TNG panel was used in the present study, due to the increased accuracy of marker order, potentially fewer double recombinants would have be observed leading to the removal of less genotype data and a more accurate reflection of IBD linkage in the 3p region.

#### 4.3.1.2. Physical maps

With the pending completion of the human genome sequence, a physical map which includes all known DNA elements (i.e. genes, ESTs, STSs) will exist. This map will replace all existing human RH and physical maps and remove the necessity of making RH and physical maps like those constructed for the 3p and 6p linkage studies and the

3p candidate gene association studies, respectively. However, because of the assumptions made regarding recombination rate during linkage analysis (Section 4.1.4.), genetic maps may continue to be used for the majority of low density STR linkage studies. Yet, if SNPs rather than STRs are used for linkage analysis in the future (Section 4.3.2.4.), the human physical map may be used for linkage analysis regardless of the inaccuracies caused by the recombination rate assumption due to the required increase in marker density and thus the inability to produce biallelic genetic maps with small populations.

## 4.3.2. Linkage analysis

4.3.2.1. Capillary electrophoresis

Originally, to analyse STR markers, PCR products were radioactively labelled and separated on large acrylamide slab gels. With this system only markers of non-overlapping size ranges could be analysed together. Then in 1991, the ABI 373 became available allowing detection and separation of up to 3 STR markers of the same size range by labelling each marker with a different fluorescent molecule. The ABI 373 could analyse up to 20 markers across 48 samples simultaneously, increasing the genotyping throughput 5-fold. Subsequently in 1994 ABI released a new acrylamide slab gel fluorescent DNA analyser called the ABI 377. The ABI 377, used in the present study, provided a better resolution of DNA fragments resulting in more accurate and reliable genotypes. It also increased the number of samples that could be analysed per gel from 48 to 96, doubling the genotyping throughput. Even so, genotype fragment analyses continued to be time consuming. Acrylamide gels needed to be poured and allowed to solidify for 2 hours. Each sample needed to be loaded by hand, making 24 hour analysis only achievable by shift work. In addition, each lane needed to be

manually tracked before the data could be extracted and analysed. Today, these issues

have been solved by the invention of fluorescent capillary electrophoresis. Instead of

using acrylamide slab gels, the DNA is separated by electrophoresis through 96 silica

capillaries filled with acrylamide. The liquid acrylamide polymer, injected into the

capillaries at the start of each run, dynamically binds to the silica. Thus, the acrylamide

does not require time to solidify. Also, because each fluorescent DNA fragment is

detected either within the capillary or immediately upon exiting the capillary, the data

lanes are constant and do not need to be tracked. Presently, two companies supply

capillary electrophoresis DNA analysers, Amersham/Pharmacia Biotech (APB) and

Applied Biosystems, Inc (ABI). However, only the ABI 3700 is completely robotic

allowing for 24 hour unattended analyses. This development has virtually eliminated

the vast amount of time spent on fragment analysis in the past. Previously, with one

ABI 377 it would have taken approximately 6 days to analyse the data required for a 10

cM genome scan across 96 subjects. Now, with one ABI 3700 the same scan can be

conducted in 3 days, requiring at most 3 hours of hands on work.


4.3.2.2. Genotype quality scores

Recently, APB and ABI have developed new genotyping software that automatically

pre-processes genotype data for quality and reliability of each genotype. A quality

score is assigned to each genotype enabling the genotypes to be ordered from most to

least reliable. Previously, without quality scores, each genotype needed to be visually

checked for miscalls because of anomalies like preferential amplification, high-stutter

and plus-A. The process was extremely time consuming. For example, in the present

linkage studies a month of full time work was required to visually check the 62,000

genotypes generated. However, when the quality score system is used less than 10% of

the genotypes require visual checking before confidence in the validity of the remaining

genotypes is achieved. Thus, with the quality score system, 40,000 genotypes generated

in a 10 cM genome scan across 96 individuals could be called and quality checked

within a day.

4.3.2.3. Genotype error detection

Genotype errors are observed genotypes that do not correspond to actual genotypes.

Genotype errors that remain undetected could have dramatic effects on linkage results.

A recent study showed that the introduction of only 1% genotype error could half the

resulting LOD score (Douglas *et al.*, 2000). Thus, in a genome scan with 1% genotype

error significant linkage (LOD > 3.6) could easily go undetected. In the present study,

PedCheck and CRI-MAP were used to detect genotype error. However, in a population

like the Oxford linkage cohort (Section 2.2.1.), where several parental genotypes are

missing, both of these methods of error detection fail to detect all genotype errors.

Recently, new programs (i.e. SIBMED, MERLIN, SIMWALK) with a hidden Markov

method for detecting genotype error have been described (Abecasis *et al.*, 2000;

Douglas *et al.*, 2000). These programs can detect up to 50% of genotype error, and

almost all errors that have an impact on linkage analysis. Also, unlike programs like

CRI-MAP, removal of questionable genotypes will not generate false evidence of

linkage. When these new programs become publicly available, the accuracy of linkage

results will be greatly improved.

4.3.2.4. Linkage analysis using SNPs

With the recent interest and developments in SNP genotyping, the option of conducting

linkage experiment using SNP markers rather STR markers has been proposed. The

main benefit of using STR markers over SNP markers is the larger number of alleles in

STR markers that result in higher marker heterozygosities and therefore higher

information content. Thus, in a genome scan, fewer STR markers need to be genotyped to thoroughly scan for linkage. It has recently been shown that 700-900 SNP markers would be required to provide equivalent results to the 300-400 STR markers currently used for genome scans (Kruglyak, 1997). Although the number of markers which need to be genotyped is greater, with the possible higher-throughput of SNP genotyping (Section 4.3.3.) combined with the benefits for association analysis of increased marker density and non-recurring mutation (Section 4.1.3.), the switch from STR to SNP genotyping for linkage analysis remains an attractive option for future experiments.

### 4.3.3. High-throughput SNP genotyping

With the recent sequencing of the human genome, thousands of common SNPs have been discovered across every chromosome. Thus, where previously high-density linkage and association experiments using SNP markers were impractical because relatively few SNPs were known and mapped, now several genetic experiments using SNPs are underway. Still, whilst this first hurdle has now been crossed, the question of how to rapidly genotype SNPs remains.

In the association studies reported here, PCR-RFLP was used for SNP genotyping. This method, although the method of choice for several years, is not a high-throughput genotyping option for several reasons. Firstly, a high-throughput SNP genotyping method should be capable of genotyping most SNPs. PCR-RFLP can only be used to genotype SNPs that create a restriction site or for which a restriction site can be created using a mismatch oligonucleotide. Thus, several SNPs are unable to be genotyped by PCR-RFLP. Secondly, a high-throughput SNP genotyping method should offer a low fixed cost per genotype. PCR-RFLP has a variable cost per genotype dependent on the cost of the restriction enzyme required, which can be as high as £5 per genotype.

Thirdly, a high-throughput SNP genotyping method should not require gels for analysis

due to the time needed for their manufacture and loading. PCR-RFLP requires the use

of agarose gels for DNA fragment analysis. Fourthly, a high throughput SNP

genotyping method should produce results that are easily and automatically scored, thus

providing results which are consistent and of a specified quality. PCR-RFLP requires

visual gel image interpretation. When this is combined with incomplete digestion,

which occurs frequently, image-to-image and day-to-day genotype calling

inconsistencies are prevalent. To avoid this problem, each subject should be genotyped

twice for each assay and only concordant results included. This corrects for the

inconsistencies arising from interpreted rather than quantified results, however it also

doubles the time and cost required for genotyping. Finally, a high-throughput SNP

genotyping method should include software that can automatically export results for

analysis without the need for human data entry. For PCR-RFLP, there is no automated

allele calling software and each genotype has to be typed into a database by hand.

Therefore, to avoid typing errors, each genotype should be entered twice and compared

for concordance to identify data entry errors. Again, this adds time to the process of

SNP genotype analysis. Thus, although PCR-RFLP was a useful SNP genotyping

method for this study and several previous studies, it is not a high-throughput

genotyping method able to handle the SNP genotyping requirements of future genetic

experiments.


Recently, several new SNP genotyping methods that fulfil most of the above

requirements of a high-throughput system have been developed. Two methods, single

base extension (SBE) and pyrosequencing™, detect SNP genotypes by sequencing off

the end of an oligonucleotide positioned adjacent to the SNP. With SBE, following the

mini-sequencing reaction, the results can either be detected for a single assay using a

fluorescent plate reader or for multiple assays using capillary electrophoresis (Section

4.3.2.1.). With pyrosequencing™, the sequencing reaction and genotype detection are

conducted simultaneously within the same instrument. Other methods, like SNiPer™,

TaqMan™ and microarrays use allele specific probes to distinguish between alleles.

With TaqMan™, annealed allele-specific oligonucleotide probes fluoresce after a

quencher is digested by the exonuclease activity of DNA Taq polymerase. With

SNiPer™, allele-specific probes are circularised by ligation if the matching allele is

present. Then, rolling circle amplification (RCA) with quenched probe-specific

fluorescent oligonucleotides is conducted and allele-specific fluorescence is detected.

With microarrays, allele-specific oligonucleotides are spotted onto a 'chip' and

subsequent genotype-specific annealing of subject DNA detected. Thousands of assays

can be spotted on one 'chip' and analysed simultaneously. Altogether, microarrays may

offer the highest throughput possible. However, to achieve this throughput, thousands

of the recently discovered SNPs would have to be verified and would need assays to be

designed and tested before a 'SNP chip' could be constructed. This process will take

some time. Therefore, until the required information for rapidly making high-density

'SNP chips' across all genomic regions is available, the other new SNP genotyping

techniques will predominately be used for SNP analysis.


### 4.3.4. Whole genome LD mapping

With the recent increase in information regarding the extent of LD and the hypotheses

regarding required marker density for detection of association to common diseases

(Section 4.2.4.1.), it has been estimated that up to 500,000 SNPs may be required to

conduct a whole-genome LD study for a common disease gene (Kruglyak, 1999). Soon,

due to the discovery of over 500,000 SNPs during the sequencing of the human genome

combined with the SNP genotyping throughput of microarrays (Section 4.3.3.) this previously impossible experiment may soon be common place.

Still, as discussed in Section 4.2.4., some forethought should be invested into the type and position of SNP used. Non-coding SNPs may provide better power to detect association to common diseases. However, because LD between two markers is not predictable both positive and negative association may not useful in narrowing a disease region or finding a disease gene. Potentially, unless the disease mutation is itself being analysed, positive association may be meaningless. Therefore, whilst non-coding SNPs provide better power for common disease association analysis, it may be more appropriate to conduct whole-genome LD studies using only potential gene, expression and transcription altering SNPs of sufficient power, rather than high frequency SNPs evenly spaced across the genome.

## 4.4. Conclusion

In the genetic studies reported here, linkage analysis was conducted across the putative IBD susceptibility regions on chromosomes 3 and 6. Linkage was only detected between chromosome 6p and IBD, replicating previous findings of IBD linkage to the region and confirming the region as the *IBD3* susceptibility locus. The failure to extend linkage to the chromosome 3 region suggests that the initial report of linkage between IBD and 3p may be a false positive. This conclusion is supported by the SIBS1&2 analysis, which also showed significantly less linkage than the original genome scan. Overall, the linkage data suggests that the 3p region confers a small effect on IBD susceptibility, if any at all, and that the *IBD3* locus may confer a larger effect on IBD susceptibility than both the formerly replicated *IBD1* and *IBD2* loci.

Concurrently, due to the initial report of IBD linkage to chromosome 3, association

studies across the region and candidate genes within it were conducted. Positive

association was seen with many markers across the region. However, only the

association detected between the 32bp deletion in CCR5 and UC would remain

significant after correcting for multiple testing. Fortunately, CCR5Δ32 is a functional

polymorphism suggesting that if the association detected is real, then CCR5Δ32 may

actually a IBD disease mutation rather than just a polymorphism associated by LD to

the mutation.

# References

Abecasis, G.R., Cherney, S.S., Cookson, W.O.C. and Cardon, L.R. MERLIN - Multipoint engine for rapid likelihood inference. *Am J Hum Genet* 2000; **67:** 327.

Aithal, G.P., Day, C.P., Leathart, J., Daly, A.K. and Hudson, M. Association of single nucleotide polymorphisms in the interleukin-4 gene and interleukin-4 receptor gene with Crohn's disease in a British population. *Genes Immun* 2001; **2:** 44-47.

Andres, P.G., Beck, P.L., Mizoguchi, E., Mizoguchi, A., Bhan, A.K., Dawson, T., Kuziel, W.A., Maeda, N., MacDermott, R.P., Podolsky, D.K. and Reinecker, H.C. Mice with a selective deletion of the CC chemokine receptors 5 or 2 are protected from dextran sodium sulfate-mediated colitis: lack of CC chemokine receptor 5 expression results in a NK1.1+ lymphocyte-associated Th2-type immune response in the intestine. *J Immunol* 2000; **164:** 6303-6312.

Andres, P.G. and Friedman, L.S. Epidemiology and the natural course of inflammatory bowel disease. *Gastroenterol Clin North Am* 1999; **28:** 225-281, vii.

Annese, V., Latiano, A., Bovio, P., Forabosco, P., Piepoli, A., Lombardi, G., Andreoli, A., Astegiano, M., Gionchetti, P., Riegler, G., Sturniolo, G.C., Clementi, M., Rappaport, E., Fortina, P., Devoto, M., Gasparini, P. and Andriulli, A. Genetic analysis in Italian families with inflammatory bowel disease supports linkage to the IBD1 locus-- a GISC study. *Eur J Hum Genet* 1999; **7:** 567-573.

Bachvarov, D.R., Landry, M., Houle, S., Pare, P., and Marceau, F. Altered Frequency of a promoter polymorphic allele of the kinin B1 receptor gene in inflammatory bowel disease. *Gastroenterol* 1998, **115:** 1286-1289.

Bansi, D.S., Lo, S., Chapman, R.W. and Fleming, K.A. Absence of antineutrophil cytoplasmic antibodies in relatives of UK patients with primary sclerosing cholangitis and ulcerative colitis. *Eur J Gastroenterol Hepatol* 1996; **8:** 111-116.

Bayless, T.M., Tokayer, A.Z., Polito, J.M., 2nd, Quaskey, S.A., Mellits, E.D. and Harris, M.L. Crohn's disease: concordance for site and clinical type in affected family members--potential hereditary influences. *Gastroenterology* 1996; **111:** 573-579.

Becker, K.G., Simon, R.M., Bailey-Wilson, J.E., Freidlin, B., Biddison, W.E., McFarland, H.F. and Trent, J.M. Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases. *Proc Natl Acad Sci U S A* 1998; **95:** 9979-9984.

Binder, V. and Orholm, M. Familial occurrence and inheritance studies in inflammatory bowel disease. *Neth J Med* 1996; **48:** 53-56.

Bioque, G., Bouma, G., Crusius, J.B., Koutroubakis, I., Kostense, P.J., Meuwissen, S.G. and Pena, A.S. Evidence of genetic heterogeneity in IBD: 1. The interleukin-1 receptor antagonist in the predisposition to suffer from ulcerative colitis. *Eur J Gastroenterol Hepatol* 1996; **8:** 105-110.

Bioque, G., Crusius, J.B., Koutroubakis, I., Bouma, G., Kostense, P.J., Meuwissen, S.G. and Pena, A.S. Allelic polymorphism in IL-1 beta and IL-1 receptor antagonist (IL-1Ra) genes in inflammatory bowel disease. *Clin Exp Immunol* 1995; **102:** 379-383.

Bjornsson, S., Johannsson, J.H. and Oddsson, E. Inflammatory bowel disease in Iceland, 1980-89. A retrospective nationwide epidemiologic study. *Scand J Gastroenterol* 1998; **33:** 71-77.

Boehnke, M., Lange, K. and Cox, D.R. Statistical methods for multipoint radiation hybrid mapping. *Am J Hum Genet* 1991; **49:** 1174-1188.

Bouma, G., Oudkerk Pool, M., Crusius, J.B., Schreuder, G.M., Hellemans, H.P., Meijer, B.U., Kostense, P.J., Giphart, M.J., Meuwissen, S.G. and Pena, A.S. Evidence for genetic heterogeneity in inflammatory bowel disease (IBD); HLA genes in the predisposition to suffer from ulcerative colitis (UC) and Crohn's disease (CD). *Clin Exp Immunol* 1997; **109:** 175-179.

Bouma, G., Poen, A.C., Garcia-Gonzalez, M.A., Schreuder, G.M.T., Felt-Bersma, R.J.F., Meuwissen, S.G.M. and Pena, A.S. HLA-DRB1*03, but not the TNFA -308 promoter gene polymorphism, confers protection against fistulising Crohn's disease. *Immunogenetics* 1998; **47:** 451-455.

Bouma, G., Xia, B., Crusius, J.B., Bioque, G., Koutroubakis, I., Von Blomberg, B.M., Meuwissen, S.G. and Pena, A.S. Distribution of four polymorphisms in the tumour necrosis factor (TNF) genes in patients with inflammatory bowel disease (IBD). *Clin Exp Immunol* 1996; **103:** 391-396.

Brant, S.R., Fu, Y., Fields, C.T., Baltazar, R., Ravenhill, G., Pickles, M.R., Rohal, P.M., Mann, J., Kirschner, B.S., Jabs, E.W., Bayless, T.M., Hanauer, S.B. and Cho, J.H. American families with Crohn's disease have strong evidence for linkage to chromosome 16 but not chromosome 12. *Gastroenterology* 1998; **115:** 1056-1061.

Brant, S.R., Panhuysen, C.I., Bailey-Wilson, J.E., Rohal, P.M., Lee, S., Mann, J., Ravenhill, G., Kirschner, B.S., Hanauer, S.B., Cho, J.H. and Bayless, T.M. Linkage heterogeneity for the IBD1 locus in Crohn's disease pedigrees by disease onset and severity. *Gastroenterology* 2000; **119:** 1483-1490.

Breslin, N.P., Todd, A., Kilgallen, C. and O'Morain, C. Monozygotic twins with Crohn's disease and ulcerative colitis: a unique case report. *Gut* 1997; **41:** 557-560.

Brownstein, M.J., Carpten, J.D. and Smith, J.R. Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *Biotechniques* 1996; **20:** 1004-1006, 1008-1010.

Carey, A.H. Use of Commercially Available Radiation Hybrid Panels; in Dracopoli, N.C., Haines, J.L., Korf, B.R., Moir, D.T., Morton, C.C., Seidman, C.E., Seidman, J.G. and Smith, D.R. (eds.): *Current Protocols in Human Genetics.* New York, 1997; **1:** 3.5.1-3.5.10.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C.R., Lim, E.P., Kalayanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G.Q. and Lander, E.S. Characterisation of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999; **22:** 231-238.

Carter, M.J., di Giovine, F.S., Jones, S., Mee, J., Camp, N.J., Lobo, A.J. and Duff, G.W. Association of the interleukin 1 receptor antagonist gene with ulcerative colitis in Northern European Caucasians. *Gut* 2001; **48:** 461-467.

Cavanaugh, J.A., Callen, D.F., Wilson, S.R., Stanford, P.M., Sraml, M.E., Gorska, M., Crawford, J., Whitmore, S.A., Shlegel, C., Foote, S., Kohonen-Corish, M. and Pavli, P. Analysis of Australian Crohn's disease pedigrees refines the localization for susceptibility to inflammatory bowel disease on chromosome 16. *Ann Hum Genet* 1998; **62:** 291-298.

Cho, J.H., Nicolae, D.L., Gold, L.H., Fields, C.T., LaBuda, M.C., Rohal, P.M., Pickles, M.R., Qin, L., Fu, Y., Mann, J.S., Kirschner, B.S., Jabs, E.W., Weber, J., Hanauer, S.B., Bayless, T.M. and Brant, S.R. Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and IBD1. *Proc Natl Acad Sci U S A* 1998; **95:** 7502-7507.

Cipolla, C., Magliocco, A., Oliva, L. and Cottone, M. Familial aggregation of inflammatory bowel disease in a Mediterranean area. *Eur J Epidemiol* 1996; **12:** 205-210.

Collins, A., Lonjou, C. and Morton, N.E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci U S A* 1999; **96:** 15173-15177.

Colombel, J.F., Grandbastien, B., Gower-Rousseau, C., Plegat, S., Evrard, J.P., Dupas, J.L., Gendre, J.P., Modigliani, R., Belaiche, J., Hostein, J., Hugot, J.P., van Kruiningen, H. and Cortot, A. Clinical characteristics of Crohn's disease in 72 families. *Gastroenterology* 1996; **111:** 604-607.

Curran, M.E., Lau, K.F., Hampe, J., Schreiber, S., Bridger, S., Macpherson, A.J., Cardon, L.R., Sakul, H., Harris, T.J., Stokkers, P., Van Deventer, S.J., Mirza, M., Raedler, A., Kruis, W., Meckler, U., Theuer, D., Herrmann, T., Gionchetti, P., Lee, J., Mathew, C. and Lennard-Jones, J. Genetic analysis of inflammatory bowel disease in a large European cohort supports linkage to chromosomes 12 and 16. *Gastroenterology* 1998; **115:** 1066-1071.

Daly, M.J., Kruglyak, L., Pratt, S., Houstis, N., Reeve, M.P., Kirby, A. and Lander, E.S. GENEHUNTER 2.0 - a complete linkage analysis system. *Am J Hum Genet Suppl* 1998; **63:** A286.

Danze, P.M., Colombel, J.F., Jacquot, S., Loste, M.N., Heresbach, D., Ategbo, S., Khamassi, S., Perichon, B., Semana, G., Charron, D. and Cezard, J.P. Association of HLA class II genes with susceptibility to Crohn's disease. *Gut* 1996; **39:** 69-72.

De La Concha, E.G., Fernandez-Arquero, M., Santa-Cruz, S., Lopez-Nava, G., Figueredo, M.A., Diaz-Rubio, M. and Garcia-Paredes, J. Positive and negative associations of distinct HLA-DR2 subtypes with ulcerative colitis (UC). *Clin Exp Immunol* 1997; **108:** 392-395.

Douglas, J.A., Boehnke, M. and Lange, K. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *Am J Hum Genet* 2000; **66:** 1287-1297.

Duerr, R.H., Barmada, M.M., Zhang, L., Davis, S., Preston, R.A., Chensny, L.J., Brown, J.L., Ehrlich, G.D., Weeks, D.E. and Aston, C.E. Linkage and association between inflammatory bowel disease and a locus on chromosome 12. *Am J Hum Genet* 1998; **63:** 95-100.

Duerr, R.H., Barmada, M.M., Zhang, L., Pfutzer, R. and Weeks, D.E. High-Density Genome Scan in Crohn Disease Shows Confirmed Linkage to Chromosome 14q11-12. *Am J Hum Genet* 2000; **66:** 1857-1862.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K.N., Beasley, O., Bird, C.P., Blakey, S., Bridgeman, A.M., Buck, D., Burgess, J., Burrill, W.D., O'Brien, K.P. and et al. The DNA sequence of human chromosome 22. *Nature* 1999; **402:** 489-495.

Durocher, F., Dunning, A., Healey, C., Teare, D., McBride, S., Xu, C.F., Lai, E., Dawson, E., Rhodes, S., Purvis, I., Easton, D. and Ponder, B. Comparison of linkage disequilibrium (LD) across long physical distances in four populations with different demographic histories. *Am J Hum Genet* 2000; **67.**

Feeney, M., Ciegg, A., Winwood, P. and Snook, J. A case-control study of measles vaccination and inflammatory bowel disease. The East Dorset Gastroenterology Group. *Lancet* 1997; **350:** 764-766.

Fernandez Arquero, M., Lopez Nava, G., De la Concha, E.G., Figueredo, M.A., Santa Cruz, S., Dumitru, C.G., Diaz Rubio, M. and Garcia Paredes, J. HLA-DR2 gene and Spanish patients with ulcerative colitis. *Rev Esp Enferm Dig* 1998; **90:** 243-249.

Folwaczny, C., Zahn, R., Brunnler, G., Albert, E., Koenig, A., Noehl, N., Loeschke, K. and Fricke, H. Lack of association between HLA-DRB1 alleles of the major histocompatibility complex and p-ANCA status or clinical characteristics in patients with ulcerative colitis. *Z Gastroenterol* 1999; **37:** 133-140.

Fonager, K., Sorensen, H.T. and Olsen, J. Change in incidence of Crohn's disease and ulcerative colitis in Denmark. A study based on the National Registry of Patients, 1981-1992. *Int J Epidemiol* 1997; **26:** 1003-1008.

Forcione, D.G., Sands, B., Isselbacher, K.J., Rustgi, A., Podolsky, D.K. and Pillai, S. An increased risk of Crohn's disease in individuals who inherit the HLA class II DRB3*0301 allele. *Proc Natl Acad Sci U S A* 1996; **93:** 5094-5098.

Hacker, U.T., Bidlingmaier, C., Gomolka, M., Keller, E., Eigler, A., Hartmann, G., Folwaczny, C., Fricke, H., Albert, E., Loeschke, K. and Endres, S. Inflammatory bowel disease: no association between allele combinations of the interleukin (IL) I beta and IL-I receptor antagonist gene polymorphisms. *Eur J Clin Invest* 1998; **28:** 214-219.

Hacker, U.T., Gomolka, M., Keller, E., Eigler, A., Folwaczny, C., Fricke, H., Albert, E., Loeschke, K. and Endres, S. Lack of association between an interleukin-1 receptor antagonist gene polymorphism and ulcerative colitis. *Gut* 1997; **40:** 623-627.

Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R. and Chakravarti, A. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet* 1999; **22:** 239-247.

Hampe, J., Hermann, B., Bridger, S., MacPherson, A.J., Mathew, C.G. and Schreiber, S. The interferon-gamma gene as a positional and functional candidate gene for inflammatory bowel disease. *Int J Colorectal Dis* 1998; **13:** 260-263.

Hampe, J., Lynch, N.J., Daniels, S., Bridger, S., Macpherson, A.J., Stokkers, P., Forbes, A., Lennard-Jones, J.E., Mathew, C.G., Curran, M.E. and Schreiber, S. Fine mapping of the chromosome 3p susceptibility locus in inflammatory bowel disease. *Gut* 2001; **48:** 191-197.

Hampe, J., Schreiber, S., Shaw, S.H., Lau, K.F., Bridger, S., Macpherson, A.J., Cardon, L.R., Sakul, H., Harris, T.J., Buckler, A., Hall, J., Stokkers, P., van Deventer, S.J., Nurnberg, P., Mirza, M.M., Lee, J.C., Lennard-Jones, J.E., Mathew, C.G. and Curran, M.E. A genomewide analysis provides evidence for novel linkages in inflammatory bowel disease in a large European cohort. *Am J Hum Genet* 1999a; **64:** 808-816.

Hampe, J., Shaw, S.H., Saiz, R., Leysens, N., Lantermann, A., Mascheretti, S., Lynch, N.J., MacPherson, A.J., Bridger, S., van Deventer, S., Stokkers, P., Morin, P., Mirza, M.M., Forbes, A., Lennard-Jones, J.E., Mathew, C.G., Curran, M.E. and Schreiber, S. Linkage of inflammatory bowel disease to human chromosome 6p. *Am J Hum Genet* 1999b; **65:** 1647-1655.

Hayward, P.A., Satsangi, J. and Jewell, D.P. Inflammatory bowel disease and the X chromosome. *Qjm* 1996; **89:** 713-718.

Heresbach, D., Alizadeh, M., Dabadie, A., Le Berre, N., Colombel, J.F., Yaouanq, J., Bretagne, J.F. and Semana, G. Significance of interleukin-1beta and interleukin-1 receptor antagonist genetic polymorphism in inflammatory bowel diseases. *Am J Gastroenterol* 1997; **92:** 1164-1169.

Heresbach, D., Alizadeh, M., Reumaux, D., Colombel, J.F., Delamaire, M., Danze, P.M., Gosselin, M., Genetet, B., Bretagne, J.F. and Semana, G. Are HLA-DR or TAP genes genetic markers of severity in ulcerative colitis? *J Autoimmun* 1996; **9:** 777-784.

Hesresbach, D., Alizadeh, M., Bretagne, J.F., Gautier, A., Quillivic, F., Lemarchand, B., Gosselin, M., Genetet, B. and Semana, G. Investigation of the association of major histocompatibility complex genes, including HLA class I, class II and TAP genes, with clinical forms of Crohn's disease. *Eur J Immunogenet* 1996; **23:** 141-151.

Hofmeister, A., Neibergs, H.L., Pokorny, R.M. and Galandiuk, S. The natural resistance-associated macrophage protein gene is associated with Crohn's disease. *Surgery* 1997; **122:** 173-178; discussion 178-179.

219

Hollander, D., Vadheim, C.M., Brettholz, E., Petersen, G.M., Delahunty, T. and Rotter, J.I. Increased intestinal permeability in patients with Crohn's disease and their relatives. A possible etiologic factor. *Ann Intern Med* 1986; **105:** 883-885.

Hugot, J.P., Laurent-Puig, P., Gower-Rousseau, C., Caillat-Zucman, S., Beaugerie, L., Dupas, J.L., Van Gossum, A., Bonait-Pellie, C., Cortot, A. and Thomas, G. Linkage analyses of chromosome 6 loci, including HLA, in familial aggregations of Crohn disease. G.E.T.A.I.D. *Am J Med Genet* 1994; **52:** 207-213.

Hugot, J.P., Laurent-Puig, P., Gower-Rousseau, C., Olson, J.M., Lee, J.C., Beaugerie, L., Naom, I., Dupas, J.L., Van Gossum, A., Orholm, M., Bonaiti-Pellie, C., Weissenbach, J., Mathew, C.G., Lennard-Jones, J.E., Cortot, A., Colombel, J.F. and Thomas, G. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 1996; **379:** 821-823.

Jawaheer, D., Seldin, M.F., Amos, C.I., Chen, W.V., Shigeta, R., Monteiro, J., Kern, M., Criswell, L.A., Albani, S., Nelson, J.L., Clegg, D.O., Pope, R., Schroeder, H.W., Jr., Bridges, S.L., Jr., Pisetsky, D.S., Ward, R., Kastner, D.L., Wilder, R.L., Pincus, T., Callahan, L.F., Flemming, D., Wener, M.H. and Gregersen, P.K. A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *Am J Hum Genet* 2001; **68:** 927-936.

Jorde, L.B., Watkins, W.S., Bamshad, M.J., Dunn, D. and Weiss, R. Snp variation and linkage disequilibrium near the CYP1A2 gene. *Am J Hum Genet* 2000; **67:** 24.

Kim, W.H., Choi, P.M., Landers, C.J. and Targan, S.R. Role of antineutrophil cytoplasmic antibodies in an ethnically distinct population: Korean patients with ulcerative colitis. *Am J Gastroenterol* 1995; **90:** 1953-1958.

Kitahora, T., Utsunomiya, T. and Yokota, A. Epidemiological study of ulcerative colitis in Japan: incidence and familial occurrence. The Epidemiology Group of the Research Committee of Inflammatory Bowel Disease in Japan. *J Gastroenterol* 1995; **30 Suppl 8:** 5-8.

Koss, K., Satsangi, J., Welsh, K.I. and Jewell, D.P. Is interleukin-6 important in inflammatory bowel disease? *Genes Immun* 2000; **1:** 207-212.

Kruglyak, L. The use of a genetic map of biallelic markers in linkage studies. *Nat Genet* 1997; **17:** 21-24.

Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 1999; **22:** 139-144.

Kruglyak, L. and Lander, E.S. Limits on fine mapping of complex traits. *Am J Hum Genet* 1996; **58:** 1092-1093.

Kuster, W., Pascoe, L., Purrmann, J., Funk, S. and Majewski, F. The genetics of Crohn disease: complex segregation analysis of a family study with 265 patients with Crohn disease and 5,387 relatives. *Am J Med Genet* 1989; **32:** 105-108.

Kyo, K., Muto, T., Nagawa, H., Lathrop, G.M. and Nakamura, Y. Associations of

distinct variants of the intestinal mucin gene MUC3A with ulcerative colitis and Crohn's

disease. *J Hum Genet* 2001; **46:** 5-20.


Kyo, K., Parkes, M., Takei, Y., Nishimori, H., Vyas, P., Satsangi, J., Simmons, J.,

Nagawa, H., Baba, S., Jewell, D., Muto, T., Lathrop, G.M. and Nakamura, Y.

Association of ulcerative colitis with rare VNTR alleles of the human intestinal mucin

gene, MUC3. *Hum Mol Genet* 1999; **8:** 307-311.


Laan, M. and Paabo, S. Demographic history and linkage disequilibrium in human

populations. *Nat Genet* 1997; **17:** 435-438.


Lander, E. and Kruglyak, L. Genetic dissection of complex traits: guidelines for

interpreting and reporting linkage results. *Nat Genet* 1995; **11:** 241-247.


Lander, E.S. and Green, P. Construction of multilocus genetic linkage maps in humans.
*Proc Natl Acad Sci U S A* 1987; **84:** 2363-2367.


Lashner, B.A. Epidemiology of inflammatory bowel disease. *Gastroenterol Clin North
Am* 1995; **24:** 467-474.


Lee, J.C. and Lennard-Jones, J.E. Inflammatory bowel disease in 67 families each with

three or more affected first-degree relatives. *Gastroenterology* 1996; **111:** 587-596.


Lennard-Jones, J.E. Classification of inflammatory bowel disease. *Scand J
Gastroenterol Suppl* 1989; **170:** 2-6; discussion 16-19.

Lesage, S., Zouali, H., Colombel, J.F., Belaiche, J., Cezard, J.P., Tysk, C., Almer, S.,

Gassull, M., Binder, V., Chamaillard, M., Le Gall, I., Thomas, G. and Hugot, J.P.

Genetic analyses of chromosome 12 loci in Crohn's disease. *Gut* 2000; **47:** 787-791.


Louis, E., Satsangi, J., Roussomoustakaki, M., Parkes, M., Fanning, G., Welsh, K. and

Jewell, D. Cytokine gene polymorphisms in inflammatory bowel disease. *Gut* 1996; **39:**

705-710.


Ma, Y., Ohmen, J.D., Li, Z., Bentley, L.G., McElree, C., Pressman, S., Targan, S.R.,

Fischel-Ghodsian, N., Rotter, J.I. and Yang, H. A genome-wide search identifies

potential new susceptibility loci for Crohn's disease. *Inflamm Bowel Dis* 1999; **5:** 271-

278.


MacDermott, R.P., Sanderson, I.R. and Reinecker, H.C. The central role of chemokines

(chemotactic cytokines) in the immunopathogenesis of ulcerative colitis and Crohn's

disease. *Inflamm Bowel Dis* 1998; **4:** 54-67.


Manousos, O.N., Giannadaki, E., Mouzas, I.A., Tzardi, M., Koutroubakis, I., Skordilis,

P., Vassilakis, S., Kouroumalis, E. and Vlachonikolis, I.G. Ulcerative colitis is as

common in Crete as in northern Europe: a 5- year prospective study. *Eur J*

*Gastroenterol Hepatol* 1996; **8:** 893-898.

Mansfield, J.C., Holden, H., Tarlow, J.K., Di Giovine, F.S., McDowell, T.L., Wilson, A.G., Holdsworth, C.D. and Duff, G.W. Novel genetic association between ulcerative colitis and the anti- inflammatory cytokine interleukin-1 receptor antagonist. *Gastroenterology* 1994; **106:** 637-642.

Martin, K., Heinzlmann, M., Borchers, R., Mack, M., Loeschke, K. and Folwaczny, C. Delta 32 mutation of the chemokine-receptor 5 gene in inflammatory bowel disease. *Clin Immunol* 2001; **98:** 18-22.

Martinson, J.J., Chapman, N.H., Rees, D.C., Liu, Y.T. and Clegg, J.B. Global distribution of the CCR5 gene 32-basepair deletion. *Nat Genet* 1997; **16:** 100-103.

Matise, T.C., Perlin, M. and Chakravarti, A. Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. *Nat Genet* 1994; **6:** 384-390.

McConnell, R.B. and Vadheim, C.M. Inflammatory Bowel Disease; in King, R.A., Rotter, J.I. and Motulsky, A.O. (eds.): *The Genetics Basis of Common Diseases.* Oxford, 1992: 326-348.

McMahon, F.J., Chen, Y.S., McInnis, M.G. and Yarnes, C. The extent of linkage disequilibrium in European Americans. *Am J Hum Genet* 2000; **67:** 225.

Mirza, M.M., Lee, J., Teare, D., Hugot, J.P., Laurent-Puig, P., Colombel, J.F., Hodgson, S.V., Thomas, G., Easton, D.F., Lennard-Jones, J.E. and Mathew, C.G. Evidence of linkage of the inflammatory bowel disease susceptibility locus on chromosome 16 (IBD1) to ulcerative colitis. *J Med Genet* 1998; **35:** 218-221.

Mohike, K.L., Lange, E.M., Valle, T., Collins, F.S. and Boehnke, M. Marker-marker linkage disequilibrium extends 1 cM on chromosome 20 in Finns. *Am J Hum Genet* 2000; **67:** 25.

Muller-Myhsok, B. and Abel, L. Genetic analysis of complex diseases. *Science* 1997; **275:** 1328-1329; discussion 1329-1330.

Mummidi, S., Ahuja, S.S., Gonzalez, E., Anderson, S.A., Santiago, E.N., Stephan, K.T., Craig, F.E., O'Connell, P., Tryon, V., Clark, R.A., Dolan, M.J. and Ahuja, S.K. Genealogy of the CCR5 locus and chemokine system gene variants associated with altered rates of HIV-1 disease progression. *Nat Med* 1998; **4:** 786-793.

Mummidi, S., Ahuja, S.S., McDaniel, B.L. and Ahuja, S.K. The human CC chemokine receptor 5 (CCR5) gene. Multiple transcripts with 5'-end heterogeneity, dual promoter usage, and evidence for polymorphisms within the regulatory regions and noncoding exons. *J Biol Chem* 1997; **272:** 30662-30671.

Munkholm, P., Langholz, E., Hollander, D., Thornberg, K., Orholm, M., Katz, K.D. and Binder, V. Intestinal permeability in patients with Crohn's disease and ulcerative colitis and their first degree relatives. *Gut* 1994; **35:** 68-72.

Mwantembe, O., Gaillard, M.C., Barkhuizen, M., Pillay, V., Berry, S.D., Dewar, J.B. and Song, E. Ethnic differences in allelic associations of the interleukin-1 gene cluster in South African patients with inflammatory bowel disease (IBD) and in control individuals. *Immunogenetics* 2001; **52:** 249-254.

Naom, I., Lee, J., Ford, D., Bowman, S.J., Lanchbury, J.S., Haris, I., Hodgson, S.V., Easton, D., Lennard-Jones, J. and Mathew, C.G. Analysis of the contribution of HLA genes to genetic predisposition in inflammatory bowel disease. *Am J Hum Genet* 1996; **59:** 226-233.

Negoro, K., Kinouchi, Y., Hiwatashi, N., Takahashi, S., Takagi, S., Satoh, J., Shimosegawa, T. and Toyota, T. Crohn's disease is associated with novel polymorphisms in the 5'-flanking region of the tumour necrosis factor gene. *Gastroenterology* 1999; **117:** 1062-1068.

Nemetz, A., Kope, A., Molnar, T., Kovacs, A., Feher, J., Tulassay, Z., Nagy, F., Garcia-Gonzalez, M.A. and Pena, A.S. Significant differences in the interleukin-1beta and interleukin-1 receptor antagonist gene polymorphisms in a Hungarian population with inflammatory bowel disease. *Scand J Gastroenterol* 1999; **34:** 175-179.

O'Connell, J.R. and Weeks, D.E. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* 1998; **63:** 259-266.

Ohmen, J.D., Yang, H.Y., Yamamoto, K.K., Zhao, H.Y., Ma, Y., Bentley, L.G., Huang, Z., Gerwehr, S., Pressman, S., McElree, C., Targan, S., Rotter, J.I. and Fischel-Ghodsian, N. Susceptibility locus for inflammatory bowel disease on chromosome 16 has a role in Crohn's disease, but not in ulcerative colitis. *Hum Mol Genet* 1996; **5:** 1679-1683.

Olavesen, M.G., Hampe, J., Mirza, M.M., Saiz, R., Lewis, C.M., Bridger, S., Teare, D., Easton, D.F., Herrmann, T., Scott, G., Hirst, J., Sanderson, J., Hodgson, S.V., Lee, J., MacPherson, A., Schreiber, S., Lennard-Jones, J.E., Curran, M.E. and Mathew, C.G. Analysis of single-nucleotide polymorphisms in the interleukin-4 receptor gene for association with inflammatory bowel disease. *Immunogenetics* 2000; **51:** 1-7.

Orholm, M., Iselius, L., Sorensen, T.I., Munkholm, P., Langholz, E. and Binder, V. Investigation of inheritance of chronic inflammatory bowel diseases by complex segregation analysis. *Bmj* 1993; **306:** 20-24.

Orholm, M., Munkholm, P., Langholz, E., Nielsen, O.H., Sorensen, I.A. and Binder, V. Familial occurrence of inflammatory bowel disease. *N Engl J Med* 1991; **324:** 84-88.

Pakstis, A.J., Zhao, H., Kidd, J.R. and Kidd, K.K. Patterns of linkage disequilibrium for multisite haplotypes at 14 genetic loci in human populations worldwide. *Am J Hum Genet* 2000; **67:** 24.

Papadakis, K.A. and Targan, S.R. The role of chemokines and chemokine receptors in mucosal inflammation. *Inflamm Bowel Dis* 2000; **6:** 303-313.

Papo, M., Quer, J.C., Pastor, R.M., Garcia-Pardo, G., Prats, E., Mirapeix, E., Rodriguez, R. and Richart, C. Antineutrophil cytoplasmic antibodies in relatives of patients with inflammatory bowel disease. *Am J Gastroenterol* 1996; **91:** 1512-1515.

Pardi, D.S., Tremaine, W.J., Sandborn, W.J., Loftus, E.V., Jr., Poland, G.A., Harmsen, W.S., Zinsmeister, A.R. and Melton, L.J., 3rd Early measles virus infection is associated with the development of inflammatory bowel disease. *Am J Gastroenterol* 2000; **95:** 1480-1485.

Parkes, M., Barmada, M.M., Satsangi, J., Weeks, D.E., Jewell, D.P. and Duerr, R.H. The IBD2 locus shows linkage heterogeneity between ulcerative colitis and Crohn disease. *Am J Hum Genet* 2000; **67:** 1605-1610.

Parkes, M., Satsangi, J. and Jewell, D. Mapping susceptibility loci in inflammatory bowel disease: why and how? *Mol Med Today* 1997; **3:** 546-553.

Peeters, M., Nevens, H., Baert, F., Hiele, M., de Meyer, A.M., Vlietinck, R. and Rutgeerts, P. Familial aggregation in Crohn's disease: increased age-adjusted risk and concordance in clinical characteristics. *Gastroenterology* 1996; **111:** 597-603.

Plevy, S.E., Targan, S.R., Yang, H., Fernandez, D., Rotter, J.I. and Toyoda, H. Tumour necrosis factor microsatellites define a Crohn's disease- associated haplotype on chromosome 6. *Gastroenterology* 1996; **110:** 1053-1060.

Pokorny, R.M., Hofmeister, A., Galandiuk, S., Dietz, A.B., Cohen, N.D. and Neibergs,
H.L. Crohn's disease and ulcerative colitis are associated with the DNA repair gene
MLH1. *Ann Surg* 1997; **225:** 718-723; discussion 723-715.

Polito, J.M.d., Rees, R.C., Childs, B., Mendeloff, A.I., Harris, M.L. and Bayless, T.M.
Preliminary evidence for genetic anticipation in Crohn's disease. *Lancet* 1996; **347:**
798-800.

Probert, C.S., Jayanthi, V., Hughes, A.O., Thompson, J.R., Wicks, A.C. and Mayberry,
J.F. Prevalence and family risk of ulcerative colitis and Crohn's disease: an
epidemiological study among Europeans and south Asians in Leicestershire. *Gut* 1993;
**34:** 1547-1551.

Radhakrishnan, S., Zubaidi, G., Daniel, M., Sachdev, G.K. and Mohan, A.N. Ulcerative
colitis in Oman. A prospective study of the incidence and disease pattern from 1987 to
1994. *Digestion* 1997; **58:** 266-270.

Rector, A., Vermeire, S., Thoelen, I., Keyaerts, E., Struyf, F., Vlietinck, R., Rutgeerts,
P. and Van Ranst, M. Analysis of the CC chemokine receptor 5 (CCR5) delta-32
polymorphism in inflammatory bowel disease. *Hum Genet* 2001; **108:** 190-193.

Reif, S., Lavy, A., Keter, D., Fich, A., Eliakim, R., Halak, A., Broide, E., Niv, Y., Ron,
Y., Patz, J., Odes, S., Villa, Y. and Gilat, T. Lack of association between smoking and
Crohn's disease but the usual association with ulcerative colitis in Jewish patients in
Israel: a multicenter study. *Am J Gastroenterol* 2000; **95:** 474-478.

Reinshagen, M., Loeliger, C., Kuehnl, P., Weiss, U., Manfras, B.J., Adler, G. and

Boehm, B.O. HLA class II gene frequencies in Crohn's disease: a population based

analysis in Germany. *Gut* 1996; **38:** 538-542.


Rioux, J.D., Daly, M.J., Green, T., Stone, V., Lander, E.S., Hudson, T.J., Steinhart,

A.H., Bull, S., Cohen, Z., Greenberg, G., Griffiths, A., McLeod, R., Silverberg, M.,

Williams, C.N. and Siminovitch, K.A. Absence of linkage between inflammatory bowel

disease and selected loci on chromosomes 3, 7, 12, and 16. *Gastroenterology* 1998;

**115:** 1062-1065.


Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., McLeod, R.S., Griffiths,

A.M., Green, T., Brettin, T.S., Stone, V., Bull, S.B., Bitton, A., Williams, C.N.,

Greenberg, G.R., Cohen, Z., Lander, E.S., Hudson, T.J. and Siminovitch, K.A.

Genomewide Search in Canadian Families with Inflammatory Bowel Disease Reveals

Two Novel Susceptibility Loci. *Am J Hum Genet* 2000; **66:** 1863-1870.


Risch, N. and Merikangas, K. The future of genetic studies of complex human diseases.

*Science* 1996; **273:** 1516-1517.


Risch, N.J. Searching for genetic determinants in the new millennium. *Nature* 2000;

**405:** 847-856.


Roberts, S.B., MacLean, C.J., Neale, M.C., Eaves, L.J. and Kendler, K.S. Replication of

linkage studies of complex traits: an examination of variation in location estimates. *Am

J Hum Genet* 1999; **65:** 876-884.

Rudolph, U., Finegold, M.J., Rich, S.S., Harriman, G.R., Srinivasan, Y., Brabet, P.,

Boulay, G., Bradley, A. and Birnbaumer, L. Ulcerative colitis and adenocarcinoma of

the colon in G alpha i2-deficient mice. *Nat Genet* 1995; **10:** 143-150.


Russel, M.G., Pastoor, C.J., Janssen, K.M., van Deursen, C.T., Muris, J.W., van

Wijlick, E.H. and Stockbrugger, R.W. Familial aggregation of inflammatory bowel

disease: a population-based study in South Limburg, The Netherlands. The South

Limburg IBD Study Group. *Scand J Gastroenterol Suppl* 1997; **223:** 88-91.


Russel, M.G. and Stockbrugger, R.W. Epidemiology of inflammatory bowel disease: an

update. *Scand J Gastroenterol* 1996; **31:** 417-427.


Sandler, R.S. The incidence of IBD is higher in the North. *Inflamm Bowel Dis* 1998; **4:**

175-176.


Satsangi, J., Grootscholten, C., Holt, H. and Jewell, D.P. Clinical patterns of familial

inflammatory bowel disease. *Gut* 1996a; **38:** 738-741.


Satsangi, J., Parkes, M., Jewell, D.P. and Bell, J.I. Genetics of inflammatory bowel

disease. *Clin Sci (Colch)* 1998; **94:** 473-478.


Satsangi, J., Parkes, M., Louis, E., Hashimoto, L., Kato, N., Welsh, K., Terwilliger,

J.D., Lathrop, G.M., Bell, J.I. and Jewell, D.P. Two stage genome-wide search in

inflammatory bowel disease provides evidence for susceptibility loci on chromosomes

3, 7 and 12. *Nat Genet* 1996b; **14:** 199-202.

Satsangi, J., Rosenberg, W.M.C. and Jewell, D.P. The prevalence of inflammatory bowel disesase in relatives of patients with Crohn's disease. *European Journal of Gastroenterology & Hepatology* 1994; **6:** 413-416.

Satsangi, J., Welsh, K.I., Bunce, M., Julier, C., Farrant, J.M., Bell, J.I. and Jewell, D.P. Contribution of genes of the major histocompatibility complex to susceptibility and disease phenotype in inflammatory bowel disease. *Lancet* 1996c; **347:** 1212-1217.

Seibold, F., Slametschka, D., Gregor, M. and Weber, P. Neutrophil autoantibodies: a genetic marker in primary sclerosing cholangitis and ulcerative colitis. *Gastroenterology* 1994; **107:** 532-536.

Sorensen, H.T. and Fonager, K.M. Myocarditis and inflammatory bowel disease. A 16-year Danish nationwide cohort study. *Dan Med Bull* 1997; **44:** 442-444.

Spiegel, A.M. G protein gene knockout hits the gut. *Nat Med* 1995; **1:** 522-524.

Stokkers, P.C., Huibregtse, K., Jr., Leegwater, A.C., Reitsma, P.H., Tytgat, G.N. and van Deventer, S.J. Analysis of a positional candidate gene for inflammatory bowel disease: NRAMP2. *Inflamm Bowel Dis* 2000; **6:** 92-98.

Stokkers, P.C., van Aken, B.E., Basoski, N., Reitsma, P.H., Tytgat, G.N. and van Deventer, S.J. Five genetic markers in the interleukin 1 family in relation to inflammatory bowel disease. *Gut* 1998; **43:** 33-39.

Su, C.G., Wen, X., Bailey, S.T., Jiang, W., Rangwala, S.M., Keilbaugh, S.A., Flanigan,

A., Murthy, S., Lazar, M.A. and Wu, G.D. A novel therapy for colitis utilizing PPAR-

gamma ligands to inhibit the epithelial inflammatory response. *J Clin Invest* 1999; **104:**

383-389.


Swallow, D.M., Vinall, L.E., Gum, J.R., Kim, Y.S., Yang, H., Rotter, J.I., Mirza, M.,

Lee, J.C. and Lennard-Jones, J.E. Ulcerative colitis is not associated with differences in

MUC2 mucin allele length. *J Med Genet* 1999; **36:** 859-860.


Tagore, A., Gonsalkorale, W.M., Pravica, V., Hajeer, A.H., McMahon, R., Whorwell,

P.J., Sinnott, P.J. and Hutchinson, I.V. Interleukin-10 (IL-10) genotypes in

inflammatory bowel disease. *Tissue Antigens* 1999; **54:** 386-390.


Thiel, B.A., Tiwari, H., Fallin, D. and Schork, N.J. The effects of misspecification of

marker allele frequencies in affected sibpair linkage analysis. *Am J Hum Genet* 2000;

**67:** 314.


Thompson, N.P., Driscoll, R., Pounder, R.E. and Wakefield, A.J. Genetics versus

environment in inflammatory bowel disease: results of a British twin study. *Bmj* 1996;

**312:** 95-96.


Tountas, N.A., Casini-Raggi, V., Yang, H., Di Giovine, F.S., Vecchi, M., Kam, L.,

Melani, L., Pizarro, T.T., Rotter, J.I. and Cominelli, F. Functional and ethnic association

of allele 2 of the interleukin-1 receptor antagonist gene in ulcerative colitis.

*Gastroenterology* 1999; **117:** 806-813.

Trachtenberg, E.A., Yang, H., Hayes, E., Vinson, M., Lin, C., Targan, S.R., Tyan, D., Erlich, H. and Rotter, J.I. HLA class II haplotype associations with inflammatory bowel disease in Jewish (Ashkenazi) and non-Jewish caucasian populations. *Hum Immunol* 2000; **61:** 326-333.

Tysk, C., Lindberg, E., Jarnerot, G. and Floderus-Myrhed, B. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut* 1988; **29:** 990-996.

Uyar, F.A., Imeryuz, N., Saruhan-Direskeneli, G., Ceken, H., Ozdogan, O., Sahin, S. and Tozun, N. The distribution of HLA-DRB alleles in ulcerative colitis patients in Turkey. *Eur J Immunogenet* 1998; **25:** 293-296.

Vasiliauskas, E.A., Plevy, S.E., Landers, C.J., Binder, S.W., Ferguson, D.M., Yang, H., Rotter, J.I., Vidrich, A. and Targan, S.R. Perinuclear antineutrophil cytoplasmic antibodies in patients with Crohn's disease define a clinical subgroup. *Gastroenterology* 1996; **110:** 1810-1819.

Vermeire, S., Peeters, M., Vlietinck, R., Parkes, M., Satsangi, J., Jewell, D. and Rutgeerts, P. Exclusion of linkage of Crohn's disease to previously reported regions on chromosomes 12, 7, and 3 in the Belgian population indicates genetic heterogeneity. *Inflamm Bowel Dis* 2000; **6:** 165-170.

Wjst, M., Immervoll, T., Loesgen, S. and Bickeboeller, H. Fine mapping of asthma genes. *Am J Hum Genet* 2000; **67:** 315.

Wright, A.F., Carothers, A.D. and Pirastu, M. Population choice in mapping genes for complex diseases. *Nat Genet* 1999; **23:** 397-404.

Xu, X., Peng, M. and Fang, Z. The direction of microsatellite mutations is dependent upon allele length. *Nat Genet* 2000; **24:** 396-399.

Yang, H., McElree, C., Roth, M.P., Shanahan, F., Targan, S.R. and Rotter, J.I. Familial empirical risks for inflammatory bowel disease: differences between Jews and non-Jews. *Gut* 1993; **34:** 517-524.

Yang, H., Ohmen, J.D., Ma, Y., Targan, S.R., Fischel-Ghodsian, N. and Rotter, J.I. Additional evidence of linkage between Crohn's disease and a putative locus on chromosome 12. *Genet Med* 1999a; **1:** 194-198.

Yang, H., Plevy, S.E., Taylor, K., Tyan, D., Fischel-Ghodsian, N., McElree, C., Targan, S.R. and Rotter, J.I. Linkage of Crohn's disease to the major histocompatibility complex region is detected by multiple non-parametric analyses. *Gut* 1999b; **44:** 519-526.

Yang, P., Jarnerot, G., Danielsson, D., Tysk, C. and Lindberg, E. P-ANCA in monozygotic twins with inflammatory bowel disease. *Gut* 1995; **36:** 887-890.

Yoshitake, S., Kimura, A., Okada, M., Yao, T. and Sasazuki, T. HLA class II alleles in Japanese patients with inflammatory bowel disease. *Tissue Antigens* 1999; **53:** 350-358.

Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W. and Arnheim, N. Whole genome amplification from a single cell: implications for genetic analysis. *Proc Natl Acad Sci U S A* 1992; **89:** 5847-5851.

# Publications

# Replication and Extension Studies of Inflammatory Bowel Disease Susceptibility Regions Confirm Linkage to Chromosome 6p (*IBD3*)

Bryan Dechairo[1,2], Claire Dimon[1], David van Heel[3,4], Ian Mackay[1], Mark Edwards[1], Peter Scambler[2], Derek Jewell[3], Lon Cardon[4], Nicholas Lench[1] and Alisoun Carey[1]

[1]Oxagen Ltd., Abingdon, Oxfordshire, UK; [2]Molecular Medicine Unit, Institute of Child Health, University College of London, London, UK; [3]Gastroenterology Unit, Nuffield Department of Medicine, Radcliffe Infirmary, Oxford, UK; [4]Wellcome Trust Centre for Human Genetics, University of Oxford, UK.

Running Title: Confirmation of linkage to *IBD3*

Correspondence to:

Bryan Michael Dechairo

Oxagen Ltd.

91 Milton Park

Abingdon, OXON

OX14 4RY

United Kingdom

Tel: +44 (0) 1235 443319

Fax: +44 (0) 1235 443301

b.dechairo@oxagen.co.uk

Inflammatory Bowel Disease (IBD) is a chronic inflammatory disease of the intestine, commonly diagnosed as either ulcerative colitis (UC) or Crohn's disease (CD). Epidemiological studies have consistently shown that both genetic and environmental factors influence the pathogenesis of IBD. A number of genome scans have been conducted in cohorts of IBD families with affected sibling pairs (ASPs) to identify chromosomal regions that harbor IBD susceptibility genes. Several putative linked loci have been identified, including two loci on chromosomes 16 and 12, *IBD1* and *IBD2*, which have subsequently been replicated by independent region-specific studies. We have conducted both a replication study on another linkage region, chromosome 6p (*IBD3*), and extension studies on two other regions, chromosomes 3p and 7q, which showed suggestive linkage in the genome scan by Satsangi et al. [14]. Microsatellite markers across each region were genotyped in 284 IBD ASPs from 234 families. A nonparametric peak multipoint LOD score of 3.0 was observed near D6S291, replicating the previous linkage to chromosome 6p (*IBD3*). Nominal evidence of linkage was observed at both the 3p and 7q regions.

**Keywords**

Inflammatory bowel disease, Crohn's disease, ulcerative colitis, linkage, sib-pair,

replication, IBD1, IBD2, IBD3

# Introduction

Crohn's disease (CD) and ulcerative colitis (UC), the two most common forms of

inflammatory bowel disease (IBD), are idiopathic inflammatory diseases of the

intestine. CD can affect any part of the gastrointestinal tract and is typically

characterized by a segmental distribution of transmural inflammation, deep ulceration,

strictures, and fistulae. In contrast, UC normally onsets in the rectum (proctitis) and is

limited to the colon, exhibiting a continuous pattern of inflammation and ulceration of

only the colonic mucosa [1,2].

Considering UC and CD together, IBD is a common disease with a prevalence in

European and North American Caucasians of 126-215 per 100,000 [3-6]. Aetiology is

unknown but is predicted to involve complex interactions between genetic susceptibility

and environmental factors. Studies have shown a first-degree relative risk for IBD of 6-

15 and 6 times higher rates of concordance between monozygotic twins versus dizygotic

twins [3,5,7-9]. In most families the increase in risk is for the same type of IBD but

studies have shown that there is also an increase for the discordant phenotype. This

overlap between CD and UC has been observed at diagnosis, where up to 10% of cases

are labeled as indeterminate colitis. In these cases, review of clinical, endoscopic, and

histological data is unable to establish with certainty the disease type [6]. Overall, genetic

epidemiological studies suggest that each disease results from its own susceptibility

loci, other loci that are shared, and environmental influences.

To date, 7 genome scans in search of IBD susceptibility loci have been

published. Each study contained a different number of families and affected sib-pairs

(ASPs) and consequently each had a different power to detect linkage. The largest

genome scan for IBD susceptibility loci contained 353 ASPs [10]; other studies contained

183 ASPs [11], 151 ASPs [12], 94 ASPs [13], 89 ASPs [14], 65 ASPs [15], and 41 ASPs [16].

The proportion of CD, UC, and mixed cases in each of these studies differed. Using the

4

criteria of Lander and Kruglyak [17], these studies resulted in significant linkage for IBD

to chromosomes 6p (*IBD3* [MIM 604519]) [18], 19p [11] and 12q (*IBD2* [MIM 601458])

[14], and significant linkage for CD to chromosomes 5q31-33 [11], 14q11-12 [13], and the

pericentromeric region of chromosome 16 (*IBD1* [MIM 266600]) [16]. Only the *IBD1*

and *IBD2* loci have been replicated by independent multipoint studies in addition to

other genome scans [14,15,19-24].

Given that 1) the power to detect linkage increases with the size of the family

cohort, and that 2) our expanded cohort of 284 IBD affected sib pairs includes families

analyzed by Satsangi et al. [14], which resulted in suggestive linkage to both 3p and 7q,

we genotyped a dense set of microsatellite markers spanning these two suggestive

linkage regions in order to increase information content and potentially extend the

significance of linkage. In addition, since the *IBD3* locus has not been confirmed by

independent region-specific replication, we conducted a replication study across the 6p

region using markers from the original study [18].


## Subjects and Methods


### Families

Ethical approval was obtained from the Central Oxford Research and Ethics Committee

(COREC) for the collection of IBD families and their subsequent participation in

genetic studies. A total of 234 UK Caucasian families, none of which were Jewish,

were identified for the present study. Of the 234 families, 151 were included in a

previous two-stage genome scan (SIBS1&2) [14]. The families consisted of 107 CD, 78

UC, and 49 mixed families. Mixed families were defined as those containing any first-

or second-degree relative with a discordant form of IBD. Two of the families

comprised two generations of affected siblings, giving a total of 236 nuclear families. Of the nuclear families, 212 contained two affected siblings and 24 contained three affected siblings. In total, there were 284 IBD ASPs, consisting of 143 CD, 90 UC and 51 mixed ASPs (Table 1). Parents were collected whenever possible, resulting in 101 nuclear families with both parents, 64 with only one parent and 71 with neither parent. Of those families containing no parents, 27 included at least one unaffected sibling. Diagnosis of IBD and sub-classification as CD or UC were determined by the use of standard diagnostic criteria [25]. Venous blood samples were taken from all subjects and DNA was prepared from whole blood using either the Puregene kit (Gentra Systems) or phenol/chloroform extraction.

**Genotyping**

A total of 41, 29 and 10 highly polymorphic, fluorescent-labeled microsatellite markers spanning the putative linkage regions on chromosome 3p, 7q and 6p respectively, were genotyped across the IBD family DNA (Table 2). Using the RAPIDGene™ automated genotyping system (Oxagen Ltd.), each marker was amplified by PCR across ten 96-well plates of DNA and pooled into sets. The pooled DNA was then denatured, loaded onto a polyacrylamide gel, and electrophoresed on an ABI 377 machine. Analysis of the genotypes was conducted using ABI GENESCAN (version 3.0) and GENOTYPER (version 2.0) software. Genotype data was stored in Discovery Manager™ (Genomica Corporation) together with the pedigree and phenotype data for each individual.

To resolve marker-marker order and generate a map inclusive of all genotyped markers, radiation hybrid (RH) mapping using the Stanford G3 Panel (Research Genetics) was carried out (Table 2). For the 3p and 7q saturation regions, marker order and distance was calculated from the RH results using RHMAP (version 2.01) [26]. However, due to the increased distance between markers analyzed in the chromosome 6

6

region, the 6p map was constructed by sending the RH data to the Stanford RH Server [27]. Resulting physical distances were converted to genetic distances under the assumptions of 1 $cR_{10,000}$ = 25 kb and 1000 kb = 1 cM.

Quality control of genotypes was performed using Pedcheck (version 1.1) [28] to check for Mendel errors and CRI-MAP [29] to check for double recombinants. Genetic maps were built using MultiMap (version 2.0) [30] and compared to the RH maps to confirm genotypes and marker order.


**Statistical genetics**

Nonparametric multipoint linkage analyses across the 3p, 6p, and 7q linkage regions were performed using MAPMAKER/SIBS implemented within GENEHUNTER (version 2.0) [31]. All affected sib pairs in each pedigree were counted, using the "weighted" option. Analysis was performed under conditions that included the possibility of dominance variance. All data were analysed under three phenotype models: 1) any form of IBD designated as affected, 2) only CD patients designated as affected and 3) only UC patients designated as affected. Allele frequencies for each microsatellite marker were calculated using available genotypes from all individuals in every pedigree, both affected and unaffected. The allele frequencies were consistent with those generated by genotyping the same markers in other independent populations. The RH map generated for each region was utilized within MAPMAKER/SIBS for marker order and spacing. The information content from each marker was calculated by the GENEHUNTER program within the linkage analysis package.

TDT association analysis was performed within GENEHUNTER (version 2.0) using the TDT function. Subsequently, to test how frequent the results could be achieved by chance, 1000 randomly generated equivalent data sets were analysed using the Perm1 function.

# Results

## Genotyping

More than 70,000 genotypes were generated from the extension studies on 3p and 7q and the replication study on 6p. The data set was 99.5% complete, with 0.2% genotypes missing due to PCR failure, 0.1% removed due to unresolved PedCheck errors, 0.1% removed due to CRI-MAP identification of double recombinants, and 0.1% removed due to microsatellite mutation. In overall length and marker order, the map output from MultiMap was comparable to the RH map generated for each region. However, only 14/41 and 19/29 markers were ordered (LOD > 3.0) in the 3p and 7q genetic maps respectively, due to the lack of informative recombinants resulting from the proximity of the markers and the number of families (Figure 1). After conversion of the RH maps from cR to cM, the 41 markers on 3p, 29 markers on 7q and 10 markers on 6p spanned a region of approximately 35, 65, and 75 cM respectively (Table 2).

## Statistical Genetics

The average information content across the 3p, 7q, and 6p region was 0.91, 0.86 and 0.76 respectively. The peak multipoint LOD score for each region under each phenotype model is given in Table 3.

Positive linkage was detected in the 6p region with a non-parametric peak multipoint LOD score of 3.04 near D6S291 for IBD (Figure 1). The mean allele sharing (MAS) at the point of peak linkage was 0.58 (Table 3). There was almost equal contribution from CD and UC ASPs to the linkage, as shown by the similar $\lambda_s$ values ($0.25/z_0$) of 1.2, 1.2, and 1.3, calculated for each of the phenotype models, IBD, CD, and UC respectively. The ability to detect linkage to a locus with a $\lambda_s$ of only 1.2 is directly

related to the power of the study, and thus when the IBD cohort was divided into sub-populations of UC and CD positive linkage was not observed.

With positive linkage to the 6p region, TDT analysis using the linkage genotype data was performed across the region with all IBD phenotype models. Positive association was detected between UC and an allele from D6S271 (p = 0.007). However, 1000 simulations with random data sets of equivalent size resulted in 187 data sets of greater significance. Association was not detected with either IBD or CD.

Linkage was not detected in either the 3p or the 7q region. The largest LOD score for each region was 1.25 for CD at D3S3640-D3S1581 and 1.26 for CD at D7S645-D7S2516 (Figure 1). Similar results were seen in the re-analysis of SIBS1&2, using the newly generated genotypes (Table 3).

## Discussion

IBD genome scans have resulted in several regions of linkage. However, only two regions have been subsequently replicated by independent region-specific studies. Recently, Hampe et al. [18] reported linkage to 6p in an extension study using the largest IBD population to date (428 ASPs). In the present study, we have replicated the 6p linkage in our IBD population containing 284 ASPs. The nonparametric peak multipoint LOD score of 3.04 near D6S291 exceeds the proposed criteria for independent replication of significant linkage (LOD = 1.5) [17]. However, while this represents the first region-specific replication study of the *IBD3* locus, both a recent genome scan by Rioux et al. [11] and an earlier candidate gene study by Yang et al. [32] also reported suggestive linkage to 6p between D6S1281 and D6S1019 in a cohort of 183 IBD ASPs (LOD = 2.3) and around TNF in a cohort of 70 CD ASPs (P = 0.002), respectively. All three IBD studies showed linkage contributed by both the UC and the CD populations. However, both the linkage peak from the Rioux et al. [11] study and the

peak from our study were approximately 15 cM proximal to the initial reports of linkage [18,32]. This discrepancy is not surprising since differences between actual gene location and peak linkage are not unexpected in complex traits [33]. Taken together the results support the existence of an IBD susceptibility locus at 6p (*IBD3*).

Confirmation of the linkage to 6p now warrants high-density association studies across the region, potentially starting with good positional candidate genes. The *IBD3* locus contains the MHC region, which has been implicated in other linkage studies of autoimmune diseases including inflammatory diseases such as asthma, arthritis and psoriasis [34]. This overlap of susceptibility regions suggests that a MHC autoimmune susceptibility gene may be responsible for the positive linkage results. In previous studies, DRB1, TNFα, and other MHC genes have already shown positive association to IBD [35-50]. However, another hypothesis is that the overlap of disease susceptibility loci could indicate an inflammation specific disease gene in the region. One potential inflammation candidate gene is cytokine suppressive anti-inflammatory binding protein 1 (CSBP1). CSBP1 RH maps directly adjacent to D6S291, the peak linkage marker in this study. Still, even with multiple overlapping disease linkage regions, the potential of finding an IBD specific disease gene in the region should not be ignored. Recently, peroxisome proliferator activated receptor gamma (PPARγ) was implicated in IBD [51], which suggests that a family member like PPARδ, which again maps adjacent to D6S291, could also be implicated. Not until all IBD, inflammation and autoimmune candidate genes within the region have been thoroughly tested for association to IBD may it be possible to understand all susceptibility conferred by the *IBD3* locus.

To better stratify the numerous IBD linkage regions by their effect on the disease population, it is important that extension studies are conducted on genome scan linkage regions when the size of the ASP cohort increases. With the exception of the IBD linkage regions on 14q and 19p, all other significant (LOD > 3.59) IBD linkages have

10

resulted from studies across suggestive loci in expanded cohorts [11,13,14,16,18]. In the

present study, we revisited the suggestive loci resulting from the Satsangi et al. [14]

genome scan on 3p and 7q by conducting extension studies in each region with an

expanded population of 284 IBD ASPs. Linkage was not detected in either region for

any phenotype model. Due to the difference in results from this extension study and the

previous genome scan by Satsangi et al. [14], the 151 families included in the previous

study were analysed again separately. Since, the additional families used in the present

study were ascertained using the same criteria as used previously (re: clinical profile,

ethnicity, family size), the differences in results may either be caused by genetic

heterogeneity or by difference in the methodology of analysis. The analysis showed

that the present SIBS1&2 data for 3p and 7q, with respective MAS of 0.55 and 0.55,

was consistent with the extension results rather than the previously published genome

scan results (MAS of 0.63 and 0.64, respectively). Thus, the differences between the

two studies are more likely to be caused by lower information content, due to the

previous use of two-point analysis rather than multipoint analysis.

Since the publication of the genome scan by Satsangi et al. [14], other independent

region-specific studies in these regions have been conducted [23,24,52,53]. These studies

also failed to see linkage. To date, 6 genome scans and 4 replication studies have failed

to detect linkage to chromosome 7. It has been noted that suggestive linkage will occur

by chance once per genome scan [17] and thus the linkage to 7q detected in our previous

genome scan may be a false positive. Conversely, a recent genome scan by Rioux et al.

[11] has reported linkage to the chromosome 3 region, with a LOD of 2.4, and thus the 3p

interval may contain a gene of small effect on IBD aetiology.

In summary, we have replicated linkage to the 6p region and have failed to

extend linkage to the 3p and 7q regions. These replication and extension studies,

together with others like them, will help in the stratification of linkage regions by their relative impact on IBD susceptibility.

## Acknowledgments

## References

1    Sorensen HT, Fonager KM. Myocarditis and inflammatory bowel disease. A 16-year Danish nationwide cohort study. *Dan Med Bull* 1997; **44**: 442-444.

2    McConnell RB, Vadheim CM. Inflammatory Bowel Disease; in King RA, Rotter JI, Motulsky AO (eds.): The Genetics Basis of Common Diseases. Oxford, 1992, pp 326-348.

3    Cipolla C, Magliocco A, Oliva L, Cottone M. Familial aggregation of inflammatory bowel disease in a Mediterranean area. *Eur J Epidemiol* 1996; **12**: 205-210.

4    Probert CS, Jayanthi V, Hughes AO, Thompson JR, Wicks AC, Mayberry JF. Prevalence and family risk of ulcerative colitis and Crohn's disease: an epidemiological study among Europeans and south Asians in Leicestershire. *Gut* 1993; **34**: 1547-1551.

5       Orholm M, Munkholm P, Langholz E, Nielsen OH, Sorensen IA, Binder V.

        Familial occurrence of inflammatory bowel disease. *N Engl J Med* 1991; **324**:

        84-88.

6       Satsangi J, Parkes M, Jewell DP, Bell JI. Genetics of inflammatory bowel

        disease. *Clin Sci (Colch)* 1998; **94**: 473-478.

7       Russel MG, Pastoor CJ, Janssen KM *et al.* Familial aggregation of inflammatory

        bowel disease: a population-based study in South Limburg, The Netherlands.

        The South Limburg IBD Study Group. *Scand J Gastroenterol Suppl* 1997;**223**:

        88-91.

8       Tysk C, Lindberg E, Jarnerot G, Floderus-Myrhed B. Ulcerative colitis and

        Crohn's disease in an unselected population of monozygotic and dizygotic twins.

        A study of heritability and the influence of smoking. *Gut* 1988; **29**: 990-996.

9       Thompson NP, Driscoll R, Pounder RE, Wakefield AJ. Genetics versus

        environment in inflammatory bowel disease: results of a British twin study. *Bmj*

        1996; **312**: 95-96.

10      Hampe J, Schreiber S, Shaw SH *et al.* A genomewide analysis provides

        evidence for novel linkages in inflammatory bowel disease in a large European

        cohort. *Am J Hum Genet* 1999; **64**: 808-816.

11      Rioux JD, Silverberg MS, Daly MJ *et al.* Genomewide Search in Canadian

        Families with Inflammatory Bowel Disease Reveals Two Novel Susceptibility

        Loci. *Am J Hum Genet* 2000; **66**: 1863-1870.

12      Cho JH, Nicolae DL, Gold LH *et al.* Identification of novel susceptibility loci

        for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for

        epistasis between 1p and IBD1. *Proc Natl Acad Sci U S A* 1998; **95**: 7502-7507.

13    Duerr RH, Barmada MM, Zhang L, Pfutzer R, Weeks DE. High-Density Genome Scan in Crohn Disease Shows Confirmed Linkage to Chromosome 14q11-12. *Am J Hum Genet* 2000; **66**: 1857-1862.

14    Satsangi J, Parkes M, Louis E *et al.* Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. *Nat Genet* 1996; **14**: 199-202.

15    Ma Y, Ohmen JD, Li Z *et al.* A genome-wide search identifies potential new susceptibility loci for Crohn's disease. *Inflamm Bowel Dis* 1999; **5**: 271-278.

16    Hugot JP, Laurent-Puig P, Gower-Rousseau C *et al.* Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 1996; **379**: 821-823.

17    Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11**: 241-247.

18    Hampe J, Shaw SH, Saiz R *et al.* Linkage of inflammatory bowel disease to human chromosome 6p. *Am J Hum Genet* 1999; **65**: 1647-1655.

19    Duerr RH, Barmada MM, Zhang L *et al.* Linkage and association between inflammatory bowel disease and a locus on chromosome 12. *Am J Hum Genet* 1998; **63**: 95-100.

20    Curran ME, Lau KF, Hampe J *et al.* Genetic analysis of inflammatory bowel disease in a large European cohort supports linkage to chromosomes 12 and 16. *Gastroenterology* 1998; **115**: 1066-1071.

21    Cavanaugh JA, Callen DF, Wilson SR *et al.* Analysis of Australian Crohn's disease pedigrees refines the localization for susceptibility to inflammatory bowel disease on chromosome 16. *Ann Hum Genet* 1998; **62**: 291-298.

22    Ohmen JD, Yang HY, Yamamoto KK *et al.* Susceptibility locus for inflammatory bowel disease on chromosome 16 has a role in Crohn's disease, but not in ulcerative colitis. *Hum Mol Genet* 1996; **5**: 1679-1683.

23    Annese V, Latiano A, Bovio P *et al.* Genetic analysis in Italian families with inflammatory bowel disease supports linkage to the IBD1 locus--a GISC study. *Eur J Hum Genet* 1999; **7**: 567-573.

24    Brant SR, Fu Y, Fields CT *et al.* American families with Crohn's disease have strong evidence for linkage to chromosome 16 but not chromosome 12. *Gastroenterology* 1998; **115**: 1056-1061.

25    Lennard-Jones JE. Classification of inflammatory bowel disease. *Scand J Gastroenterol Suppl* 1989; **170**: 2-6.

26    Boehnke M, Lange K, Cox DR. Statistical methods for multipoint radiation hybrid mapping. *Am J Hum Genet* 1991; **49**: 1174-1188.

27    Carey AH. Use of commercially available radiation hybrid panels; in Dracopoli NC, Haines JL, Krof BR *et al.* (eds.): Current Protocols in Human Genetics. New York, 1997, vol 1, pp 3.5.1-3.5.10.

28    O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* 1998; **63**: 259-266.

29    Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* 1987; **84**: 2363-2367.

30    Matise TC, Perlin M, Chakravarti A. Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. *Nat Genet* 1994; **6**: 384-390.

31    Daly MJ, Kruglyak L, Pratt S *et al.* GENEHUNTER 2.0 - a complete linkage analysis system. *Am J Hum Genet Suppl* 1998; **63**: A286.

32    Yang H, Plevy SE, Taylor K *et al.* Linkage of Crohn's disease to the major histocompatibility complex region is detected by multiple non-parametric analyses. *Gut* 1999; **44**: 519-526.

33      Kruglyak L, Lander ES. Limits on fine mapping of complex traits [letter]. *Am J Hum Genet* 1996; **58**: 1092-1093.

34      Becker KG, Simon RM, Bailey-Wilson JE *et al.* Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases. *Proc Natl Acad Sci U S A* 1998; **95**: 9979-9984.

35      Bouma G, Xia B, Crusius JB *et al.* Distribution of four polymorphisms in the tumour necrosis factor (TNF) genes in patients with inflammatory bowel disease (IBD). *Clin Exp Immunol* 1996; **103**: 391-396.

36      Bouma G, Oudkerk Pool M, Crusius JB *et al.* Evidence for genetic heterogeneity in inflammatory bowel disease (IBD); HLA genes in the predisposition to suffer from ulcerative colitis (UC) and Crohn's disease (CD). *Clin Exp Immunol* **109**, 175-179 (1997).

37      Bouma G, Poen AC, Garcia-Gonzalez MA *et al.* HLA-DRB1*03, but not the TNFA -308 promoter gene polymorphism, confers protection against fistulising Crohn's disease. *Immunogenetics* 1998; **47**: 451-455.

38      De La Concha EG, Fernandez-Arquero M, Santa-Cruz S *et al.* Positive and negative associations of distinct HLA-DR2 subtypes with ulcerative colitis (UC). *Clin Exp Immunol* 1997; **108**: 392-395.

39      Danze PM, Colombel JF, Jacquot S *et al.* Association of HLA class II genes with susceptibility to Crohn's disease. *Gut* 1996; **39**: 69-72.

40      Forcione DG, Sands B, Isselbacher KJ, Rustgi A, Podolsky DK,Pillai S. An increased risk of Crohn's disease in individuals who inherit the HLA class II DRB3*0301 allele. *Proc Natl Acad Sci U S A* 1996; **93**: 5094-5098.

41      Fernandez Arquero M, Lopez Nava G, De la Concha EG *et al.* HLA-DR2 gene and Spanish patients with ulcerative colitis. *Rev Esp Enferm Dig* 1998; **90**: 243-249.

42    Heresbach D, Alizadeh M, Reumaux D *et al.* Are HLA-DR or TAP genes

genetic markers of severity in ulcerative colitis? *J Autoimmun* 1996; **9**: 777-784.

43    Louis E, Satsangi J, Roussomoustakaki M *et al.* Cytokine gene polymorphisms

in inflammatory bowel disease. *Gut* 1996; **39**: 705-710.

44    Negoro K, Kinouchi Y, Hiwatashi N *et al.* Crohn's disease is associated with

novel polymorphisms in the 5'-flanking region of the tumor necrosis factor gene.

*Gastroenterology* 1999; **117**: 1062-1068.

45    Plevy SE, Targan SR, Yang H, Fernandez D, Rotter JI,Toyoda H. Tumor

necrosis factor microsatellites define a Crohn's disease- associated haplotype on

chromosome 6. *Gastroenterology* 1996; **110**: 1053-1060.

46    Reinshagen M, Loeliger C, Kuehnl P *et al.* HLA class II gene frequencies in

Crohn's disease: a population based analysis in Germany. *Gut* 1996; **38**: 538-

542.

47    Satsangi J, Welsh KI, Bunce M *et al.* Contribution of genes of the major

histocompatibility complex to susceptibility and disease phenotype in

inflammatory bowel disease. *Lancet* 1996; **347**: 1212-1217.

48    Trachtenberg EA, Yang H, Hayes E *et al.* HLA class II haplotype associations

with inflammatory bowel disease in Jewish (Ashkenazi) and non-Jewish

caucasian populations. *Hum Immunol* 2000; **61**: 326-333.

49    Uyar FA, Imeryuz N, Saruhan-Direskeneli G *et al.* The distribution of HLA-

DRB alleles in ulcerative colitis patients in Turkey. *Eur J Immunogenet* 1998;

**25**: 293-296.

50    Yoshitake S, Kimura A, Okada M, Yao T,Sasazuki T. HLA class II alleles in

Japanese patients with inflammatory bowel disease. *Tissue Antigens* 1999; **53**:

350-358.

51      Su CG, Wen X, Bailey ST *et al.* A novel therapy for colitis utilizing PPAR-

        gamma ligands to inhibit the epithelial inflammatory response. *J Clin Invest*

        1999; **104**: 383-389.

52      Rioux JD, Daly MJ, Green T *et al.* Absence of linkage between inflammatory

        bowel disease and selected loci on chromosomes 3, 7, 12, and 16.

        *Gastroenterology* 1998; **115**: 1062-1065.

53      Vermeire S, Peeters M, Vlietinck R *et al.* Exclusion of linkage of Crohn's

        disease to previously reported regions on chromosomes 12, 7, and 3 in the

        Belgian population indicates genetic heterogeneity. *Inflamm Bowel Dis* 2000; **6**:

        165-170.

**Table 1**

**Summary of IBD Cohort Used for Linkage Analysis**

| Family Type/ Affected Sibs | CD (ASPs) | UC (ASPs) | Mixed (ASPs) | IBD (ASPs) |
|---|---|---|---|---|
| CD/2 | 94 (94) | - | - | 94 (94) |
| CD/3 | 13 (39) | - | - | 13 (39) |
| UC/2 | - | 72 (72) | - | 72 (72) |
| UC/3 | - | 5 (15) | - | 5 (15) |
| UC/2G-2 | - | 1 (2) | - | 1 (2) |
| Mixed/2 | 2 (2) | - | 40 (40) | 42 (42) |
| Mixed/3 | 5 (7) | 1 (1) | 5 (10) | 6 (18) |
| Mixed/2G-2 | 1 (1) | - | 1 (1) | 1(2) |
| Total | 115 (143) | 79 (90) | 46 (51) | 234 (284) |

Note - "2G-2" denotes an extended family with two generations of ASPs. Mixed family type refers to any family with a least one first- or second-degree relative with a discordant form of IBD. Mixed ASPs denotes any sibling pair with discordant forms of IBD.

**Table 2**

**Genotyped Markers and Corresponding Inter-Marker Distance and Order**

| Chr. 3 Markers | RH $cR_{10,000}$ | cM |
|---|---|---|
| D3S3727 | 0 | 0 |
| D3S3518 | 63.1 | 1.58 |
| D3S1277 | 39.9 | 1 |
| D3S1768 | 16.4 | 0.41 |
| D3S1619 | 66.3 | 1.66 |
| D3S1298 | 192.6 | 4.82 |
| D3S3521 | 16.3 | 0.41 |
| D3S3527 | 5.3 | 0.13 |
| D3S3685 | 66.4 | 1.66 |
| D3S3559 | 19.6 | 0.49 |
| D3S3678 | 30.9 | 0.77 |
| D3S3647 | 4.8 | 0.12 |
| D3S1767 | 89.9 | 2.25 |
| D3S3640 | 49.6 | 1.24 |
| D3S2420 | 14.4 | 0.36 |
| D3S1581 | 31.5 | 0.79 |
| D3S2384 | 45 | 1.13 |
| D3S3629 | 26.1 | 0.65 |
| D3S1568 | 45.5 | 1.14 |
| D3S1621 | 15.4 | 0.39 |
| D3S1573 | 15.4 | 0.39 |
| D3S1578 | 78.3 | 1.96 |
| D3S1588 | 13.2 | 0.33 |
| D3S3672 | 17.2 | 0.43 |
| D3S1582 | 32.3 | 0.81 |
| D3S1613 | 29.1 | 0.73 |
| D3S3660 | 0 | 0.01 |
| D3S3719 | 20.5 | 0.51 |
| D3S1606 | 7.4 | 0.19 |
| D3S3621 | 33.5 | 0.84 |
| D3S3616 | 32.5 | 0.81 |
| D3S1295 | 20.1 | 0.5 |
| D3S1766 | 22.7 | 0.57 |
| D3S2452 | 8.3 | 0.21 |
| D3S3577 | 22 | 0.55 |
| D3S1547 | 13.4 | 0.34 |
| D3S1234 | 4.4 | 0.11 |
| D3S1600 | 122.9 | 3.07 |
| D3S3635 | 22.9 | 0.57 |
| D3S3571 | 0 | 0.01 |
| D3S1285 | 37.7 | 0.94 |

| Chr. 6 Markers | RH $cR_{10,000}$ | cM |
|---|---|---|
| D6S344 | 0 | 0 |
| D6S309 | 172 | 4.3 |
| D6S470 | 148 | 3.7 |
| D6S289 | 456 | 11.4 |
| D6S276 | 296 | 7.4 |
| D6S265 | 312 | 7.8 |
| D6S291 | 548 | 13.7 |
| D6S426 | 252 | 6.3 |
| D6S271 | 136 | 3.4 |
| D6S257 | 540 | 13.5 |

| Chr. 7 Markers | RH $cR_{10,000}$ | cM |
|---|---|---|
| D7S2548 | 0 | 0 |
| D7S2428 | 49.2 | 1.23 |
| D7S2427 | 48 | 1.2 |
| D7S478 | 4 | 0.1 |
| D7S519 | 78 | 1.95 |
| D7S2506 | 63.2 | 1.58 |
| D7S670 | 5.2 | 0.13 |
| D7S679 | 9.2 | 0.23 |
| D7S2451 | 79.2 | 1.98 |
| D7S2422 | 100 | 2.5 |
| D7S2552 | 158 | 3.95 |
| EGFR | 12 | 0.3 |
| D7S499 | 56 | 1.4 |
| D7S494 | 30 | 0.75 |
| D7S2429 | 197.2 | 4.93 |
| D7S2512 | 137.2 | 3.43 |
| D7S502 | 41.2 | 1.03 |
| D7S2503 | 8 | 0.2 |
| D7S645 | 76 | 1.9 |
| D7S2516 | 78 | 1.95 |
| D7S653 | 238 | 5.95 |
| D7S1870 | 36 | 0.9 |
| D7S672 | 54 | 1.35 |
| ELN | 74 | 1.85 |
| D7S2455 | 182 | 4.55 |
| D7S2540 | 177.2 | 4.43 |
| D7S524 | 170 | 4.25 |
| D7S2410 | 108.4 | 2.71 |
| D7S491 | 440.4 | 11.01 |

Note – Markers appear in order from p to q arm. 1 $cR_{10,000}$ is approximately equivalent to 25 kb. The genetic distances were derived under the assumption of 1000 kb = 1 cM.

# Table 3

## Peak Multipoint Nonparametric Linkage Results

| Chromosomal Region / Phenotype Model | Marker(s) extended cohort | LOD Score extended cohort | MAS extended cohot | Marker(s) Sibs 1&2 | LOD Score Sibs 1&2 | Mas Sibs 1&2 |
|---|---|---|---|---|---|---|
| 6p/IBD | D6S291 | 3.04 | 0.58 | - | - | - |
| 6p/CD | D6S291 | 1.23 | 0.57 | - | - | - |
| 6p/UC | D6S426 | 0.89 | 0.58 | - | - | - |
| 3p/IBD | D3S1588-D3S3672 | 0.68 | 0.54 | D3S3672-D3S1582 | 1.19 | 0.55 |
| 3p/CD | D3S3640-D3S1581 | 1.25 | 0.58 | D3S1581-D3S2384 | 1.21 | 0.59 |
| 3p/UC | D3S1588-D3S3672 | 0.09 | 0.53 | D3S3672-D3S1582 | 0.74 | 0.57 |
| 7q/IBD | D7S2505-D7S670 | 1.15 | | D7S2427 | 1.12 | 0.55 |
| 7q/CD | D7S645-D7S2516 | 1.26 | 0.58 | D7S1870-D7S672 | 1.05 | 0.59 |
| 7q/UC | D7S2428 | 1.06 | | D7S2428 | 0.83 | 0.57 |

Figure 1

Chromosome 6
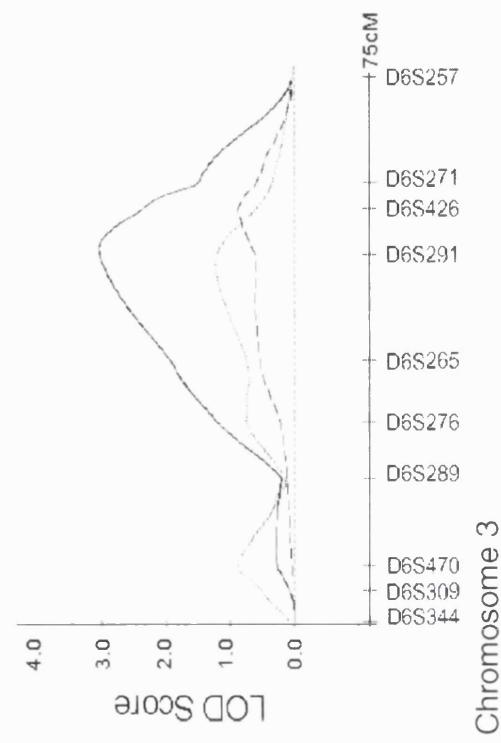
Chromosome 3
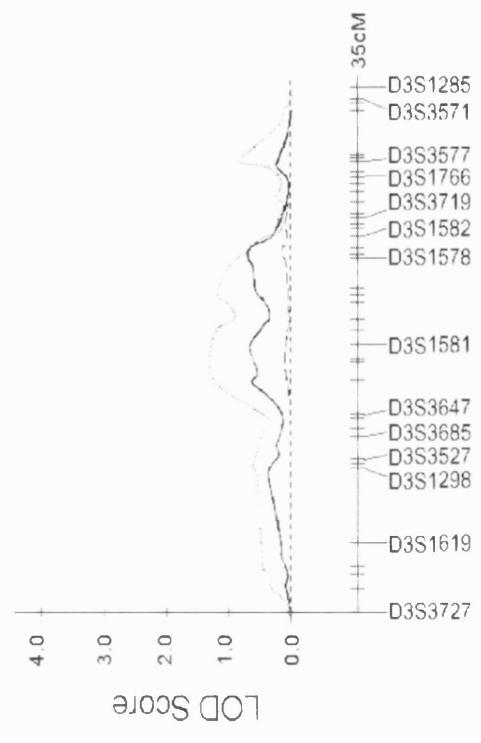
Chromosome 7
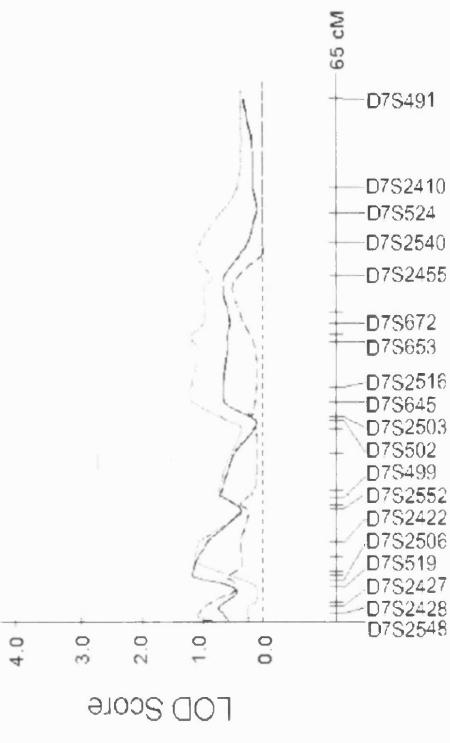
**Figure 1**    Nonparametric multipoint MLS curves for the chromosome 6, 3 and 7

IBD linkage regions.  Results for the IBD (bold solid line), CD (thin solid line), and UC

(dashed line) phenotype models are shown.  Marker-marker order and distance was

calculated from RH mapping results (Table 2).  The markers shown for the

chromosomes 3 and 7 regions correspond to the markers resulting from the MultiMap

output for each interval.