LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

**Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin**

**Copyright statement**

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

**Liability statement**

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

**Access Agreement**

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

TRINITY COLLEGE
P 2 1 NOV 2007
LIBRARY DUBLIN
THOS D
8211

# Genetic Association Study of Coeliac Disease in the Irish Population

A thesis submitted to the University of Dublin, Trinity College

*for the degree of*

Doctor of Philosophy (Ph.D.)

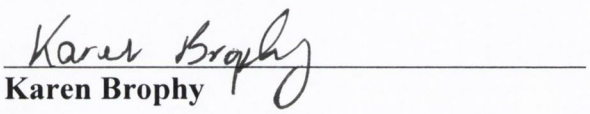
*by*

Karen Brophy, B.A. (Mod. Human Genetics)

Supervisor
Ross McManus, Ph.D.
Dept. Clinical Medicine
University of Dublin
Trinity College
*March, 2006*

**DECLARATION**

I hereby certify that the work in this thesis has not been submitted for any other degree or diploma at this, or any other university, and that all the work described herein is entirely my own except where otherwise acknowledged. The library may lend or copy this thesis upon request.



**Karen Brophy**

TABLE OF CONTENTS

*GENERAL INTRODUCTION* ..... *1*

1.1 Historical background of coeliac disease ..... 2

1.2 Epidemiology ..... 2

1.3 The role of gluten in coeliac disease ..... 4

1.4 Pathology ..... 5

1.5 Clinical Presentation ..... 7

1.6 Diagnosis ..... 8

1.7 Coeliac disease as a genetic disorder ..... 10

1.8 Human leukocyte antigen (HLA) complex and coeliac disease ..... 11

1.9 Non-HLA linked genes and coeliac disease ..... 14

1.10 Genomewide linkage studies in coeliac disease ..... 15

1.11 Candidate gene studies in coeliac disease ..... 20

1.12 Association studies - Identification of candidate genes ..... 22

1.13 The case-control association study design ..... 24

1.14 Factors influencing genetic association studies ..... 26

1.15 Factors influencing genetic association studies - Odds Ratio (OR) of the disease variant ..... 27

1.16 Factors influencing genetic association studies – Linkage disequilibrium ..... 27

1.17 Factors influencing genetic association studies – Disease and marker allele frequencies ..... 30

1.18 Influence of factors on the power of single marker tests in association studies . 31

1.19 Genotyping accuracy ..... 34

1.20 Haplotypes and association studies - Advantages of haplotype analysis ..... 34

1.21 Haplotype structure in the human genome ..... 36

1.22 Haplotype blocks and association mapping ..... 37

1.23 Haplotype tagging ..... 39

1.24 Use of public resources for association studies ..... 41

**1.25 Haplotype reconstruction** ..... **42**

**1.26 Rational and objectives of this study** ..... **44**

***GENERAL MATERIALS AND METHODS*** ..... ***46***

**2.1 Coeliac disease patient classification and control collection** ..... **47**

**2.2 DNA extraction from blood samples** ..... **47**

**2.3 DNA quality control and quantification** ..... **47**

**2.4 Power calculations** ..... **47**

**2.5 HLA-DRB1, HLA-DQB1 genotyping in coeliac disease and patient samples** ..... **48**

**2.6 Identification of candidate genes for analysis** ..... **48**

**2.7 Identification of genetic polymorphisms for analysis in candidate genes in regions 2q33 and 11q23.3** ..... **50**

2.7.1 SNP selection criteria ..... 50

2.7.2 Haplotype-tagging SNP identification ..... 50

2.7.3 Transcription factor binding site altering SNPs ..... 53

2.7.4 Non-synonymous SNP selection ..... 53

**2.8 Identification of candidate polymorphisms in MHC class III region** ..... **55**

**2.9 Analysis of the accuracy of haplotype reconstruction from haplotype-tagging SNPs** ..... **57**

**2.10 Analysis of the ability of haplotype-tagging SNPs to capture previously unidentified genetic variations on tagged haplotypes** ..... **58**

**2.11 Genotyping of polymorphisms** ..... **60**

2.11.1 Polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) analysis ..... 60

2.11.2 Restriction fragment length polymorphism (RFLP) analysis ..... 60

2.11.3 Agarose gel electrophoresis ..... 61

2.11.4 Allelic discrimination using Taqman technology ..... 61

2.11.5 Genotyping by Amplifluor ..... 63

2.11.6 Confirmation of Ampliflour genotyping results ..... 66

**2.12 Data analysis** ..... **66**

2.12.1 Genetic case-control differentiation analysis of allele, genotype and carrier status frequencies ..... 66

2.12.2 Hardy-Weinberg Equilibrium (HWE) testing ..... 66

2.12.3 Haplotype differentiation analysis ..... 66

2.12.4 Inference of allele frequencies of non-genotyped SNPs ..... 67

2.12.5 Gene-gene interaction effects ..... 68
2.12.6 Linkage disequilibrium analysis ..... 69
2.12.7 Statistical adjustment for multiple testing ..... 69

*GENETIC ASSOCIATION ANALYSIS OF THE 2q33 REGION* ..... *70*

**3.1 Introduction** ..... **71**

**3.2 Objectives** ..... **72**

**3.3 Results** ..... **73**

3.3.1 Gene selection on chromosomal region 2q33 ..... 73
3.3.2 SNP selection results ..... 76
3.3.2.1 CTLA4 ..... 76
3.3.2.2 ICOS ..... 77
3.3.2.3 CD28 ..... 78
3.3.3 Haplotype-tagging SNP (htSNP) selection and accuracy testing. ..... 78
3.3.3.1 CTLA4 ..... 78
3.3.3.2 ICOS ..... 81
3.3.4 Power calculations ..... 82
3.3.5 Hardy-Weinberg Equilibrium (HWE) Testing ..... 83
3.3.6 Allele, Genotype and Carrier Status frequencies and disease association testing in genotyped polymorphisms ..... 84
3.3.7 Linkage Disequilibrium Analysis ..... 91
3.3.8 Haplotype Analysis of the 2q33 region ..... 93
3.3.9 Inference of non-genotyped SNP frequencies using observed haplotype frequencies ..... 97
3.3.10 Evolutionary Conserved Regions Analysis ..... 97

**3.4 Discussion** ..... **99**

*GENETIC ASSOCIATION ANALYSIS OF THE 11q23 REGION AND THE TNFSF14 GENE* ..... *104*

**4.1 Introduction** ..... **105**

4.1.1 Coeliac disease and chromosomal region 11q23 ..... 105
4.1.2 Candidate genes in the 11q23 region ..... 105
4.1.3 LIGHT (TNFSF14) ..... 107

**4.2 Objectives** ..... **108**

**4.3 Results** ..... **109**

4.3.1 Gene selection on chromosomal region 11q23.3 ..... 109
4.3.2 SNP selection results ..... 110
4.3.2.1 CD3E ..... 110
4.3.2.2 CD3D and CD3G ..... 111

4.3.2.3 THY1 ..... 112
4.3.2.4 IL10RA ..... 112
4.3.2.5 IL18 ..... 113
4.3.2.6 LIGHT (TNFSF14) ..... 114
4.3.3 Haplotype-tagging SNP (htSNP) selection and accuracy testing ..... 115
4.3.4 Hardy-Weinberg Equilibrium (HWE) Testing ..... 117
4.3.5 Allele, Genotype and Carrier Status frequencies and disease association testing of genotyped polymorphisms in chromosome 11 candidate genes ..... 118
4.3.6 Allele, Genotype and Carrier Status frequencies and disease association testing in genotyped polymorphisms in *TNFSF14* ..... 118
4.3.7 Linkage Disequilibrium Analysis ..... 128
4.3.8 Haplotype analysis of the candidate genes on chromosome 11 ..... 130
4.3.9 Haplotype analysis of the *TNFSF14* gene ..... 132
4.3.10 Inference of non-genotyped SNP frequencies using observed haplotype frequencies ... 137

**4.4 Gene-gene interaction effects in coeliac disease susceptibility genes ..... 139**

**4.5 Discussion ..... 147**

***GENETIC ASSOCIATION ANALYSIS OF THE MHC CLASS III REGION... 151***

**5.1 Introduction ..... 152**
5.1.1 Major Histocompatibility Complex Structure ..... 152
5.1.2 HLA association with disease ..... 152
5.1.3 Additional MHC autoimmune susceptibility loci ..... 153
5.1.4 Linkage Disequilibrium patterns and haplotypes in the MHC ..... 155
5.1.5 Rational of this study ..... 156

**5.2 Objectives ..... 157**

**5.3 Results ..... 159**
5.3.1 HLA-DRB1, HLA-DQB1 typing ..... 159
5.3.2 SNP selection results ..... 161
5.3.2.1 TNF block ..... 161
5.3.2.2 SNP variation between the AH 8.1 and AH 7.1 haplotypes ..... 162
5.3.3 Hardy-Weinberg Equilibrium (HWE) Testing ..... 165
5.3.4 Allele, genotype and carrier status frequencies and disease association testing in genotyped polymorphisms in MHC class III in DQ2 heterozygous individuals ..... 166
5.3.5 Haplotype reconstruction accuracy and testing ..... 177
5.3.6 Linkage disequilibrium analysis ..... 180
5.3.7 Analysis of SNP frequency differentials in high risk haplotypes ..... 188
5.3.8 Analysis of low risk (non-DQB1*0201) haplotypes ..... 196

**5.4 Discussion ..... 207**

*GENERAL DISCUSSION* ..... *214*

**6.1 Coeliac disease and chromosomal region 2q33** ..... **215**

**6.2 Coeliac disease and chromosomal region 11q23** ..... **217**

**6.3 Coeliac disease and the *TNFSF14* gene** ..... **218**

**6.4 Gene-gene interaction effects in coeliac disease susceptibility genes** ..... **219**

**6.5 Coeliac disease and the MHC class III region/*MIC* genes** ..... **220**

**6.6 Future Prospects** ..... **223**

**Appendix I** ..... **226**

**Appendix II** ..... **227**

**Appendix III** ..... **231**

**Appendix IV** ..... **233**

**Appendix V** ..... **238**

*REFERENCES* ..... *240*

## LIST OF TABLES

TABLE 1.1 COELIAC DISEASE SUSCEPTIBILITY LOCI IDENTIFIED BY GENOME-WIDE LINKAGE STUDIES (NON-HLA REGIONS) ..... 19

TABLE 2.1 GENOTYPE REFERENCE INFORMATION SOURCES FOR CHROMOSOME 11 GENES AND *TNFSF14* ..... 52

TABLE 2.2 PCR CYCLING CONDITIONS AND RESTRICTION ENZYME USED IN DIGESTION FOR RFLP GENOTYPE ASSAYS ..... 65

TABLE 3.1 GENES LOCATED 100 KB FROM THE *CD28/CTLA4/ICOS* LOCUS AS IDENTIFIED BY HUMAN CONTIGVIEW IN ENSEMBL (V32) ..... 75

TABLE 3.2 COMPARISON OF HAPLOTYPE FREQUENCIES BETWEEN JOHNSON *ET AL* STUDY AND PRESENT CONTROL POPULATION ..... 81

TABLE 3.3 COMPARISON OF HAPLOTYPE FREQUENCIES BETWEEN HAANING ANDERSEN *ET AL* STUDY AND PRESENT CONTROL POPULATION ..... 82

TABLE 3.4 POWER CALCULATIONS OF VARIOUS GENOTYPE RELATIVE RISKS IN THE PRESENT STUDY ..... 83

TABLE 3.5 HARDY WEINBERG EQUILIBRIUM ANALYSIS RESULTS FOR ALL SNPS IN 2Q33 ..... 84

TABLE 3.6 GENOTYPE AND ALLELE FREQUENCIES IN THE COELIAC DISEASE PATIENT AND CONTROL POPULATIONS ..... 87

TABLE 3.7 CARRIER STATUS FREQUENCIES IN THE COELIAC DISEASE PATIENT AND CONTROL POPULATIONS ..... 89

TABLE 3.8 LD BLOCK HAPLOTYPE ASSOCIATION TESTING AND FREQUENCY ESTIMATIONS (>5%) FOR PATIENT AND CONTROL POPULATIONS ..... 94

TABLE 3.9 *CTLA4* HAPLOTYPE ASSOCIATION TESTING AND FREQUENCY ESTIMATIONS (>5%) FOR PATIENT AND CONTROL POPULATIONS ..... 94

TABLE 3.10 HT*CTLA4* HAPLOTYPE ASSOCIATION TESTING AND FREQUENCY ESTIMATIONS (>5%) FOR PATIENT AND CONTROL POPULATIONS ..... 94

TABLE 3.11 PROMOTER *CTLA4* HAPLOTYPE ASSOCIATION TESTING AND FREQUENCY ESTIMATIONS (>5%) FOR PATIENT AND CONTROL POPULATIONS ..... 95

TABLE 3.12 *ICOS* HAPLOTYPE ASSOCIATION TESTING AND FREQUENCY ESTIMATIONS (>5%) FOR PATIENT AND CONTROL POPULATIONS ..... 95

TABLE 3.13 HT*ICOS* HAPLOTYPE ASSOCIATION TESTING AND FREQUENCY ESTIMATIONS (>5%) FOR PATIENT AND CONTROL POPULATIONS ..... 95

TABLE 3.14 PROMOTER *ICOS* HAPLOTYPE ASSOCIATION TESTING AND FREQUENCY ESTIMATIONS (>5%) FOR PATIENT AND CONTROL POPULATIONS ..... 96

TABLE 3.15 *CD28* HAPLOTYPE ASSOCIATION TESTING AND FREQUENCY ESTIMATIONS (>5%) FOR PATIENT AND CONTROL POPULATIONS ..... 96

TABLE 4.1 HARDY WEINBERG EQUILIBRIUM ANALYSIS RESULTS FOR ALL SNPS IN CANDIDATE GENES ON CHROMOSOME 11 ..... 117

TABLE 4.2 HARDY WEINBERG EQUILIBRIUM ANALYSIS RESULTS FOR ALL SNPS IN TNFSF14 ..... 117

TABLE 4.3 GENOTYPE AND ALLELE FREQUENCIES FOR CANDIDATE GENES ON CHROMOSOME 11 IN THE COELIAC DISEASE PATIENT AND CONTROL POPULATIONS ..... 121
TABLE 4.4 CARRIER STATUS FREQUENCIES FOR CANDIDATE GENES ON CHROMOSOME 11 IN THE COELIAC DISEASE PATIENT AND CONTROL POPULATIONS ..... 123
TABLE 4.5 GENOTYPE AND ALLELE FREQUENCIES IN THE *TNFSF14* GENE IN THE COELIAC DISEASE PATIENT AND CONTROL POPULATIONS ..... 126
TABLE 4.6 CARRIER STATUS FREQUENCIES IN THE *TNFSF14* GENE IN THE COELIAC DISEASE PATIENT AND CONTROL POPULATIONS ..... 127
TABLE 4.7 CD3E HAPLOTYPE FREQUENCY ESTIMATIONS (FOR HAPLOTYPES WITH PREDICTED FREQUENCY >5%) AND HAPLOTYPE ASSOCIATION TEST RESULTS FOR PATIENT AND CONTROL POPULATIONS ..... 133
TABLE 4.8 CD3D/CD3G HAPLOTYPE FREQUENCY ESTIMATIONS (FOR HAPLOTYPES WITH PREDICTED FREQUENCY >5%) AND HAPLOTYPE ASSOCIATION TEST RESULTS FOR PATIENT AND CONTROL POPULATIONS ..... 133
TABLE 4.9 ALL CD3 GENES HAPLOTYPE FREQUENCY ESTIMATIONS (FOR HAPLOTYPES WITH PREDICTED FREQUENCY >5%) AND HAPLOTYPE ASSOCIATION TEST RESULTS FOR PATIENT AND CONTROL POPULATIONS ..... 134
TABLE 4.10 THY1 HAPLOTYPE FREQUENCY ESTIMATIONS (FOR HAPLOTYPES WITH PREDICTED FREQUENCY >5%) AND HAPLOTYPE ASSOCIATION TEST RESULTS FOR PATIENT AND CONTROL POPULATIONS ..... 134
TABLE 4.11 IL10RA HAPLOTYPE FREQUENCY ESTIMATIONS (FOR HAPLOTYPES WITH PREDICTED FREQUENCY >5%) AND HAPLOTYPE ASSOCIATION TEST RESULTS FOR PATIENT AND CONTROL POPULATIONS ..... 135
TABLE 4.12 IL18 HTSNP HAPLOTYPE FREQUENCY ESTIMATIONS (FOR HAPLOTYPES WITH PREDICTED FREQUENCY >5%) AND HAPLOTYPE ASSOCIATION TEST RESULTS FOR PATIENT AND CONTROL POPULATIONS ..... 135
TABLE 4.13 IL18 PROMOTER HAPLOTYPE FREQUENCY ESTIMATIONS (FOR HAPLOTYPES WITH PREDICTED FREQUENCY >5%) AND HAPLOTYPE ASSOCIATION TEST RESULTS FOR PATIENT AND CONTROL POPULATIONS ..... 136
TABLE 4.14 TNFSF14 HTSNP HAPLOTYPE FREQUENCY ESTIMATIONS (FOR HAPLOTYPES WITH PREDICTED FREQUENCY >5%) AND HAPLOTYPE ASSOCIATION TEST RESULTS FOR PATIENT AND CONTROL POPULATIONS ..... 136
TABLE 4.15 TNFSF14 HAPLOTYPE FREQUENCY ESTIMATIONS (FOR HAPLOTYPES WITH PREDICTED FREQUENCY >5%) AND HAPLOTYPE ASSOCIATION TEST RESULTS FOR PATIENT AND CONTROL POPULATIONS ..... 136
TABLE 4.16 INFERRED ALLELE FREQUENCIES AND ASSOCIATION TESTING FOR IL18_RS5744230C/T, IL18_RS4988359A/G, IL18_RS5744249T/G, IL18_RS1834481G/C, IL18_RS5744258G/C, IL18_RS5744276G/C, IL18_RS3882892G/A ..... 138
TABLE 4.17 INFERRED ALLELE FREQUENCIES AND ASSOCIATION TESTING FOR IL18_RS3882891 ..... 138

TABLE 4.18 CROSSTABULATION OF DISEASE STATUS AND IL18HT4-607, CONDITIONING FOR TNFSF14RS2279627 ..... 140
TABLE 4.19 RISK ESTIMATES FOR THE CROSSTABULATION OF DISEASE STATUS AND IL18HT4-607, CONDITIONING FOR TNFSF14RS2279627 ..... 140
TABLE 4.20 TESTS OF HOMOGENEITY OF THE ODDS RATIO FOR CROSSTABULATION OF DISEASE STATUS AND IL18HT4-607, CONDITIONING FOR TNFSF14RS2279627 ..... 141
TABLE 4.21 CROSSTABULATION OF DISEASE STATUS AND CTLA4_CT60, CONDITIONING FOR TNFSF14RS2279627 ..... 142
TABLE 4.22 RISK ESTIMATES FOR THE CROSSTABULATION OF DISEASE STATUS AND CTLA4_CT60, CONDITIONING FOR TNFSF14RS2279627 ..... 142
TABLE 4.23 TESTS OF HOMOGENEITY OF THE ODDS RATIO FOR CROSSTABULATION OF DISEASE STATUS AND CTLA4_CT60, CONDITIONING FOR TNFSF14RS2279627 ..... 143
TABLE 4.24 CROSSTABULATION OF DISEASE STATUS AND IL18HT4_607, CONDITIONING FOR CTLA4_CT60 ..... 144
TABLE 4.25 RISK ESTIMATES FOR THE CROSSTABULATION OF DISEASE STATUS AND IL18HT4-607, CONDITIONING FOR CTLA4_CT60 ..... 144
TABLE 4.26 TESTS OF HOMOGENEITY OF THE ODDS RATIO FOR CROSSTABULATION OF DISEASE STATUS AND IL18HT4-607, CONDITIONING FOR CTLA4_CT60 ..... 145
TABLE 4.27 SIGNIFICANCE OF INDICATOR VARIABLES IN THE LOGISTIC REGRESSION MODEL FOR DISEASE STATUS USING SNPS TNFSF14RS2279627 AND IL18HT4-607 ..... 146
TABLE 4.28 SIGNIFICANCE OF INDICATOR VARIABLES IN THE LOGISTIC REGRESSION MODEL FOR DISEASE STATUS USING SNPS TNFSF14RS2279627 AND CTLA4_CT60 ..... 146
TABLE 4.29 SIGNIFICANCE OF INDICATOR VARIABLES IN THE LOGISTIC REGRESSION MODEL FOR DISEASE STATUS USING SNPS IL18HT4-607 AND CTLA4_CT60 ..... 146
TABLE 5.1 DQ2 STATUS OF INDIVIDUALS GENOTYPED FOR THE PRESENCE OF THE HIGH RISK DQ2 HETERODIMER ..... 161
TABLE 5.2 NON-SYNONYMOUS SNPS IDENTIFIED BY STEWART *ET AL.* TO DIFFER BETWEEN THE MHC EXTENDED HAPLOTYPES AH 8.1 AND AH 7.1 ..... 163
TABLE 5.3 SNPS WHICH ALTER POTENTIAL TRANSCRIPTION FACTOR BINDING SITES, WHICH DIFFER BETWEEN THE MHC EXTENDED HAPLOTYPES AH 8.1 AND AH 7.1, IDENTIFIED IN THIS STUDY FROM THE DATA OF STEWART *ET AL.* ..... 164
TABLE 5.4 NON-SYNONYMOUS SNPS IN THE *MICA* AND *MICB* GENES IDENTIFIED IN THIS STUDY FROM THE DATA OF STEWART *ET AL.* TO DIFFER BETWEEN THE MHC EXTENDED HAPLOTYPES AH 8.1 AND AH 7.1 ..... 165
TABLE 5.5 HARDY WEINBERG EQUILIBRIUM ANALYSIS RESULTS FOR ALL CANDIDATE SNPS IN THE MHC CLASS III REGION AND *MICB* ..... 166
TABLE 5.6 GENOTYPE AND ALLELE FREQUENCIES FOR CANDIDATE POLYMORPHISMS IN THE MHC CLASS III REGION AND *MICB* IN THE DQ2 HETEROZYGOUS COELIAC DISEASE PATIENT AND CONTROL POPULATIONS ..... 170

TABLE 5.7 CARRIER STATUS FREQUENCIES FOR CANDIDATE POLYMORPHISMS IN THE MHC CLASS III REGION AND *MICB* IN THE DQ2 HETEROZYGOUS COELIAC DISEASE PATIENT AND CONTROL POPULATIONS ..... 173
TABLE 5.8 COMPARISON OF HAPLOTYPE FREQUENCIES BETWEEN ALLCOCK *ET AL* STUDY AND PRESENT CONTROL POPULATION ..... 178
TABLE 5.9 TNF BLOCK HAPLOTYPE FREQUENCIES (>5%) IN COMBINATION WITH THE HLA-DQB1 LOCUS IN DQ2-CONTROLLED PATIENT AND CONTROL POPULATIONS ..... 180
TABLE 5.10 P-VALUES ASSOCIATED WITH LINKAGE DISEQUILIBRIUM LEVELS AS MEASURED BY D' VALUES BETWEEN DQB1 AND ALL MHC CLASS III SNPS STUDIED IN THE TOTAL PATIENT AND CONTROL GROUPS ..... 182
TABLE 5.11. SNP ALLELE FREQUENCIES AT EACH LOCUS ON THE HIGH RISK HAPLOTYPE OF THE DQ2 HETEROZYGOUS PATIENT AND CONTROL GROUPS ..... 189
TABLE 5.12 SNP ALLELE FREQUENCIES AT EACH LOCUS ON THE HIGH RISK HAPLOTYPE OF THE DQ2 HETEROZYGOUS AND DQ2 HOMOZYGOUS PATIENTS ..... 192
TABLE 5.13 DQB1*0201 ALLELE-CARRYING MHC HAPLOTYPE FREQUENCIES (>5%) IN DQ2 HETEROZYGOUS AND DQ2 HOMOZYGOUS PATIENT POPULATIONS ..... 195
TABLE 5.14 NUMBER OF HLA-DQB1*0201 ALLELE-CARRYING CONSERVED HAPLOTYPES IN DQ2 HOMOZYGOUS PATIENTS ..... 196
TABLE 5.15 DQB1 ALLELE FREQUENCIES IN THE DQ2 HETEROZYGOUS PATIENT AND CONTROL GROUPS ..... 197
TABLE 5.16 DQB1*0602 ALLELE-CARRYING MHC HAPLOTYPE FREQUENCIES AS A PROPORTION OF THE LOW RISK (NON DQB1*0201-CARRYING) IN DQ2 HETEROZYGOUS PATIENT AND CONTROL POPULATIONS ..... 198
TABLE 5.17 SNP ALLELE FREQUENCIES AT EACH LOCUS ON THE LOW RISK HAPLOTYPE OF THE DQ2 HETEROZYGOUS PATIENT AND CONTROL GROUPS ..... 202
TABLE 5.18 SNP ALLELE FREQUENCIES AT EACH LOCUS ON THE DQB1*0602-CARRYING HAPLOTYPES ONLY FOR THE DQ2 HETEROZYGOUS PATIENT AND CONTROL GROUPS ..... 204
TABLE 5.19 SNP ALLELE FREQUENCIES ON NON DQB1*0602–CARRYING, LOW RISK HAPLOTYPES, IN THE DQ2 HETEROZYGOUS PATIENT AND CONTROL GROUPS ..... 206

## LIST OF FIGURES


FIGURE 1.1 NORMAL AND COELIAC DISEASE SMALL INTESTINE MUCOSAE ..... 7

FIGURE 1.2 HLA-DQ2 HETERODIMER FORMATION. THE HLA-DQ2 (DQA1*05–DQB1*02) MOLECULE CAN BE ENCODED EITHER IN CIS INDIVIDUALS WHO HAVE THE DR3-DQ3 HAPLOTYPE, OR IN TRANS IN INDIVIDUALS WHO ARE DR5–DQ7 AND DR7–DQ2 HETEROZYGOUS. (SOURCE : (SOLLID 2002)) ..... 14

FIGURE 1.3 THE JOINT EFFECTS OF LINKAGE DISEQUILIBRIUM, EFFECT SIZE AND MARKER DISEASE ALLELE FREQUENCIES ON THE POWER TO DETECT DISEASE ASSOCIATION ..... 33

FIGURE 1.4 SNPS, HAPLOTYPES AND TAG SNPS ..... 40

FIGURE 2.1 OUTLINE OF THE SNP DISCOVERY PROCESS (IN NON-MHC REGIONS) ..... 54

FIGURE 2.2 OUTLINE OF THE SNP DISCOVERY PROCESS FOR MHC CLASS III REGION ..... 57

FIGURE 2.3 PROCEDURE FOR A TAQMAN ALLELIC DISCRIMINATION ASSAY ..... 62

FIGURE 2.4. PROCEDURE FOR AMPLIFLUOR SNPS DETECTION OF A G/T GENOTYPE ..... 64

FIGURE 3.1 CHROMOSOME 2Q33 INFORMATION AND GENES AS IDENTIFIED BY HUMAN CONTIGVIEW IN THE ENSEMBL (V32) GENOME BROWSER ..... 74

FIGURE 3.2 CTLA4 HAPLOTYPE TAGGING SNP SELECTION ..... 76

FIGURE 3.3 ICOS HAPLOTYPE TAGGING SNP SELECTION ..... 77

FIGURE 3.4 PERCENTAGE ERROR RATES ASSOCIATED WITH RECONSTRUCTING TOTAL HAPLOTYPE INFORMATION IN THE *CTLA4* GENE USING 5 ORIGINAL HTSNPS ..... 79

FIGURE 3.5 PERCENTAGE ERROR RATES ASSOCIATED WITH RECONSTRUCTING TOTAL HAPLOTYPE INFORMATION IN THE *CTLA4* GENE USING 5 ORIGINAL HTSNPS ..... 80

FIGURE 3.6 P-VALUES AND ODDS RATISO FOR ALLELE FREQUENCY DIFFERENCES BETWEEN PATIENT AND CONTROL GROUPS FOR SNP LOCI AT 2Q33 ..... 86

FIGURE 3.7 LINKAGE DISEQUILIBRIUM ANALYSIS OF THE 2Q33 REGION IN COELIAC DISEASE CASES (A) AND CONTROLS (B) ..... 92

FIGURE 3.8 GRAPHICAL SUMMARY OF EVOLUTIONARY CONSERVED REGION (ECR) ANALYSIS IN THE 2Q33 REGION USING THE ECR BROWSER SOFTWARE ..... 98

FIGURE 4.1 CD3E HAPLOTYPE TAGGING SNP SELECTION ..... 110

FIGURE 4.2 CD3D/CD3G HAPLOTYPE TAGGING SNP SELECTION ..... 111

FIGURE 4.3 THY1 HAPLOTYPE TAGGING SNP SELECTION ..... 112

FIGURE 4.4 IL10RA HAPLOTYPE TAGGING SNP SELECTION ..... 113

FIGURE 4.5 IL18 HAPLOTYPE TAGGING SNP SELECTION ..... 114

FIGURE 4.6 TNFSF14 HAPLOTYPE TAGGING SNP SELECTION ..... 115

FIGURE 4.7 P-VALUES AND ODDS RATIOS FOR ALLELE FREQUENCY DIFFERENCES BETWEEN PATIENT AND CONTROL GROUPS FOR SNP LOCI AT 11Q23 ..... 120

FIGURE 4.8 P-VALUES AND ODDS RATIOS FOR ALLELE FREQUENCY DIFFERENCES BETWEEN PATIENT AND CONTROL GROUPS FOR SNP LOCI AT TNFSF14 ..... 125

FIGURE 4.9 LINKAGE DISEQUILIBRIUM ANALYSIS OF THE 11Q23 REGION IN COELIAC DISEASE CASES (A) AND CONTROLS (B) ..... 129

FIGURE 4.10 LINKAGE DISEQUILIBRIUM ANALYSIS OF THE *TNFSF14* (*LIGHT*) GENE IN COELIAC DISEASE CASES (A) AND CONTROLS (B) ..... 130
FIGURE 5.1 TNF BLOCK HAPLOTYPE TAGGING SNP SELECTION ..... 162
FIGURE 5.2 P-VALUES AND ODDS RATIOS FOR ALLELE FREQUENCY DIFFERENCES BETWEEN PATIENT AND CONTROL GROUPS FOR SNP LOCI IN THE MHC CLASS III REGION AND MIC GENES ..... 168
FIGURE 5.2 P-VALUES AND ODDS RATIOS FOR ALLELE FREQUENCY DIFFERENCES BETWEEN PATIENT AND CONTROL GROUPS FOR SNP LOCI IN THE MHC CLASS III REGION AND MIC GENES (CONTINUED) ..... 169
FIGURE 5.3 ENTROPY MAXIMISATION METHOD ALLOWING IDENTIFICATION OF THE SNP LOCI WHICH CONTRIBUTE MOST INFORMATION ON HAPLOTYPE DIVERSITY. ..... 179
FIGURE 5.4A LINKAGE DISEQUILIBRIUM ANALYSIS OF THE MHC REGION IN COELIAC DISEASE CASES 183
FIGURE 5.4B LINKAGE DISEQUILIBRIUM ANALYSIS OF THE MHC REGION IN THE CONTROL POPULATION ..... 184
FIGURE 5.4C LINKAGE DISEQUILIBRIUM ANALYSIS OF THE MHC REGION IN THE DQ2 HETEROZYGOUS PATIENT POPULATION ..... 185
FIGURE 5.4D LINKAGE DISEQUILIBRIUM ANALYSIS OF THE MHC REGION IN THE DQ2 HOMOZYGOUS PATIENT POPULATION ..... 186
FIGURE 5.4E LINKAGE DISEQUILIBRIUM ANALYSIS OF THE MHC REGION IN THE DQ2 HETEROZYGOUS CONTROL POPULATION ..... 187

**ACKNOWLEDGEMENTS**

I would like to thank Dr. Ross McManus and Prof. Dermot Kelleher for giving me the opportunity to undertake this project and for the unwavering support over the last four years, it is very much appreciated.

To all within the Genomic Research group, past and present; you have been an invaluable support. A special thank you to Anthony Ryan, for being a constant source of information and for answering all my questions, no matter how trivial. Thank you to Jacinta Thornton, who took the new girl under her wing, I would never have survived SPD without you. To Kaukab Azim, always willing to take the time to answer any question with a smile. To Emma Caraher, Arun Kumar Mankan, Matthew Lawless, Graham Turner, Edel Duggan, Carol Goulding, Micheal O'Dwyer, thank you for all your help over the years.

To my partners in crime, Megan Dring and Erica Mullaney, thank you for all the support, sharing your skills and knowledge with me and helping me out so many times. But most importantly, thank you for your friendship.

To all the present members of the Kelleher group; Danijela Petrovic, Shane Duggan, Dennis Finn, Arun Chandran, Siobhan Mitchell, Deirdre Ni Eidhin, Henry Windle, Yuri Volkov, Anthony Davis, Sinead Smith, Shonagh O'Leary, Deirdre Kellya and Avril O'Riordan. Thank you also to all past members of the group, a special mention to fellow students Julie O'Brien, Eileen Looby, Aine Fox, Aine Fanning and Blathnid Mee, who all made the "basement" a better place to be. And to everyone in the IMM, in particular those in the Immunology, Neuropsychiatric genetics and Haematology groups, who have helped me in many ways over the years.

To Anthony Fitzgerald, for providing such excellent statistical assistance and for his endless patience.

To the doctors, nurses and patients of the coeliac clinic, thank you for your continued support in providing clinical samples for this project. To the Irish Blood Transfusion Service, in particular Richard Hagan, for providing healthy volunteer blood samples and for the assistance in the HLA-typing of these samples.

And finally, to all my family and friends, thank you for endless support over the years, I couldn't have survived this without you. A special thank you to my mother, Catherine and my brothers, David and Philip. David, I would still be looking at the computer wondering how to turn it on without you.

**Summary**

The aim of the present study was to investigate the genetic contribution of genes or regions other than the HLA to coeliac disease (CD) susceptibility in the Irish population. Genes selected for analysis in this study were chosen based on evidence from previous studies suggesting a genetic or functional association between the gene of interest and autoimmune disease. Genomic regions were also selected for analysis and subsequently screened for candidate genes, based on evidence from previous genome scans which identified regions linked to CD. All genes selected for analysis in this study have potential immunological functions which may contribute to the pathogenesis of coeliac disease. These include *CTLA4*, *ICOS* and *CD28* on chromosome 2q33, *IL10RA*, *CD3D*, *CD3E*, *CD3G*, *THY1* and *IL18* on chromosome 11q23, and *TNFSF14* (*LIGHT*) on chromosome 19p13.3, and a number of MHC class III genes located between the *BAT1* and *DDAH* genes.

394 coeliac disease patients and 354 ethnically matched healthy controls were genotyped at a total of 61 single nucleotide polymorphisms (SNPs) over 20 genes using RFLP-PCR, Taqman or Amplifluor technologies. SNPs were selected based on either their ability to tag haplotypes (haplotype tagging SNPs), their ability to change the amino acid sequences of a gene (non-synonymous SNPs) or their potential to alter gene expression (transcription factor binding site altering SNPs). Allele, genotype and carrier status frequencies for each individual SNP were compared statistically between patient and control groups using the chi-square or fishers exact test. Haplotype frequencies were estimated from unphased genotype data using computational methods and were statistically compared between groups. Gene-gene interaction analysis was also carried out using crosstabulation and logistic regression analysis methods. Analysis of MHC class III SNPs was carried out in a DQ2 controlled environment to avoid spurious disease associations based on linkage disequilibrium between the polymorphism of interest and the known disease associated locus HLA-DQB1. High risk and low risk HLA haplotypes were also analysed for the presence of additional risk factors in the MHC class III region which contribute to CD.

Results from this study have revealed a number of statistically significant associations between genetic variation and CD in the Irish population. In the 2q33 region, though several individual polymorphisms including CTLA4_CT60, were

found to be associated with the disease, the largest frequency differential between patient and control groups was observed for an extended haplotype in the region which incorporated *CTLA4/CD28* and 5'*ICOS*. In the 11q23 region, an association with disease was observed for the promoter polymorphisms -137 and -607 in the *IL18* gene. Haplotypes displaying a significant association with disease were also observed for *IL18*. At 19p13.3, the *TNFSF14* (*LIGHT*) gene was also found to be associated with disease with three individual polymorphisms displaying statistically significant disease associations (TNFSF14_rs2279627, ht2_rs3760746 and ht3_rs11878563). Haplotype analysis of the gene also revealed significant disease associations. This is the first genetic association study to date which has investigated the genetic contribution of *LIGHT* to autoimmune disease, and is the first to identify a significant association between the *LIGHT* gene and coeliac disease.

Gene-gene interaction effects between disease associated polymorphisms in each non-MHC genomic region were also studied. No significant gene-gene interaction effects were observed in this study.

Analysis of genes in the MHC class III region identified additional risk factors on the high risk extended haplotype at two polymorphisms in the *MICB* gene (MICBrs3134900 and MICB-7595). Analysis of this region also identified an excess of the HLA-DQB1*0602 allele in CD patients, a result which has not been previously observed in this disease.

In this study, we have identified a number of genetic variants which are significantly associated with coeliac disease, including newly investigated variants which have not been studied previously in association studies of CD. Such identified genetic factors may in the future be utilised to help elucidate the complete pathogenesis of CD and hence, assist in the developments of new effective treatments.

**ABBREVIATIONS**

| | |
|---|---|
| µl | Microlitre |
| CD | Coeliac disease |
| cM | centimorgan |
| dNTP | Deoxyribonucleotide triphosphate |
| ECR | Evolutionary Conserved Regions |
| EM | Expectation maximisation |
| EMM | Entropy maximisation method |
| g | Gram |
| GD | Graves disease |
| htSNP | Haplotype tagging single nucleotide polymorphism |
| HWE | Hardy-Weinberg Equilibrium |
| kb | Kilobase |
| L | Litre |
| LD | Linkage disequilibrium |
| MB | megabase |
| mg | Milligram |
| ml | Millilitre |
| MS | Multiple sclerosis |
| ng | Nanogram |
| RA | Rheumatoid arthritis |
| RFLP | Restriction fragment length polymorphism |
| SEM | Standard error mean |
| SNP | Single nucleotide polymorphism |
| SSD | Stephens-Smith-Donnelly |
| T1D | Type 1 diabetes |
| TCR | T cell receptor |
| TF | Transcription factor |
| UTR | Untranslated region |

## PUBLICATIONS FROM THIS THESIS

### Original papers

**Brophy K**, Ryan AW, Thornton JM, Abuzakouk M, Fitzgerald AP, McLoughlin RM, O'Morain C, Kennedy NP, Stevens FM, Feighery C, Kelleher D, McManus R. Haplotypes in the CTLA4 region are associated with coeliac disease in the Irish population. *Genes and Immun*ity 2006 Jan;7(1):19-26

**Brophy K**, Ryan AW, Thornton JM, Abuzakouk M, McLoughlin RM, O'Morain C, Kennedy NP, Stevens FM, Feighery C, Kelleher D, McManus R. Analysis of coeliac disease candidate genes in chromosome 11q23. In preparation.

**Brophy K**, Ryan AW, Thornton JM, Abuzakouk M, McLoughlin RM, O'Morain C, Kennedy NP, Stevens FM, Feighery C, Kelleher D, McManus R. Association between the TNFSF14 gene and coeliac disease in the Irish population. In preparation.

### Presentations at Scientific Meetings or Conferences

54th Annual Meeting of the American Society of Human Genetics, Toronto, Canada. October 2004. Poster Presentation: Analysis of the 2q33 region in celiac disease. Brophy K, Ryan AW, Feighery C, O'Morain C, Kennedy N, McLoughlin R, Stevens F, Kelleher D, McManus R. *Proceedings of the American Soc Human Genetics, Abs 2250*

7th Annual Meeting of the Irish Society of Human Genetics, Dublin, Ireland. September 2004. Oral Presentation: CTLA4, ICOS and CD28 haplotypes and coeliac disease in the Irish population. Brophy K, Ryan A, Feighery C, O'Morain C, Kennedy N, Abuzakouk M, McLoughlin R, Stevens F, Kelleher D, McManus R.

Digestive Disease Week, New Orleans, U.S.A. May 2004. Poster Presentation: ICOS and CD28 haplotypes and celiac disease. Brophy K, Ryan A, Feighery C, O'Morain C, Kennedy N, McLoughlin R, Stevens F, Kelleher D, McManus R.

*Gastroenterology* 126 (4): A247-A247 Suppl. 2 Apr 2004

11th International Symposium on Coeliac Disease, Belfast, Northern Ireland. April 2004.
Oral presentation: CTLA4, ICOS and CD28 haplotypes and coeliac disease in the Irish population. Brophy K, Ryan A, Feighery C, O'Morain C, Kennedy N, Abuzakouk M, McLoughlin R, Stevens F, Kelleher D, McManus R.

Irish Society of Gastroentrology Meeting, Belfast, Northern Ireland. April 2003. Poster Presentation: Cytotoxic T-lymphocyte Antigen 4 gene polymorphisms and coeliac disease in the Irish population. Brophy K, Ryan A, Feighery C, O'Morain C, Kennedy N, Abuzakouk M, McLoughlin R, Stevens F, Kelleher D, McManus R.

This thesis is dedicated to my mother, Catherine

# CHAPTER I

## GENERAL INTRODUCTION

### 1.1 Historical background of coeliac disease

Coeliac disease is an inflammatory disorder characterised by villous atropy and crypt cell hyperplasia, which results from a permanent intolerance to gluten. The central role of dietary gluten was outlined by Dicke (Dicke *et al.* 1953) in 1953, who observed that, during World War II when there was a scarcity of cereals and bread, the incidence of coeliac sprue was remarkably reduced. This observation in addition to a description by a mother of the improvement in her child's health upon removal of cereals from the diet, identified the toxic nature of wheat to coeliac disease patients. Dicke also went on the show that toxicity was found in the gluten fraction of the wheat.

### 1.2 Epidemiology

Screening studies have revealed that the incidence of coeliac disease (CD) approaches 1% of the population. CD was originally recognised as a disorder known to affect European Caucasians, however, there is increasing awareness of CD in other regions and among other ethnic groups with incidence prevalence reported around the world including Northern Africa (Boudraa *et al.* 1996), Brazil (Gangolfi *et al.* 2000), the US (Hill *et al.* 2000), New Zealand (Walsh *et al.* 2003) and the Middle East (Al-Tawaty and Elbargathy 1998). Numerous studies have been carried out to detect the prevalence of this disease. The ability to carry out screening studies on seemingly healthy individuals for the presence of CD is now possible with the development of serological tests for diagnosis. These tests include tests for endomysial and tissue transglutaminase antibodies. An extensive EU wide study has been undertaken to examine the prevalence rate of CD in the European community (This project is called COELIAC-EU/CLUSTER N° QLK1-1999-00037; data from year 2001 and 2002 annual reports). 29,077 sera were collected from individuals across Europe including countries such as Ireland, Italy, Germany and Finland. The prevalence of disease was calculated as 1 in 72 when calculated on human based kit results and 1 in 122 when calculated on the endomysial antibody (EMA) test results (methods of disease detection will be discussed in detail later). Originally, the incidence of coeliac disease was thought to be much lower, however, this is now known to be due to under-diagnosis of the disease and the presence of non-symptomatic individuals who

are nonetheless susceptible to developing disease. Indeed, the large proportion of individuals suffering from CD who are not recognised as sufferers, has led to the concept of the "coeliac iceberg". This concept states that a smaller number of individuals lie above the symptom threshold, that is, they are floating above the surface and are recognised as having CD than those below the surface who have undiagnosed, silent or latent forms of the disease. Indeed, a comprehensive study (Catassi *et al.* 1996) carried out in an Italian population which screened 17,201 individuals for CD, identified the presence of undiagnosed CD was 1 in 210 individuals, and they concluded that the ratio of known to undiagnosed CD in Italy was 1 to 7. Other studies have also investigated the prevalence of undiagnosed CD. In a Norwegian study (Hovdenak *et al.* 1999), the prevalence of symptomatic CD was calculated to be 1 in 674 individuals, however, after a screening study of 2,096 seemingly healthy individuals, 6 were identified as CD sufferers, indicating a disease prevalence of 1 in 340 in asymptomatic, seemingly healthy, individuals. A similar pattern was observed in other populations including a Finnish study (Kolho *et al.* 1998), where 8 undiagnosed CD patients were identified among 1,070 tested implying a prevalence rate for undiagnosed CD of 1 in 130. And in a study on a UK population (West *et al.* 2003), the prevalence of undiagnosed CD was estimated to be 1.2% of the total population.

The incidence of CD has been approximated at less than 1 in 100 to 1 in 300 in a variety of European populations (McLoughlin *et al.* 2003), with the incidence in Ireland estimated at 1 in 122 (Johnston *et al.* 1997). Most individuals with coeliac disease are currently undiagnosed, although the rate of diagnosis is increasing. Patients do, however, seem to have a long duration of symptoms prior to diagnosis. This may be attributed to physicians' delay in diagnosis, rather than patients delay in seeking health care.

Most patients who present with coeliac disease are female (3:1 ratio) (Green *et al.* 2001), although men seem to present with a more severe form of disease at presentation (Nistico *et al.* 1996).

### 1.3 The role of gluten in coeliac disease

Dicke *et al* (Dicke *et al.* 1953) originally suggested a role for a wheat protein in the pathogenesis of coeliac disease and it was shown that the toxic component of wheat lies in the endosperm. There are two main protein fractions in the endosperm of wheat; the prolamins (gliadins) and the glutenins. The gliadins are alcohol-soluble and can be subdivided into α, β, γ and Ω fractions based on size or according to their N-terminal amino acid sequence. Studies have shown that all fractions of gliadin exacerbates coeliac disease both in vitro and in vivo (Hekkens *et al.* 1970).

Gliadin contains high levels of the amino acid glutamine (~35%), and studies (Sollid *et al.* 2001) have shown that deamidation by the enzyme tissue transglutaminase (tTG), converts glutamine resides to negatively charged glutamic acid, which results in the creation of epitopes which bind to HLA-DQ2 (or HLA-DQ8) on antigen presenting cells and are recognised by gut-derived T cells in coeliac disease patients. Although there was a proliferative response in certain T cell lines from peripheral blood, only the gut-derived T cells could be expanded several-fold by the deamindated gliadin, suggesting that this response resulted in a local but not a systemic immune reaction.

The interaction between HLA-DQ2 and the gliadin epitope is pivotal to the commencement of the immune reaction in CD pathogenesis, and the binding properties of DQ2 have been extensively studied (Shidrawi *et al.* 1998), with a number of DQ2-restricted gliadin epitopes described, which have been shown to stimulate T cells derived from the small intestinal mucosa of coeliac disease patients. α-gliadin has been shown to contain two overlapping immunodominant epitopes to which the majority of coeliac disease T cell clones responded (Ellis *et al.* 2003). It was identified, however, that approximately half of the T cells from children suffering with coeliac disease were not stimulated with these immunodominat epitopes and it was also noted that several minor gliadin epitopes have also been identified which elicit an immune response, suggesting that coeliac disease intestinal T cell clones are stimulated by a variety of gliadin epitopes (Vader *et al.* 2002). Given these findings, it was also suggested that T cell reactivity evolves from a wide array

of native gluten epitopes in the early phase of the disease (as in the disease presented in childhood), to more immunodominant epitopes as the disease progresses.

### 1.4 Pathology

Coeliac disease pathology is observed in the tissue of the small intestine. It affects the mucosa of the proximal small intestine and in severe cases can extend into the ileum and can also affect the rectal mucosa (Ruben *et al.* 1962). Normal human intestinal mucosa has an oscillating villous structure which extends into the lumen of the gut and which increases the surface area of the lining of the intestine to aid absorption. Villi are normally three times taller than they are wide, however, in untreated coeliac disease patients, the small intestine displays a loss of this villous structure, with flattening of the villous mucosa taking place. This abnormality in villous appearance can range from mild partial villous atrophy to a total absence of villous structure. Crypt cell hyperplasia and infiltration of the lamina propria by lymphocytes and plasma cells takes place, which may result in an expansion in the thickness of the mucosal layer. The surface epithelial cells become pseudo-stratified in comparison to their normal tall columnar formation and surface enterocyte height is reduced. Thickening of the membrane basement may also take place. Due to the inflammatory nature of the disease, it is not surprising that there is an increase in the number of chronic inflammatory cells found in the surface epithelium and underlying lamina propria, in addition to the observation of plasma cells in the lamina propria and an increased number of lymphocytes in the surface epithelium (Ciclitira *et al.* 2005). Figure 1.1 shows the normal and coeliac disease small intestine mucosae.

Coeliac disease is known as an inflammatory disorder characterised by a permanent intolerance to gluten. This immune response to gluten, which is responsible for the pathology observed in the small intestine of coeliac disease patients, is known to occur in two compartments, the lamina propria and the epithelium. This compartmentalisation can distinguish the two different components of the immune response, the specific and innate response. In the lamina propria, $CD4^+$ T cells are important in disease pathogenesis. From

genetic linkage studies, it is known that DQ2 and DQ8 are associated with coeliac disease and it has been shown that DQ2 and DQ8 anti-gluten restricted T cells are present in the lamina propria of patients with CD (Sollid *et al.* 2001). These T cells secrete γ-interferon, which promotes inflammation and villous atrophy in the intestine. It is also known that gut fibroblasts, stimulated by activated $CD4^+$ T cells secrete matrix degrading enzymes (matrix metalloproteinases-1 and -3) which degrade fibrillar collagen, matrix glycoproteins and proteoglycans, hence playing a role in villous atrophy and crypt cell hyperplasia (Pender and al. 1997).

The second component of the immune response is intraepithelial lymphocytosis by $CD8^+$ T cells. It has been proposed that these T cells are not directed against the gliadin peptides themselves, but they recognise gliadin-damaged epithelial cells which produce stress-induced MIC and HLA-E molecules (Maiuri *et al.* 2001), and also secrete interleukin-15 (Roberts *et al.* 2001). Increased levels of IL-15, in turn, induces the expression of natural killer receptors NKG2D and CD94 on the intraepithelial lymphocytes, which then recognise the MIC and HLA-E molecules on the damaged epithelial cells and destroys them.

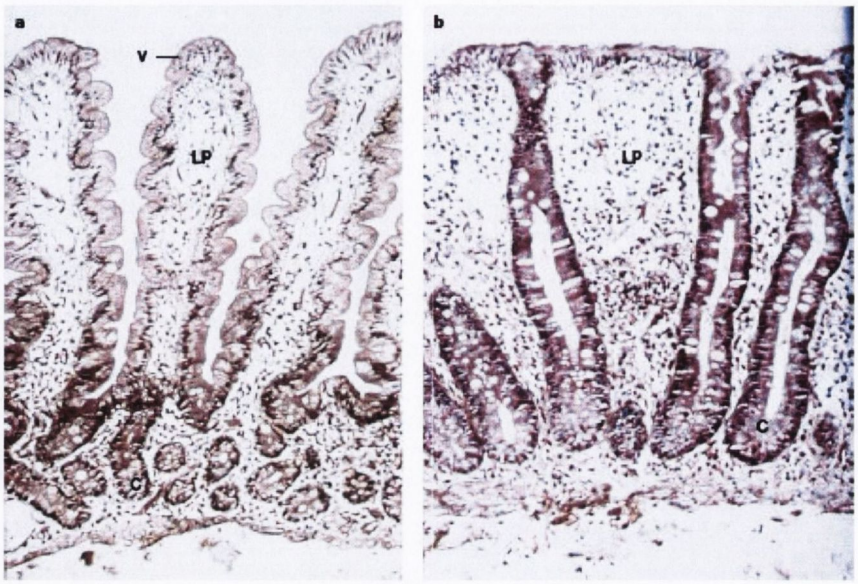



**Figure 1.1 Normal and coeliac disease small intestine mucosae. a. Histology of the normal small intestine mucosa showing normal villi (V), crypt (C) and lamina propria (LP) structure. b. The coeliac mucosa is characterised by villous atrophy, enlarged hyperplastic crypts and increased infiltration of lymphoid cells in the lamina propria and epithelium.** (Source: (Sollid 2002))

### 1.5 Clinical Presentation

In coeliac disease, disease location is in the proximal small intestine. This often results in malabsorption of nutrients such as calcium, iron, folic acid and fat-soluble vitamins, leading coeliac disease patients to experience reduced bone density, iron deficiency and folate deficiency. Diarrhoea, another factor of this disease, is mostly due to the progression of the disease into the distal small bowel (Ruben 1960). Hence, the classical picture of coeliac disease consists of malabsorption with diarrhoea and a failure to thrive in children.

However, asymptomatic, atypical or silent forms of coeliac disease also exist and these forms exhibit varied clinical features. Though the clinical presentation of coeliac disease in infancy consists of diarrhoea, abdominal distension and failure to thrive, other symptoms which can also be present include irritability, vomiting and constipation. Older children can also present with extra-intestinal symptoms including short stature, neurological symptoms

and anaemia (Green and Jabri 2003). The majority of cases of coeliac disease are, however, diagnosed in later life, with the peak age of diagnosis in the fourth and fifth decades (Green *et al.* 2001), and one-third of patients are diagnosed over the age of 65. Often, adults do not present with severe malabsorption but they present with a variety of vague symptoms; diarrhoea is still the main clinical feature but various other symptoms may also be present such as weight loss and anaemia. A small number of individuals may also present with neurological or psychiatric problems such as schizophrenia. Problems relating to osteomalacia are also seen which include bleeding diatheses, spontaneous fractures and infertility (Ciclitira *et al.* 2005). Skin disorders are also a common factor of coeliac disease and include conditions such as eczema and psoriasis. The most common of these is dermatitis herpetiformis, which is characterised as an intensely pruritic vesicular rash located symmetrically on the elbows, knees, buttocks, face, neck, trunk, and occasionally, within the mouth. Almost all patients with dermatitis herpetiformis have coeliac disease.

**1.6 Diagnosis**

A definitive diagnosis of coeliac disease can only be made using duodenal or jejunal biopsy samples to gain histological evidence of typical small intestinal mucosal abnormalities including villous atrophy, crypt cell hyperplasia and intraepithelial lymphocytosis. It should be noted that there is a range in the level of histological damage present in coeliac cases and this has been characterised on a scale as classified by Marsh (Marsh 1992). The mildest form of disease, a Marsh I lesion, is identified by normal villous structure with intraepithelial lymphocytosis, a Marsh II lesion is characterised by intraepithelial lymphocytosis in addition to crypt cell hyperplasia, a Marsh III lesion includes villous atrophy, either partial, subtotal or total. These histological profiles can be milder than expected if the patient's diet is low in gluten or if the patient is on immunosuppressants. Poorly orientated biopsies can also hinder diagnosis of disease, hence, biopsies must be interpreted with caution. It is also necessary in the diagnosis of disease to ensure that the histological abnormalities are alleviated with the removal of gluten from the diet.

Serological tests are a less invasive method of testing for the presence of coeliac disease and are, therefore, the first investigations carried out in screening for disease. Tests used in coeliac disease diagnosis are based on the IgA isotype, and include antigliadin antibodies and the connective tissue antibodies, endomysial and tissue transglutaminase antibodies. The antigliadin antibody (AGA) test was initially used in screening, however, this has been now more routinely replaced by the use of the IgA endomysial antibody (EMA) test which has higher specificity (approaching 100%) and a sensitivity of >90% (Rostom *et al.* 2005). The detection of EMAs requires indirect immunofluorescence with either monkey oesophagus or human umbilical cord as the substrate, hence making this an expensive assay. This factor and the recognition that the enzyme tissue transglutaminase (tTG) is the autoantigen for EMA, has led to the development of an enzyme-linked immunoabsorbent assay which originally used guineapig-tTG (Sulkanen *et al.* 1998), but more recently, recombinant human-tTG is more routinely used for its higher specificity (Sblattero *et al.* 2000). Overall, the sensitivity of IgA anti-tTG is >90%. It does not achieve the near 100% specificity as seen with EMAs (Rostom *et al.* 2005).

A complicating factor in the use of these tests is the fact that selective IgA deficiency occurs relatively frequently in coeliac patients (at a rate of 1.7%-2.6% (Greco *et al.* 1998)) compared with the general population. To test for coeliac disease in these patients, a test for total serum IgA needs to be incorporated into the panel of tests. In addition to this, a test for IgG antibodies, either IgG endomysial antibodies, IgG anti-tTG or IgG-AGA should be carried out. Due, to the factors outlined above, including the failure of these tests to reach 100% specificity and sensitivity in the clinical setting and the presence of selective IgA deficiency in a number of coeliac patients, it is proposed that the most effect option in serological diagnosis is a combined panel of tests which include the IgA-tTG, IgG-tTG, EMA and total IgA levels (Green and Jabri 2006).

### 1.7 Coeliac disease as a genetic disorder

Evidence that coeliac disease is a genetic disorder has been provided by familial and twin studies. Familial aggregation studies have shown that the prevalence rate of CD is higher among relatives of affected individuals than in the general population suggesting that there is a genetic component to this disease. As previously stated, the incidence of CD has been approximated at between 1% to 0.3% in a variety of European populations (McLoughlin *et al.* 2003). Differing values for the prevalence rate in first-degree relatives have been published, and range from under 5% to over 20%, however the greatest number of studies have shown the value to range from 10 to 12% (Macdonald *et al.* 1965; Mylotte *et al.* 1974; Shipman *et al.* 1975; Stokes *et al.* 1976; Stenhammar *et al.* 1982). These differing values may be explained by the heterogeneity of this disease; not only genetic but also environmental. Differing sets of diagnostic criteria between studies may also prove to be a factor in the differing percentage prevalence rates. Sibling relative risk analysis has also been used to establish the contribution of the genetic component to disease and this is defined as the risk to an affected individual's sibling divided by the population risk. Sibling relative risk has been estimated to be 30 in a British study (Ueda *et al.* 2003). However, it should be noted that in recent times, earlier estimates of sibling relative risk have changed as a result of an acceptance that the population prevalence of disease is higher than originally thought.

Lending support to the familial predisposition theory of CD are twin studies. The first comprehensive twin study was carried out in a Spanish population, pooling data from 18 centres (Fernandez *et al.* 2002). However, the value of such twin studies depends on correct methodology and ascertainment of samples. There are, in fact, two methods of recruitment of individuals for study; volunteer based and population based recruitment. Interestingly, it has been suggested that the volunteer recruitment method over-selects for monozygotic twins and also female sets of twins, and this usually results in overestimation of the genetic component of disease (Likken *et al.* 1978). Hence, a population based study was later carried out to overcome this issue (Greco *et al.* 2002). This Italian study was conducted using the national twin

registry which contains over 1.6 million twin pairs, and this was cross-matched with a coeliac patient database to identify cases. Concordance rates for disease for monozygotic (MZ) twins were 87% using probandwise concordance and 75% using pairwise condordance. For dizygotic (DZ) twins, concordance rates were 20% and 11% for probandwise and pairwise concordance rates, respectively. The pairwise concordance rate is a descriptive statistic and gives the proportion of affected pairs that are concordant for the disease and the probandwise concordance rate gives the proportion of affected individuals among the co-twins of previously ascertained index cases. The difference between the monozygotic versus dizygotic concordance rate is a good estimate of the size of the genetic component as each twin pair is presumed to share an identical environment, hence any difference in concordance rate is due to the difference in the level of genetic similarity between monozygotic and dizygotic twins (100% and 50% similarity, respectively). These results suggest that coeliac disease has one of the highest concordance rates of complex diseases, as monozygotic concordance rates for other such diseases have been estimated at 25% in multiple sclerosis (Willer *et al.* 2003), 36% in type 1 diabetes (Olmos *et al.* 1988) and 33% in Crohn's disease (van Heel *et al.* 2000).

**1.8 Human leukocyte antigen (HLA) complex and coeliac disease**

As coeliac disease has been characterised as an autoimmune disorder, an association between disease and the HLA genes of the major histocompatibility complex on chromosome 6p21.3 seems a quite likely proposition. The HLA complex is located on the short arm of chromosome 6, and is approximately 4 MB in length. The complex comprises regions encoding class I and class II cell membrane molecules. "Classic" class I molecules are designated A, B and C. Class II molecules are called DR, DQ, DM and DP. These genes are extremely polymorphic and the whole region is characterised by high levels of linkage disequilibrium. A class III region also exists in the HLA complex and consists of a number of genes including *TNF*, *IkBL*, *BAT2* and *BAT4* (Sollid and Thorsby 1993).

A large number of studies have shown an unequivocal association between disease and certain HLA alleles. The first HLA gene reported to have an

association with coeliac disease was the class I molecule B8 (Falchunk *et al.* 1972). Stronger associations were then found with the class II molecule DQ2 (Sollid *et al.* 1989), which is in strong linkage disequilibrium with the DR3 molecule, also located in the class II region. An association with DR7 was also reported (DeMorchi *et al.* 1979) but almost only in conjunction with DR3 or DR5. Studies on the haplotypes of CD patients have shown that most carry either the DR3-DQ2 or DR5-DQ7 / DR7-DQ2 haplotype at the HLA locus. The DR3-DQ2 haplotype carrys the alleles DQA1*0501 and DQB1*0201 which constitute the disease associated DQ2 heterodimer. An interesting similarity between the DR3-DQ2 and DR5-DQ7 / DR7-DQ2 haplotypes, is that both express the HLA-DQ (α1*0501, β1*02) molecules (i.e. the high risk DQ2 heterodimer). This is encoded *in cis* in DR3-DQ2 individuals. In DR5-DQ7 / DR7-DQ2, this molecule is encoded *in trans*. The DR5-DQ7 haplotype carries the DQA1*0501 and DQB1*0301 alleles and the DR7-DQ2 haplotype carries the DQA1*0201 and DQB1*0202 alleles. This allows for the encoding of the HLA-DQ (α1*0501, β1*02) transcript *in trans* (Figure 1.2). Although the cis and trans forms of the DQ2 molecule are known to differ at one amino acid in each subunit, it is not thought that this has a functional effect on the molecule as the variants occur in non-critical regions and both confer equal disease risk. There is also evidence of a gene dose effect with a number of studies providing evidence that homozygosity for the DR3-DQ2 haplotype or a second DQB1*02 allele resulted in a 4 to 6-fold increased risk of disease (van Belzen *et al.* 2004; Yamashita *et al.* 2004).

The DQA1*0501 and DQB1*0201 alleles have been sequenced, as have other HLA class II alleles in coeliac disease and they do not show any disease-specific sequences and are thought to be normal structural variants. But the fact that both disease haplotypes, DR3-DQ2 and DR5-DQ7 / DR7-DQ2, share a particular pair of DQA1 and DQB1 genes either *in cis* or *in trans* and the over presentation of such alleles in CD, with 95% of coeliac patients possessing these alleles, compared to 20 - 30% of controls (Polski *et al.* 1993), are strong indications of their involvement in CD.

The association with disease is, in fact, the result of the particular amino acid sequence identified in the peptide binding groove of the DQ2 molecule variant. The primary role of the DQ molecule is to present exogenous antigens to helper T cells in order to elicit an immune response. The DQ2 variant has been shown to contain a specific immunodominant, tissue transglutaminase – modified, gliadin epitope, hence this molecule is responsible for the presentation of the disease associated exogenous antigen to the immune system and the initiation of the inflammatory response observed in the small intestine of individuals susceptible to coeliac disease.

Though the above-mentioned haplotype seems to be the most prevalent in coeliac disease, the DQ8 molecule has also been associated with disease and the majority of sufferers who do not carry the DQ2 allele are known to carry the DQ8 molecule. The disease associated DQ8 molecule is found to occur in linkage disequilibrium with the DR4 allele and is constructed from DQA1*03 and DQB1*0302 chains. A recent study found that 6% of a European cohort of 1,008 coeliac patients carried this DQ8 molecule (Louka *et al.* 2003).

A number of association studies have been carried out on other genes located in the HLA gene complex. An association with genes in the TNF region has been proposed in a study of Irish CD patients (McManus *et al.* 1996; McManus *et al.* 1996). This finding was not replicated in a Finnish study (Polui *et al.* 1998), but has been shown replicated more recently a large Scandinavian family study (Louka *et al.* 2003). The *MICA* gene has also been the subject of investigation and has been postulated to be associated with disease, with positive associations observed in some studies (Fernandez *et al.* 2002; Louka *et al.* 2003). It is possible that there are other genes located within the HLA complex, possibly in the class III region, which are associated with disease.

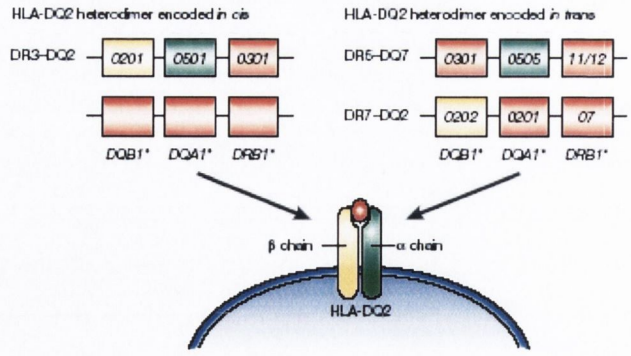



**Figure 1.2 HLA-DQ2 heterodimer formation. The HLA-DQ2 (DQA1*05–DQB1*02) molecule can be encoded either in cis individuals who have the DR3-DQ3 haplotype, or in trans in individuals who are DR5–DQ7 and DR7–DQ2 heterozygous. (Source : (Sollid 2002))**

### 1.9 Non-HLA linked genes and coeliac disease

As previously described, the involvement of HLA genes in the predisposition to coeliac disease is well characterised. However, the involvement of the HLA complex cannot account for the complete predisposition to CD as the high risk DQ2 heterodimer is present in 20 – 30% of healthy Caucasian individuals and as the incidence of coeliac disease varies from 0.3 – 1%, this implies that only about 1-5% of carriers develop disease. Further evidence is seen in the difference in concordance rates observed between monozygotic twins (87%) and HLA identical siblings (20%) (Greco *et al.* 2002). Clearly, a non-HLA linked factor is involved. The sib relative risk value for CD has been estimated to be 30, while the value attributed to the MHC region is 3.3 (Bevan *et al.* 1999). Therefore, this study states that the MHC genes contribute no more than 40% of the sib familial risk of coeliac disease and the non-HLA linked gene (or genes) are likely to be the stronger determinant of coeliac disease susceptibility.

The first proposal for the involvement of genes outside the HLA locus was in 1978 (Wu *et al.* 2004). This genetic model hypothesised that the development of coeliac disease was dependant on homozygosity at the HLA-unlinked locus and the involvement of an independently inherited gene or genes associated with the HLA region acting (subsequently identified as the DQ2 locus) in a

dominant fashion. This model found some support (Greenberg *et al.* 1982; Hernadez *et al.* 1991) but there was some question over the dominance or recessivity of the HLA-linked disease locus. Today, the generally accepted model of disease, involves a major genetic contribution from the HLA region plus the involvement of multiple non-HLA genetic loci, each of which contribute a minor genetic effect to disease.

### 1.10 Genomewide linkage studies in coeliac disease

Genomewide linkage analysis studies can be undertaken by two distinct methodologies. For monogenic disorders, if the mode of inheritance, disease allele frequencies and penetrance is known, linkage analysis is usually based on the co-segregation of a DNA marker and a disease through a family pedigree that cannot simply be attributed to chance. For complex disorders, however, where the mode of inheritance is not known or unclear, a non-parametric approach is undertaken. This method involves typing DNA markers in affected family members and identifies those that are shared between affected family members more frequently than expected. The success of this method is dependent on the contribution made by the particular locus to disease pathology.

From the methods described, the most suitable and widely used method, so far, for investigating the association of HLA-unlinked genes and coeliac disease has been the non-parametric genetic linkage analysis. Reasons for this include the lack of knowledge concerning reliable candidate genes for involvement in CD (with the exception of the *CTLA4* gene) and the uncertainty over the mode of inheritance of the disease. The success of genomewide scans has been quite limited until recent times, where improvements in methodology which recognise the need for tightly defined disease phenotypes and adequately sized cohorts have led to the identification of several disease genes, including the *NOD2* gene for Crohn's disease (Hugot *et al.* 2001) and the *ADAM33* gene for asthma (Van Eerdewegh *et al.* 2002). Genome searches for coeliac disease susceptibility loci have, on the whole, failed to identify beyond doubt any regions of the genome outside the HLA locus, although a number of regions have been replicated several times. Reasons for this may include the fact that, a

number of genes, each having a minor effect on disease pathology, may be involved, making the task of finding regions with clear association difficult to detect.

A number of independent genome screens have been carried out to try identify non-HLA candidate regions for coeliac disease (table 1.1). However inconsistent and sometimes contradictory results have been found. An early study (Zhong *et al.* 1996) on a population from the western counties of Ireland which analysed 15 families with 40 affected siblings found linkage at 6p, 30 cM from the HLA locus with a multiple linkage score (MLS) of 4.66, and evidence for linkage at 5qter, 7q31, 11p11, 15q26 and 22cen. An Italian study also identified candidate gene regions by analysing 110 families using genome scans methods (Greco *et al.* 1998). They found evidence for linkage at 5q. Using a subset of 39 sib-pairs, in which both siblings had a symptomatic form of the disease, they also found suggestive evidence for linkage to 11q23. This finding was not, however, replicated in the 71 sib-pairs in which one sibling displayed symptomatic CD and the other sibling has an asymptomatic form of the disease. They were not able to identify linkage at the other regions previously identified in the Irish sample. They postulated that this difference may be due to a difference in the genetic component of CD between the Irish and Italian populations, or the MLS values in the Irish populations may be high due to the familial structure used, where 31 of the 45 sib pairs used belong to just three nuclear families. It has been hypothesised that such a structure leads to artificially high MLS values (Babron *et al.* 1997). An additional study by the Italian group, in an independent dataset, presented further evidence for linkage to 5q33 (Greco *et al.* 2001). A study in a Finnish population attempted to replicate the previously reported linkages of chromosomal loci, 5q33 and 11q23, in addition to region 2q33 (Holopainen *et al.* 2001). They found evidence of linkage to the 11q23 region but not the 5q region using the entire dataset of 102 families. Division of families by phenotypic characteristics revealed linkage to the 5q33 and 11q regions in families with dermatitis herpetiformis, and linkage to the 2q33 region was observed in families who only developed the intestinal disease. A UK study (King *et al.* 2000) utilised a family based linkage method including 16 multiply affected pedigrees and

found evidence for linkage at 10q23 and 16q23. They also identified regions with lower levels of linkage at 6q14, 11p11 and 19q13, all regions which have been implicated in the previous Irish genome scan. Another study, this time using 106 Swedish and Norwegian families identified 8 non-HLA regions with point wise p-values below 0.05 (Naluai *et al.* 2001). Included in these regions were the previous identified 5q31-33 and 11q23 regions, in addition to 2q11-13, Xp11, 9p21, 17q22, 3p24 and 11p15. Again this study provides further evidence that 5q33 and 11q23 are likely to be true susceptibility regions in coeliac disease. An American study found intermediate linkage evidence at 18q and at 3p and suggestive evidence for linkage at 5p (Neuhausen *et al.* 2001). Linkage to regions 1p36, 4p15, 5q31, 7q21, 9p21-23 and 16q12 was observed in a genomewide scan of CD in 60 Finnish families (Liu *et al.* 2002) and regions 19p13 and 4p14 showed evidence of linkage to disease in a study of 24 multiplex families of northern European ancestry (Popat *et al.* 2002). A meta and pooled analysis of European coeliac disease data provided additional evidence for the presence of a genetic risk factor for disease at 5q31-33 (Percopo *et al.* 2003), and a more recent Finnish study has shown significant linkage at 2q23-32 and suggestive linkage to 10p (Rioux *et al.* 2004). Woolley *et al* (Woolley *et al.* 2002) identified a linkage locus at position 15q12 using a population isolate in the Koilliskaira region of Finland and Van Belzen *et al* (Van Belzen *et al.* 2003) found convincing evidence to suggest a linkage region at chromosomal position 19p13.1 in a Dutch population. This is the only region to date to reach significant linkage ($p < 2 \times 10^{-6}$) as defined by the widely employed Lander and Kruglyak criteria (Lander and Kruglyak 1995) (the evidence in the Woolley *et al* study suggested the 15q12 region also reached this level of significance, however, this population isolate is thought to be a unique case). The authors also identified association of a microsatellite marker (D19S899) showing peak linkage in an independent case-control cohort using a multi-allelic test.

Examination of results from all genome scans indicates that the body of evidence supports linkage at the non-HLA regions, 5q31-33, 11q23 and 19p13. The 5q31-33 region has been implicated in six independent genome scans (Zhong *et al.* 1996; Greco *et al.* 1998; Greco *et al.* 2001; Holopainen *et al.*

2001; Naluai *et al.* 2001; Liu *et al.* 2002) in addition to a meta-analysis of 442 European coeliac families (Percopo *et al.* 2003). The 11q23 region has been implicated in three independent scans (Greco *et al.* 1998; Holopainen *et al.* 2001; Naluai *et al.* 2001) and the 19p13 region has been shown to have the highest level of significant linkage reported to date for any genomic region in genome scans for coeliac disease (Van Belzen *et al.* 2003). All other reported loci have not shown as convincing evidence for true linkage to disease, and replication of these regions in further independent genome scans is required before these loci can be considered for further analysis.

**Table 1.1 Coeliac disease susceptibility loci identified by genome-wide linkage studies (non-HLA regions)**

| Study | Families scanned/ follow up studies | Population | Linkage regions |
|---|---|---|---|
| Zhong *et al* (1996) | 15 | Irish | 6p, 5qter, 7q31, 11p11, 15q26 and 22cen |
| Greco *et al* (1998) | 110 / 39asp | Italian | 5q, 11q23 |
| Greco *et al* (2001) | 71asp | Italian | 5q33 |
| Holopainen *et al* (2001) | 102 | Finnish | 2q33, 5q33, 11q23 |
| King *et al* (2000) | 16 | UK | 10q23, 16q23, 6q14, 11p11, 19q13 |
| Naluai *et al* (2001) | 106 | Swedish/Norwegian | 5q31-33, 11q23, 2q11-13, Xp11, 9p21, 17q22, 3p24, 11p15 |
| Neuhausen *et al* (2002) | 62 | US | 18q23, 3p26, 5p14 |
| Liu *et al* (2002) | 60 | Finnish | 1p36, 4p15, 5q31, 7q21, 9p21-23, 16q12 |
| Popat *et al* (2002) | 24 | Northern European | 19p13 and 4p14 |
| Rioux *et al* (2004) | 54 | Finnish | 2q23-32, 10p |
| Wooley *et al* (2002) | 9 / 1 | Finnish (Koilliskaira) | 15q12 |
| Van Belzen *et al* (2003) | 67 / 15 | Dutch | 19p13.1 |

asp = Affected sib pairs

### 1.11 Candidate gene studies in coeliac disease

Candidate gene studies have also been undertaken in an attempt to identify susceptibility loci which contribute to the pathogenesis of coeliac disease. Without doubt, the most studied candidate gene to date is the *CTLA4* gene on chromosome 2q33, in conjunction with its neighbouring genes, *CD28* and *ICOS*. All threes genes can be described as T lymphocyte regulatory genes; *CD28* and *ICOS* are involved in T-cell activation while *CTLA4* is a negative regulator of activation as it competes with CD28 for the same receptors, CD80 and CD86 (Carreno and Collins 2002). The initial candidate gene study chose to investigate any genetic association between the *CTLA4* gene and disease based on the immunological function of the gene (Djilali-Saiah *et al.* 1998). A positive association was observed between a polymorphism (+49A/G) in the gene and disease and this has since inspired numerous additional studies on this gene and has expanded to incorporate the neighbouring *CD28* and *ICOS* genes. Results have been varied with association between markers within or near the *CTLA4* gene and CD being observed in a number of different populations (Holopainen *et al.* 1999; Naluai *et al.* 2000; King *et al.* 2002; Popat *et al.* 2002; Popat *et al.* 2002; Mora *et al.* 2003; Amundsen *et al.* 2004; Rioux *et al.* 2004; van Belzen *et al.* 2004; Hunt *et al.* 2005), while other studies have failed to find any association (Clot *et al.* 1999; Wijmenga *et al.* 2000; King *et al.* 2003; Martin-Pagola *et al.* 2003; Haimila *et al.* 2004).

Other studies have also been carried out on various genes selected based on their immunological function and possible role in disease progression. The Matrix Metalloproteinase 1 and 3 *(MMP-1* and *MMP-3)* genes have been identified as potential candidate genes. These genes are known to degrade the extracellular matrix and expression levels of both MMP-1 and MMP-3 have been shown to be significantly elevated in the gluten-challenged CD patient gut, compared with biopsies taken from healthy individuals (Daum *et al.* 1999). Two polymorphisms associated with elevated expression of these genes were analysed in a population containing Norwegian and Swedish individuals, however, no significant association was observed between these variants and disease (Louka *et al.* 2002).

Another study investigating genetic susceptibility to CD focused on the *IL12B* gene. This gene is located within chromosomal region 5q31.1–33.1 which has been shown in a number of genome scans to be linked to disease, as previously discussed. IL12 is known to promote maturation of T-lymphocytes into the Th1 subtype and it also regulates γ-interferon production, which is known to be highly expressed in the coeliac disease lesion and is thought to play a central role in disease pathogenesis (Nilsen *et al.* 1998). A variant in the 3'UTR of *IL12B* has been shown to alter IL12 expression (Morahan *et al.* 2001) and this SNP was tested for genetic association with CD in a Norwegian and Swedish population, however, no significant association was observed (Louka *et al.* 2002).

An association study between the inducible nitric oxide synthase (*NOS2*) gene and CD also failed to identify any significant association. The villous atrophy and inflammation observed in the CD lesion are known to be associated with enhanced nitric oxide (NO) production. Though the exact role of NO is not clear, studies have shown that individuals suffering from coeliac disease have increased production of intestinal NO compared with normal controls (Beckett *et al.* 1998). An additional association study was carried out to investigate the contribution of the natural killer cell immunoglobulin-like receptors (KIRs) and the leukocyte immunoglobulin-like receptor (LILR) gene cluster to disease status. A case-control study carried out in a UK population failed to find association with any of five KIR genotype categories, the A or B KIR haplotype, the LILRA3 gene deletion or a LILRA3 exon 3 SNP (Moodie *et al.* 2002).

Efforts have also been made to identify the disease susceptibility gene or genes responsible for the linkage previously identified at 5q31. Genetic association studies have been carried out on a number of the most plausible candidate genes in the region, namely, *IL4, IL5, IL13, IL17B, NR3C1* and also at the *IBD5/SLC22A4* locus (Ryan *et al.* 2004; Ryan *et al.* 2004a). Variants within these genes were chosen based on functional relevance, prior association with autoimmune disease and also, where possible, on the ability of such variants to

tag the haplotype variation present in the region. No evidence of a significant association with disease was observed with any variant studied.

**1.12 Association studies - Identification of candidate genes**

Association studies rely on the careful selection of good candidate genes for analysis. As it has been approximated that there between 18,000 - 20,000 genes located within the human genome, the task of identifying the handful of genes conferring susceptibility to any complex disorder would be an immense undertaking without a systematic selection process or some prior information to assist in gene selection. A number of sources of information have been identified to aid in the identification of candidate genes for analysis (Zeggini *et al.* 2005).

The biology of a protein is a useful indicator of potential candidacy of a gene. If the the protein function indicates that it may be involved in the pathogenesis of the disease, this would suggest that the gene is a good potential candidate. This method is advantageous when the molecular mechanisms of a disease are known, however, it is not useful when the primary disease pathways are unclear. Use of such a priori information on the biological pathways involved in disease has been attributed to the successful identification of the Factor V Leiden gene in thrombophilia pathogenesis (Bertina and al. 1994).

Animal studies also offer useful *a priori* information on possible gene candidacy. These allow the identification of genes involved in the pathogenesis of related diseases in animals. These methods provide real evidence of the effect of gene dysfunction on an actual phenotype, however, it should be noted that there is the possibility that there are species differences in physiology and in patterns of genetic variation. The association between the *ApoAV* gene and hypertriglyceridaemia was identified by the utilisation of information gained through animal studies (Pennacchio *et al.* 2001).

Pharmacological analysis is also a useful tool in candidate gene selection. This method suggests that a gene encoding a protein involved in the process of a disease-modifying drug may indeed be a good candidate gene for examination.

Advantages of this method are identified from evidence that modification of particular pathways by certain small molecules can affect the outcome of disease and this would suggest that genetic variation in certain genes could do likewise. However, with the use of this method, it must be recognised that pharmacological treatment may not always act on the actual disease pathway. An example of a successful genetic association study using the *a priori* information gained from pharmacology studies is the association of the *PPARG* gene in type 2 diabetes (Stram *et al.* 2003).

Monogenic forms of disease are also a useful source of information. In this situation, genes which lead to monogenic or syndromal forms of disease, sometimes also show common genetic variation which confers susceptibility to the more common polygenic forms of the disease. This method is advantageous as such rarer monogenic disorders provide evidence that a particular gene has a critical function in disease pathogenesis, and indicates lack of compensation or plasticity of the gene function. However, for the majority of complex disorders, there is a lack of a monogenic model.

Positional information is also extremely useful in candidate gene identification. Genome-wide linkage scans, like those carried out for coeliac disease, identify regions with are likely to contain disease susceptibility genes. This method is advantageous as genes which generate linkage signals, reproducible in independent studies, most probably present major susceptibility genes. However, such linked regions generated from genome scans are usually quite large in size and can contain hundreds of genes, and also not all susceptibility genes will be located within regions which are shown to be linked to disease.

Another useful resource for the detection of possible candidate genes is prior association studies. These studies have identified genes which have been shown to have a significant association with disease in particular populations and this leads to the suggestion that identified genes may also play a role in disease pathogenesis in the population under investigation.

To best utilise all the *a priori* information available, the most beneficial strategy for the identification of candidate genes, is to combine information from as many external sources as possible, to identify the maximum amount of information, which will, in turn, lead to a more informed candidate gene selection choice.

### 1.13 The case-control association study design

Genetic association studies are the basis for detection of genetic variants which are associated with disease. Association studies generally take the form of either of two options; the family based study design or the case-control study design. The case-control study design is the most widely used population-based design for genetic association studies and is the method employed in this study. The objective of this study design is to compare exposure to risk factors (genetic or environmental) between a patient population and unaffected healthy control individuals, who have been chosen from the same population as the patient group, to identify associations between risk factors and disease.

An essential factor in a successful genetic association study of disease is the appropriate selection of individuals for inclusion in the case-control study design. A well designed case-control study will contain individuals from a population that is relatively homogenous, and this is generally achieved through selection of individuals from the same source population. Correct sampling of the population should ensure that selected control individuals represent the spectrum of allele frequencies and environmental exposures present in the underlying total population from which the sample was selected. Additionally, the selected patient sample should represent the spectrum of allele frequencies and environmental exposures present in the underlying disease population. It should be noted however, that the strict definition of random sampling of cases from a population can generally only be achieved for conditions where occurrence is automatically registered (e.g. a cancer incidence register) (Zondervan and Cardon 2004). Therefore, most disease studies do not lend themselves to true random sampling as they are not centrally registered, and another important consideration is the fact that some diseases can remain undiagnosed in certain individuals (e.g. coeliac disease)

and these will not be selected in the sampling set. For most case-control studies, the only feasible option is to recruit patients from individual hospitals or medical centres. In these cases, the control population should then be selected from the catchment area of the patient recruitment centre. It has been suggested that recruitment of patients from multiple centres may led to increased genetic and environmental heterogeneity compared with single-centre studies if there is population substructure between the regions in which the recruitment centres are located (Zondervan and Cardon 2004). This may result in a decrease in the increase in power expected with the addition of extra patients from the extra centres.

Indeed, the issue of population stratification is important when considering the design of case-control association studies (Cardon and Palmer 2003; Marchini *et al.* 2004). The degree of importance of this factor and the problem, if indeed any, posed to genetic association studies is still up for debate, with limited data until recently available to examine the issue. However, a recent study analysed data from 11 previous case-control studies and did not detect statistically significant evidence for stratification in any study (Freedman *et al.* 2004). However, after increasing the number of markers and samples in one study of individuals with African American ancestry, they found that stratification was in fact present. A similar increase in study design in a sample of European Americans did not reveal any evidence for statistically significant stratification. As the African American population has more genetic heterogeneity that the European American population, these results may suggest that this increased level of genetic diversity is susceptible to population stratification. They also suggest that stratification is probably most problematic in populations whose ancestors recently mixed due to intercontinental migrations. Given a marker which displays a significant association with disease and which has a substantial magnitude of effect and a highly significant p-value, the authors of the study suggest genotyping of a few dozen unlinked markers, which are likely not to have a disease effect, as a method to rule out gross stratification as the explanation for the positive disease association.

The common measure of the effect of the risk factor in the patient population relative to the control population in a case-control study is the odds ratio (OR). This is defined as the odds of exposure among cases divided by the odds of exposure among controls. An OR of 1 means that individuals exposed to the risk factor (e.g. genetic variant) are at no increased risk of disease compared with those who are not exposed to the risk factor. An OR of >1 implies an increased risk and $0 < OR < 1$ implies a protective effect. Statistical analysis is used to determine whether the odds ratio implies a significant association between the risk factor and disease incidence.

### 1.14 Factors influencing genetic association studies

After the sequencing of the human genome and the subsequent availability of large amounts of information on genetic variation present within that genome, much hope was placed on the success of population-based genetic association studies for the identification of the genetic variants responsible for the many inheritable diseases present in the human population. However, to date, the success of association studies to conclusively identify genes involved in disease processes has been quite limited. There have however been some notable exceptions, for example, the successful identification of the factor V leiden gene in deep vein thrombosis (Bertina and al. 1994). A number of factors have been suggested as the reasons for the lack of success of such studies and these include genetic heterogeneity, phenotypic complexity, sub-optimal sampling, environmental influences, genotyping error, small sample sizes and over-interpretation of marginal findings (Weiss and Terwilliger 2000; Cardon and Bell 2001). Careful design of the association study and consideration of a number of factors which influence such studies is necessary to optimise the ability of these studies to detect disease susceptibility variants.

An important aspect to realise and consider when designing association studies is that such studies are, in fact, indirect association studies. This means that the genetic variants being tested in these studies are probably not those which are directly involved in disease pathogenesis. As for most complex diseases, like coeliac disease, there are generally not specific hypotheses available for the involvement of individual genetic variants in disease susceptibility. The actual

functional effects of polymorphisms, where such exist, have been demonstrated for only an extremely limited number of genetic variants. So, while testing variants in genes or regions of interest, it is likely that the actual susceptibility allele is not genotyped but might be located near the genotyped marker or markers which are of no direct consequence to disease pathology. With the knowledge that this is the approach undertaken in genetic association studies, a number of factors, involving the contribution of genetic variants to disease and the nature and characteristics of the genetic material surrounding the susceptibility variant need to be considered when designing experiments, determining the power of such studies to detect significant association between variants and disease, and the interpretation of the outcome of association testing.

### 1.15 Factors influencing genetic association studies - Odds Ratio (OR) of the disease variant

The OR of the disease variant is an important factor in the successful identification of such variants in association studies. Complex diseases by definition are multifactorial disorders, with a complex interplay between numerous genetic and environmental factors. It is likely that genetic variants are, in general, incompletely penetrant, and individual variants do not contribute a major risk to disease and hence, have a relatively small odds ratios. If a common risk factor conferred a high OR, it would have to be the principal contributor to the disease (Zondervan and Cardon 2004) and this is very rarely observed in association study findings although many association studies have been preformed to date (Altmuller *et al.* 2001). A notable exception, however, is the *APOE(*4)* allele in Alzheimer's disease, which has an disease susceptibility allele frequency of 15% and an OR of 3.3 (Roses 1997).

### 1.16 Factors influencing genetic association studies – Linkage disequilibrium

As genetic association studies use on indirect method of testing for association which relies on the ability of one marker to detect association at a neighbouring marker, this method, hence, relies on a level of co-segregation between the two

markers. That is, the marker variant will be more often co-inherited with the disease susceptibility variant than would be expected under the principal of independent assortment. This lack of independence is termed linkage disequilibrium (LD) and is often found to occur at varying degrees between loci on the same chromosome. This LD between loci is the result of a shared ancestry and the levels of LD can be influenced by varying factors which have had an effect on the ancestral population, including forces such as recombination, natural selection, mutation, genetic drift, ancestral population demographics and mating patterns (Zondervan and Cardon 2004). An example of alteration in the levels of LD as a result of population history is the differing levels of LD observed between African populations (or populations with African ancestry) and non-African populations (Gabriel and al. 2002). LD, on average, extends over regions of 11 kb in African populations, while this extends to an average of 22 kb in non-African populations, therefore, LD patterns in non-African populations have been shown to extend to longer distances when compared to African populations. This may reflect a population bottleneck at the time when modern humans first left Africa. There has much recent interest in characterising the pattern and level of LD present in the human genome. In general, genetic variants which are close together are more highly correlated than those that are located further apart. However, patterns of LD are known to be quite variable within and between regions and this variability is quite unpredictable (Reich and al. 2002). For example, genetic variants adjacent to each other on a chromosome may display low levels or no LD between them, while markers which are located tens of kilobases or more from each other may be in complete LD. Hence, it is necessary for association studies to be aware of the levels of LD which exist locally in candidate regions under investigation if it is to be utilised in such studies. This problem is being addressed presently, with the focus of LD studies shifting from investigating average LD levels over the genome, to studies focusing on specific patterns of LD which exist in different regions of the genome (Daly *et al.* 2001; Dawson and al. 2002).

To measure the levels of linkage disequilibrium between genetic variants, the most common statistical measures used are the D' and $r^2$ statistics (Devlin and

Risch 1995). The D' statistic tends to be favoured in medical genetic studies, possibly as a result of the ease of interpretation of the resulting scale from 0 – 1, where a value of 0 means total independence between the markers and a value of 1 indicates total linkage, that is, all copies of the rarer allele occur exclusively with one of the two possible alleles at the other marker. The D' statistic is a result of the equation $D' = |D/D_{max}|$, where D is the difference between the probability of observing two marker alleles on the same haplotype and the probability of observing them independently in the population.

$$D = f(A_1B_1)-f(A_1)(B_1)$$

where A and B are genetic variants with alleles 1 and 2, and f = frequency. Whether D is positive or negative depends on the arbitrary labelling of alleles. $D_{max}$ is the lesser of $f(A_1)(B_2)$ or $f(A_2)(B_1)$ if D is positive, or $f(A_1)(B_1)$ or $f(A_2)(B_2)$ if D is negative. Because the sign is arbitrary, the absolute value of $D/D_{max}$ is used to calculate D'. If only three of the possible four haplotypes are present in the population for the loci under question, then LD is complete using the D' statistic (D' = 1). D' is not dependent on allele frequencies and can reach a value of 1 when the allele frequencies vary widely, as it only reflects the correlation since the most recent mutation occurred.

However, another measure of LD termed $r^2$, does in fact take allele frequencies into consideration.

$$r^2 = D^2/f(A_1)f(A_2)f(B_1)f(B_2)$$

$r^2$ is only equal to one when the marker variants have identical allele frequencies and every occurrence of an allele at each of the markers perfectly predicts the allele at the other locus. No conclusive answer has been established as to what level of LD is required for successful analysis in association studies. For D', various levels of LD have been suggested as useful in such studies, for example, values of D' = 0.5 or D' half-length (the midpoint between minimal and maximal LD (Abecasis 2001; Reich and al. 2001). The interpretation of $r^2$ values is a little simpler. For this measure, for a given power

and significance threshold, the required increase in sample size to allow for reduced LD between the marker and disease variants is inversely proportional to $r^2$. For example, if 1,000 cases and controls were needed to detect significant disease association in an association study where there was perfect LD ($r^2 = 1$) between marker and disease variants, in a case where there was imperfect LD (e.g. $r^2 = 0.4$), then the study would need 1,000/0.4 (2,500) cases and controls to achieve the same likelihood of detecting association.

### 1.17 Factors influencing genetic association studies – Disease and marker allele frequencies

There is large variation in the number and population frequency of disease susceptibility genetic variants. This ranges from monogenic disorders where only one genetic variant is responsible for disease, to multifactorial disorders such as coeliac disease, where there are multiple genetic variants responsible for disease, although the exact number of these variants or genes is largely unknown. Indeed, even the frequency of disease-associated alleles at susceptibility loci may vary widely with both rare and common variants postulated to be responsible for disease. However, little is known about the range of allelic frequencies for disease variants in many disorders such as coeliac disease. An important hypothesis which considers the frequency of disease-susceptibility variants in complex disease is the Common Disease/Common Variant (CD/CV) model. This proposes that the genetic risk for common diseases will often be due to disease-susceptibility alleles with relatively high frequencies. As previously mentioned, such common variants have been identified for some complex diseases, such as the *APOE(*4)* allele in Alzheimer's disease (Corder and al. 1993). While the CD/CV hypothesis has gained support (Lohmueller *et al.* 2003), an alterative hypothesis has also been proposed; the rare variant model, which suggests that rare variants are instead responsible for the majority of susceptibility to complex disease (Nistico *et al.* 1996; Weiss and Terwilliger 2000). The rare variant model poses problems for the use of association studies to detect disease-susceptibility variants. These variants (proposed to be usually <1%) are initially hard to detect in the population and will not be identified in general mutation studies, hence their presence will generally be unknown. Also, rare variants will often result in

allelic heterogeneity, for example, if a gene contains many low frequency mutations at varying sites, as is sometimes the case in Mendelian disorders (Green *et al.* 1999), this can greatly reduce the power of current statistical tests to detect association. Hence, common variants are more susceptible to identification as disease-causing variants using the present association study design. Indeed, a recent meta-analysis investigating replication rates in association studies, had to limit the investigation to studies which identified common variants as susceptibility alleles, as there were not enough published reports to permit meta-analysis of rare variant studies, probably owing to the greater difficulties in studying rare variants (Lohmueller *et al.* 2003).

Indeed, it is not just the disease susceptibility variant allele frequencies which are important to consider in association studies, but it is also important to consider the marker allele frequency. Whatever the disease-susceptibility allele frequency, regardless of whether it is common or rare, the power of an association study is greatest when the marker and disease allele frequencies are the same (Muller-Myhsok and Abel 1997; Abecasis *et al.* 2001).

### 1.18 Influence of factors on the power of single marker tests in association studies

We have discussed so far, a number of factors which have an influence on the outcome of association studies, that is, the apparent effect size observed at the marker locus of interest. As mentioned above, these factors include the actual odds ratio of the disease variant, the disease allele frequency, the marker allele frequency and the extent of linkage disequilibrium between the marker and disease loci. Disease variants with common frequencies are easier to detect than rare variants with the same OR. It has also been stated that a lack of correlation between disease and marker allele frequencies leads to a loss of information and the ability to detect the disease variant, and this loss of information is, in turn, exacerbated by low or incomplete levels of LD between markers. A recent study has utilised a sample population of 5,000 case alleles and 5,000 control alleles to determine the power of a case-control study to detect significant association with disease, in relation to the factors described above which determine apparent effect size (Zondervan and Cardon 2004).

Figure 1.3 shows the level of power associated with this study design in relation to alterations in the marker versus disease allele frequencies for models of large effect size of the disease variant (OR = 4.0), moderate effect size (OR = 2.0) and small effect size (OR = 1.2), assuming relatively high LD (D' = 0.8) versus moderate/low LD (D' = 0.5) and very low LD (D' = 0.3). For disease-susceptibility variants with large effect sizes, marker alleles of all frequencies produce a high power of detection, at all levels of LD. For disease-susceptibility variants with moderate effect size, common disease variants (>1%) have a high power of detection when using common markers, exact correlation between disease allele frequencies and marker allele frequencies is not crucial, particularly for higher levels of LD. For disease-susceptibility variants with low effect sizes, the power to detect significant association for common disease variants is only present when disease allele frequencies and marker allele frequencies are correlated and when LD is high.

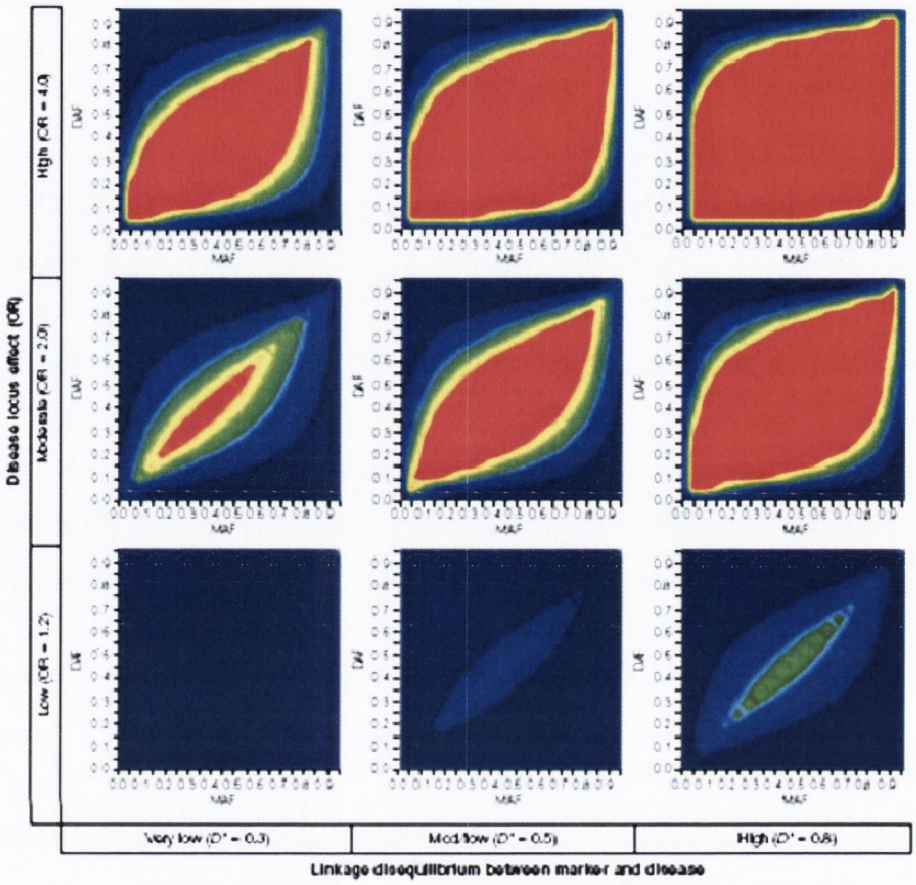



**Figure 1.3 The joint effects of linkage disequilibrium, effect size and marker disease allele frequencies on the power to detect disease association.** (Red indicated the highest power to detect apparent allelic odds ratios, navy indicates the lowest power to detect apparent allelic odds ratios) (Source: (Zondervan and Cardon 2004))

It must be noted, however, that while such information on the effects of various factors on the power and outcome of case-control association studies is an important consideration, in reality, for the majority of association studies, we do not know in advance the OR of the disease variant, the exact location or frequency of the disease allele and the level of LD between this variant and the marker of interest. Such factors can indeed be estimated, and these estimates are likely to reflect the actual situation much closer as our knowledge of the genomic architecture and behaviour of the human genome increases.

### 1.19 Genotyping accuracy

A good association study design will endeavour to ensure high genotyping accuracy, and will test and report this accuracy within the publication of results. Though most studies make every effort to ensure high levels of genotyping accuracy, empirical data has shown some assays to be unreliable and have found a residue level of error in most studies (Ewen *et al.* 2000; Little *et al.* 2002). In general, most journals have failed to specify the need for authors to report genotyping error rates, or even to identify the steps undertaken to ensure error rates are kept acceptably low. This has led to the inability of readers to assess the quality of the data presented in the study. Even well-designed experiments using new technologically advanced genotyping methods have reported error rates close to 1%, and SNP-specific error rates of up to 3% (Mein *et al.* 2000), so the occurrence of error rates of a few percent are not atypical even under ideal conditions. Studies have been carried out to assess the extent to which random genotyping error reduces the power to detect true associations with disease in the case-control study design (Gordon *et al.* 2002; Kim *et al.* 2005). These studies have shown that for each 1% rise in genotyping error, a sample size increase of 2 – 8% is required to maintain constant type I and type II error rates. This shows that random genotyping error does have an appreciable effect on the power of such a study design, however, of more concern, is the effect which systematic genotyping error has on the outcome of association studies. Such errors may lead to spurious genetic associations with disease.

### 1.20 Haplotypes and association studies - Advantages of haplotype analysis

This discussion, thus far, has focused on the use of single marker tests, and their suitability to the investigation of the genetic component of complex diseases. However, the analysis of single markers with disease association does not take into account the entire background of genetic variation present in the candidate region and, hence, may fail to identify the actual contributing factors to disease, or may possibly fail to utilise additional genetic polymorphisms which will more readily identify the actual disease variant. Haplotype analysis has now become an important component in association studies. Haplotype analysis involves the simultaneous investigation of multiple closely linked

polymorphic markers on the same chromosome and the use of haplotype analysis is considered to have numerous advantages over single marker analysis. These advantages include an improvement in the probability of detecting a disease association as it provides a more comprehensive and informative analysis of the genetic variation in the region and in certain cases will increase the power of a study to detect associations.

One advantage is that this approach takes into account the biologic function of genes. The biological and physical properties of a protein such as the folding kinetics and stability of the protein may depend on interactions between pairs or higher-order combinations at the amino-acids sites within that protein, and the sequence of these amino-acids will be a direct result of the DNA sequence of that gene with its associated polymorphic variants. Hence, if these interactions are important, then haplotypes are of direct biological relevance (Clark 2004). Another consideration for the use of haplotype analysis over single marker testing is that haplotypes arise as an intrinsic attribute of population genetic variation. That is, variation in populations is inherently structured into haplotypes, with haplotypes with multiple variable genetic sites having shared common ancestry (Clark 2004), and this may be an important factor to consider when trying to identify susceptibility factors for diseases which are population specific or are found to have differing prevalence rates in different populations.

An issue which presents an advantage for the use of haplotype analysis over that of single marker testing is the improvement in the statistical power of association testing using haplotypes compared to single marker analysis. The majority of studies on the power of such tests have usually compared single marker tests corrected for multiple testing using the Bonferroni correction, to a global statistic of haplotype associations. These studies showed that when the single marker test under investigation included the disease causative variant or variants, and the number of disease causative variants was less than the number of haplotypes in the region, these single-locus tests were more powerful than haplotype based tests (Bader 2001). However, as has been discussed previously, it is most probable that the variants under investigation in

association tests are not the actual disease causative variants but are most likely to be in the vicinity of these variants and experience some level of LD with the said variants. Haplotype based methods are likely to be more powerful than single marker methods when the SNPs assayed are not the causative variants but are in LD with the causative variant (Akey and Xiong 2001).

### 1.21 Haplotype structure in the human genome

A haplotype can be defined as a segment of chromosomal DNA, which contains a number of polymorphic loci which are in LD and hence co-segregate more often than would be expected by chance. There has been much interest in the pattern of LD in the human genome and this, in turn, has led to, the identification of haplotype patterns and structure. A number of recent studies have identified the haplotype structure in differing regions of the genome as a "block-like" pattern, with a series of high LD regions (blocks) separated by short segments of very low levels of LD.(Daly *et al.* 2001; Patil *et al.* 2001; Dawson and al. 2002; Gabriel and al. 2002; Phillips *et al.* 2003) Hence, this block like pattern of haplotypes seems to be a common feature throughout the genome.

There have been, however, a number of methods described to define a haplotype block, and inevitably this leads to differences in the patterns and sizes of the haplotype blocks identified. There are two general methods of haplotype block inference; the first uses pairwise LD measures and the second makes use of known or inferred haplotypes. When pairwise LD measures are used, a haplotype block is defined whenever all pairwise coefficients (adjacent and non-adjacent) within a region exceed some pre-defined threshold. This method has been implemented in a number of studies (Gabriel and al. 2002; Phillips *et al.* 2003), however, it must be noted that the thresholds used to define the level of LD which identifies a block are arbitrary and subject to the authors discretion. For the method which makes use of known or inferred haplotypes, the definition of a haplotype block is usually defined when a small number of haplotypes (e.g. three or four) account for a high proportion of the observations (75 – 90%) (Cardon and Abecasis 2003). Again, however, this method is arbitrary and subject to author discretion.

An important aspect of haplotype block pattern and structure to consider before utilising such haplotypes in disease association studies, is the impact of genotyped marker density on this pattern of haplotype blocks. The density of SNPs used in the identification of haplotype block patterns have ranged from approximately one marker per kilobase (Jeffreys *et al.* 2001; Patil *et al.* 2001) to one marker per 15 kb (Dawson and al. 2002) and it has been shown that studies using denser marker sets have resulted in shorter haplotype blocks. This would suggest that by genotyping more markers, long haplotype blocks are being broken down into shorter blocks, hence, with the use of a sparse genotyping marker set, it would seem that some haplotype diversity will be overlooked, with the convergence of two or more haplotype blocks, which would have been distinguished by the use of non-genotyped markers, into one longer block. One study has suggested that marker density of one marker per 7.8 kb is sufficient to identify the haplotype block pattern throughout the genome (Gabriel and al. 2002), however, extra studies need to be carried out to conclusively answer this question.

### 1.22 Haplotype blocks and association mapping

The block like structure of haplotypes present in the human genome has an impact on the way in which genetic association studies can be implemented. A number of agencies including the National Institute of Health have funded a project called the Haplotype Mapping (HapMap) project, with the aim of identifying genetic variation in the human genome and cataloguing this in terms of the haplotype structure which it creates. It thus aims to characterise the structure of all common haplotypes in the genome and is an invaluable tool in association studies (Kruglyak 2005; Zeggini *et al.* 2005). This tool will allow for the identification of the main haplotypes present in each haplotype block and the identification of the smallest set of SNPs (haplotype tagging SNPs) which will be needed to distinguish these haplotypes from each other. The haplotype block based approach has a number of advantages in genetic association studies. The presence of regions which display high levels of LD enclosed by 'hotspots' of recombination, acts to delimit potential candidate regions and enclose the search for disease causative variants to a specific

region of the genome. In addition, once all the common haplotypes in the region have been distinguished from each other, there is no need to genotype extra variation in the region as this can be inferred from the data already gathered. This can greatly reduce the genotyping costs associated with such studies allowing maximum information on a greater number of genetic markers to be extracted with minimum input.

A common criticism of the HapMap project is that the majority of SNPs encompassed by the project are common variants (Couzin 2002) suggesting that the use of this approach will fail to detect any rare variants which are associated with disease as only the most common haplotypes will be identified. However, if the Common Disease, Common Variant (CDCV) hypothesis is true, the failure to detect rare variants will not be a major issue, as CDCV holds that rare variants are not the main contributing factors to common disease. The haplotype based approach is a very useful tool in genetic association studies and is at present, the only conceivable method to achieve complete SNP genotyping coverage over the human genome, with one study (Gabriel and al. 2002) estimating that approximately 300,000 and 1,000,000 SNPs would be needed to scan the non-African and African populations, respectively, by this approach. While this approach is possibly, to date, the best method for association study design, it should be noted that, while a great proportion of the genetic variation present in the genome is organised within the haplotype block pattern, and hence can be detected by this method, there will remain a proportion of variants, both common and rare, which will not be organised into this simple block like pattern, and these variants will need modified methods for detection and analysis (Cardon and Abecasis 2003).

Also, the long stretches of LD which characterise haplotype blocks and allow for the genotyping of relatively few SNPs to capture all variation within a defined area, though extremely valuable in the detection of regions or genes associated with disease, can also cause problems in the identification of the exact location of the disease causal variant. When there is a significant level of LD between all variants, the exact causal variant cannot often be identified. One suggested solution to this problem is after the initial detection of an

association in the non-African population of interest, examine the area and perform fine mapping in African or African-American populations, which have reduced levels of LD, to further reduce the area containing the disease susceptibility variant or variants (Cardon and Abecasis 2003).

### 1.23 Haplotype tagging

The idea that all, or a large proportion, of the genetic variation present within a haplotype can be tagged by a smaller subset of SNPs, called haplotype tagging SNPs (htSNPs), has been mentioned in the proceeding discussion. The use of these htSNPs is an important advance in association studies. The ability to use these htSNPs comes from the fact that within haplotype blocks there is limited haplotype diversity, and extensive LD within markers. How small this subset of SNPs (htSNPs) can be depends on the degree of redundancy amongst the markers. To identify a group of htSNPs to tag the genetic variation in a haplotype of interest, the entire SNP set identified in the haplotype or region are genotyped in a limited number of control individuals, and from this data, the haplotype structure of the region is constructed, and finally, the smallest subsets of SNPs, which will capture all, or a specified proportion of the variation, will be identified. This selected subset of SNPs can then be genotyped in the full study and the full haplotypes can be inferred from this. The strategy for SNP selection is outlined in figure 1.4.

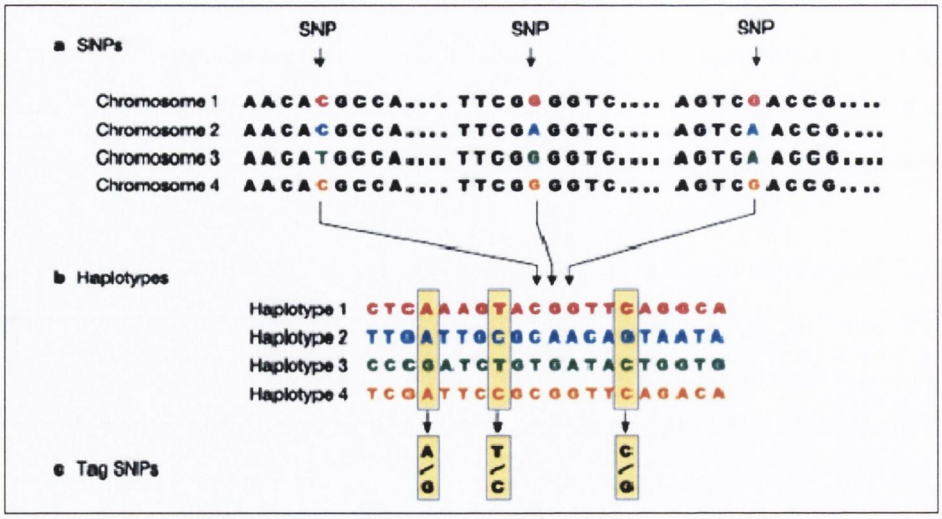



**Figure 1.4 SNPS, Haplotypes and Tag SNPs. a. SNPs. SNPs are highlighted within a chromosomal sequence from four individuals. b. Haplotypes. Haplotypes are composed of a combination of alleles at nearby SNPs. c. Tag SNPs. Genotyping of just the three tag SNPs from the 20 variants present, is sufficient to identify these four haplotypes from each other.** (Source: (The International HapMap Consortium 2003))

Knowledge of the haplotype block structure, though useful, is not necessary for htSNP selection, and htSNPs have been identified using two different approaches. One method utilises the haplotype block structure of the region and uses a unique htSNP set to tag each block separately (Daly *et al.* 2001; Patil *et al.* 2001; Gabriel and al. 2002). The second method does not take the haplotype block structure into account, and identifies a htSNP set which is most informative across the region of interest as a whole (Ackerman *et al.* 2003; Ke and Cardon 2003). This latter method of htSNP selection usually results in the need for fewer htSNPs. With regard to the differences in tagging efficacy using the haplotype block approach or by incorporating the region as a whole, a recent study found that though the haplotype block based approach required a larger number of SNPs than did the unstructured approach, it gave lower levels of error in haplotype inference (Forton *et al.* 2005). This was only of particular concern however, in regions where there were lower levels of LD present, and also when there was a significant amount of missing data to consider during haplotype reconstruction.

Using this htSNP method, theoretically, the htSNP set chosen should reflect the overall variation in the region. This is an assumption however, and if, indeed, it is not true, and the htSNPs capture less variation than is believed, it is probable that this will lead to an under-powered study which excludes an unexpected proportion of the region's genetic variation. A number of studies have been carried out to assess the effect of SNP marker density on the efficacy of htSNPs to detect all or a specified proportion of variation in a given region. One study examined SNPs with an average spacing of just over 1.5 kb, and concluded that, on average, 80% of haplotypes could be captured using only 25% of the full SNP set, with a loss of power of only 4% (Zhang *et al.* 2005). An additional study found that whatever density of SNPs was used, htSNP analysis captures at least 80% of the variation in most cases (Iles 2004).

**1.24 Use of public resources for association studies**

As stated, with the completion of the sequencing of the human genome, large efforts have been expended in identifying and characterising the genetic variation present within this genome. With the establishment of large collaborative groups, such as the International HapMap Consortium, a considerable amount of variation data is available on publicly assessable databases. These databases are an invaluable and often referenced and utilised source of information to identify genetic variation for use in association studies which use both a single marker and haplotype based approach.

To assess how comprehensive any association study is, with regards to the proportion of genetic variation captured and assessed in any candidate region, the comprehensiveness of the reference material used to design the study must be known to some degree. A number of studies have been carried out to address whether there are enough SNPs in the public databases to create adequately dense maps of local linkage disequilibrium patterns and thereby to identify disease associations. Other studies have compared the numbers of SNPs found in public databases in specific regions in which researchers have independent data regarding the number of SNPs present. One study examined 173 kb in 17 genomic locations which had been independently re-sequenced in a relatively large number of individuals of west African and European ancestry,

and these independently identified SNPs were compared with those already listed in public databases (Reich *et al.* 2003). Of the independently identified common SNPs (with a minor allele frequency >10%), a considerable number (54%) were already found listed in the public databases. Another study examined whether the SNPs listed in the public databases were adequate to fully describe all the LD relationships in the region (Li *et al.* 2003). This study found that it was possible to use a set of predictive SNPs made up solely from SNPs present in the public arena to test for association to nearly 80% of all common SNPs in a Caucasian population, indicating that these SNPs represent a good presentation of the LD relationships present in the region.

The construction of these public databases is still a work in progress with new variation being added regularly. There is reason to suggest that, at present, particularly with the latest release of HapMap data, there is enough information within these databases to detect at least some genetic variants which cause susceptibility to disease. A good example of this was a study by Vermeire *et al.*, which demonstrated that previously identified loci shown to confer susceptibility to Crohn's disease could have been detected indirectly using haplotypes reconstructed solely from SNPs available on public databases (Vermeire *et al.* 2002).

**1.25 Haplotype reconstruction**

The advantages of the use of haplotypes in genetic association studies has been discussed above. To identify haplotype structure experimentally would involve considerable cost, as routine genotyping efforts in unrelated individual do not provide phase information. To obtain such data on the phase of alleles at certain loci would involve sequencing from clones large enough to directly obtain phasing from multiple sites, or it would involve the genotyping of one or more additional family member(s), which may only result in partial identification of the phase state. Alternatively, to avoid the cost and effort associated with the identification of phased haplotypes through experimental means, statistical methods have been developed to infer haplotype phase and frequencies and such methods are used widely at the present time in haplotype analysis. These methods include the maximum likelihood method implemented

via the expectation-maximisation (EM) algorithim (Excoffier and Slatkin 1995). This robust and highly efficient approach estimates the frequencies of the haplotypes in the population from unphased genotypes counts. This algorithm attempts to find the set of (unknown) population haplotype frequencies which maximises the likelihood;

$$L(F) = Pr(G|F)$$

where $G = (G_1,\ldots, G_n)$ denotes the (known) genotypes for the individuals and $F = (F_1,\ldots, F_M)$ denotes the set of (unknown) population haplotype frequencies. This likelihood is simply the probability of observing the sample genotypes, as a function of the population haplotype frequencies, under the assumption of Hardy-Weinberg equilibrium.

Another method for haplotype inference is Clark's algorithm (Clark 1990). This method begins with the identification of individuals who are homozygous for every genotype under investigation or are single-site heterozygotes as these provide unambiguous haplotypes which exist in the population. This provides a list of all "known" haplotypes in the population. Once this list has been constructed, the haplotypes on this list are considered one at a time, to see whether any of the remaining unresolved genotypes can be resolved into a "known" haplotype plus a remaining complementary "unknown" haplotype. This set of genotypes is then considered resolved and the complementary haplotype is added to the list of "known" haplotypes. The algorithm continues to cycle through the list in this fashion until all genotypes are resolved or no further genotypes can be resolved in this way.

The third method uses Bayesian statistics to infer haplotype frequencies. Bayesian statistics incorporate prior knowledge into algorithms to obtain a more accurate outcome. Haplotype inference using Bayesian methods treat the unknown haplotypes as random quantities and combine prior information, which is a belief about the sort of pattern of haplotypes which would be expected to be observed in population samples, with the likelihood, which is the actual information in the data. These are used to calculate the posterior

distribution, which is the conditional distribution or frequencies of the unobserved haplotypes given the observed genotype data (Stephens and Donnelly 2003). There have been a number of different algorithms produced using the Bayesian framework (Stephens *et al.* 2001; Lin *et al.* 2002; Niu *et al.* 2002; Stephens and Donnelly 2003). These methods differed slightly in the prior information which they used to construct the model and the computational algorithm which they employed. Niu *et al* (Niu *et al.* 2002) and Lin *et al* (Lin *et al.* 2002) both employed the Dirichlet prior in their model. The Dirichlet prior assumes parent independent mutation, that is, the genetic sequence of the disease affected offspring does not depend on the progenitor sequence. This assumption, though simple to implement does not reflect the situation observed in natural human evolution. Both algorithms by Stephens *et al* (Stephens *et al.* 2001; Stephens and Donnelly 2003) make use of the approximate coalescent model as a prior. This model reflects the more realistic situation that the genetic sequence present in the disease affected offspring will differ only slightly (often by a single base change) from the progenitor sequence. Differences in the four studies have also been identified in the computational algorithm used. Both Stephens *et al* (2001) and Lin *et al* use Gibbs sampling, a type of Markov chain – Monte Carlo algorithm while Niu *et al* and Stephens *et al* (2003) uses the prior annealing and partition ligation (PL) algorithm. The latter method requires much less computational effort, hence the superior Bayesian algorithm is that employed by Stephens *et al* (2003), as they have implemented a more realistic prior and have complemented their efforts by using a much less computationally intensive algorithm.

### 1.26 Rational and objectives of this study

This study has been designed to investigate the genetic component responsible for the pathogenesis of coeliac disease. There has been much evidence from familial and twin studies to suggest that such a genetic component exists. The contribution of the HLA locus in disease susceptibility has been well established with the majority of coeliac patients carrying the high risk DQ2 heterodimer. There is also evidence to suggest that non-HLA genes are also involved in the disease process. Numerous genome scans have been carried out, and this has led to the identification of a number of genomic regions which

have shown linkage with disease. There is strong evidence to suggest that some of these regions do, indeed, harbour true susceptibility loci as linkage to these regions has been replicated independently in a number of different populations. At the present time, we are better positioned that ever before to investigate these regions and identify the genes and genetic variants within these genes which are responsible for disease. This is due to great advances, in both the understanding of the pathology of the disease in question and in the field of genetics. In recent times, through the sequencing of the human genome and efforts such as the HapMap project, we have gained access to vast quantities of genetic data, with regards to the genetic variation present in the genome and the relationship between these variants. This data, in combination with the development of high throughput genotyping methods and increased understanding of the need for good association study design, and the factors which need to be considered during this study design, have improved the prospects that we may identify disease susceptibility alleles using these methods. Nonetheless, this remains a difficult task with more advances needed, particularly in dealing with the problem of analysing relatively large data sets in relatively small numbers of individuals (the problems of multiple testing and dimensionality).

In this study, all the above knowledge and advances will be combined in an investigative study to facilitate identification of susceptibility genes in coeliac disease. Genomic regions which have the greatest body of evidence to suggest that they are truly linked to disease will be investigated for the presence of potential candidate genes (region 5q31-33 will not be included in this study, as this region is presently under investigation by another member of this laboratory group). Genetic variants in these genes will then be selected and subject to extensive analysis to examine if such genes confer susceptibility to coeliac disease. Such analysis will utilise the large amount of actual raw data available and will avail of the insights and advances in association study methodology available at the present time.

# CHAPTER II

## GENERAL MATERIALS AND METHODS

### 2.1 Coeliac disease patient classification and control collection

Blood samples from 394 unrelated Irish individuals suffering from coeliac disease were collected from the Department of Gastroenterology, St. James Hospital, Dublin, from the Adelaide and Meath Hospital, Tallaght, Dublin and from the University College Hospital, Galway. Patients were subject to biopsy and diagnosed on the basis of histologic appearance and response to a gluten-free diet. 354 unselected ethnically matched unrelated healthy controls were also collected for a case-control study. Biological samples and clinical data were collected with informed consent and ethical review board approval.

### 2.2 DNA extraction from blood samples

DNA was extracted from whole blood samples using the QIAamp DNA Mini Kit (Qiagen, Colonge, Germany) using 200 µl of whole blood according to the manufacturer's instructions. The extracted DNA was then stored at -20°C in the supplied elution buffer (buffer AE). Additional blood samples were extracted externally using the Gentra automated DNA extraction system (Gentra Systems, Inc., MN, U.S.A.).

### 2.3 DNA quality control and quantification

A subset of the extracted DNA samples was visualised on a 2% agarose gel (agarose gel electrophoresis is described in section 2.11.3) to establish the quality of extracted DNA. Samples were run alongside a 1 kb DNA marker ladder (New England Biolabs, MA, USA). DNA concentration was assayed using spectrophotometery on an Eppendorf BioPhotometer (Eppendorf GA, Hamburg, Germany). 5 µl of the extracted DNA was added to 95 µl of distilled water in an Eppendorf UVette (Eppendorf GA, Hamburg, Germany) and optical density readings at 260 nm and 280 nm were taken to identify the quantity and purity of the DNA extracted.

### 2.4 Power calculations

Power calculations were performed using the Genetic Power Calculator (Purcell *et al.* 2003). The power of this study was explored using a disease allele frequency of 10%, disease prevalence of 1%, required level of statistical significance, $\alpha$, of 0.05, and varying genotype relative risks ranging from 1.3 to

2.0. Power was calculated for the 394 cases available in this study with a control:cases ratio of 1. The number of cases necessary to provide 80% power to successfully detect a disease effect at $\alpha = 0.05$ was also calculated for the various genotype relative risks.

### 2.5 HLA-DRB1, HLA-DQB1 genotyping in coeliac disease and patient samples

281 coeliac patients and 305 healthy control individuals were genotyped to establish their HLA status. HLA-typing was carried out periodically by the Irish Blood Transfusion Service, St. James's Hospital, Dublin. Samples were genotyped at either the HLA-DRB1, HLA-DQB1 (and HLA-DPB1) loci; the HLA-DRB1 and HLA-DQB1 loci; or the HLA-DQB1 loci only.

### 2.6 Identification of candidate genes for analysis

The process of candidate gene identification and SNP discovery and selection is outlined in figure 2.1. The initial approach to identifying candidate genes for analysis involved examining genomic regions which have been shown in previous genome scans to be linked to coeliac disease. Several such scans are known to date (Zhong *et al.* 1996; Houlston *et al.* 1997; Greco *et al.* 1998; King *et al.* 2000; Greco *et al.* 2001; Holopainen *et al.* 2001; Naluai *et al.* 2001). Priority was given to regions which have been independently replicated in separate studies and regions which cover an area of manageable size for this scale of study. This method identified chromosomal region 11q23 as a good candidate region as it has been identified in three independent genome scans (Greco *et al.* 1998; Holopainen *et al.* 2001; Naluai *et al.* 2001).

A literature review of previous genetic association studies in immune-mediated diseases was also undertaken to identify genes which have already been shown to be associated with coeliac disease, even if these genes do not necessarily fall within a previously linked genomic region. This identified the *CTLA4* gene in chromosomal region 2q33, which was first tested for association with disease based on its immune function and was shown to have a significant association with disease status (Djilali-Saiah *et al.* 1998).

Previous studies also have suggested that regions with a well-established link to CD (i.e. the MHC region, in which HLA-DQ2 has been shown to have a functional role in CD) may also harbour a secondary disease susceptibility locus. Examples of such studies include, the TNF locus (McManus *et al.* 1996), or the MICA locus (Lopez-Vazquez *et al.* 2002). These findings have led to a more systematic investigation of the genes in the MHC class III region being undertaken in this study.

Candidate genes were also selected on the basis of experimental evidence suggesting a potential role in the pathogenesis of an autoimmune or immune-mediated, inflammatory disorder such as coeliac disease. (An example in this study is the *LIGHT* (*TNFSF14* gene) as described in chapter 4.)

For genomic region 11q23, the microsatellite markers which delimit the linked region were identified from a previous genome screening study (Holopainen *et al.* 2001). These marker identifiers were entered into the Ensembl Human MapView site (http://www.ensembl.org). This identifies all genes listed between any two microsatellite markers. A literature review of the functions of these genes was then carried out to identify those with an immune role. Such genes with an immune or inflammatory function may play a part in the pathogenesis of coeliac disease and hence are potentially good candidate genes for investigation.

To identify any genes of potential functional interest surrounding the previously implicated CD susceptibility gene, *CTLA4*, on chromosome 2q33, the NCBI Map Viewer utility (http://www.ncbi.nlm.nih.gov/mapview/) was employed to identify surrounding genes. A literature review was conducted to distinguish those with functions which may suggest a role in the pathogenesis of CD. A recent published study (Horton *et al.* 2004) has listed all genes in the MHC class III region and this study was used as a resource for gene location and function.

## 2.7 Identification of genetic polymorphisms for analysis in candidate genes in regions 2q33 and 11q23.3

### 2.7.1 SNP selection criteria

Once a list of candidate genes was established, the next step in the systematic approach to the identification of gene polymorphisms in CD candidate susceptibility genes involves screening these genes for SNPs in the most comprehensive, but cost-effective, manner. SNPs were selected for genotyping based on the following criteria;

1. The ability to tag a haplotype and other SNPs present on that haplotype using haplotype-tagging SNPs (htSNPs).
2. SNPs which alter a transcription factor binding site in the promoter region of a candidate gene were included.
3. SNPs which change the amino acid sequence of a candidate gene (i.e. non-synonymous SNPs).
4. SNPs shown in previous studies to have a significant association with CD or another immune-mediated disease. Examples of these SNPs include, CTLA4+49 (Djilali-Saiah *et al.* 1998; Naluai *et al.* 2000; Popat *et al.* 2002; Mora *et al.* 2003), IL18-607 and IL18-137 (Kretowski *et al.* 2002; Sivalingam *et al.* 2003).

### 2.7.2 Haplotype-tagging SNP identification

To investigate if a haplotype-tagging approach could be undertaken in the genetic analysis of a candidate gene, an examination of publicly available genotype information was carried out. These sources include the HapMap project and this information is freely available on the World Wide Web. All such identified projects (listed in Appendix I), as well as published studies on haplotype-tagging of genes of interest for this study were screened. Haplotype information was available for the *CTLA4* gene from a previous gene re-sequencing and polymorphism identification study (Johnson *et al.* 2001). This study listed haplotypes composed of 12 SNP loci and identified the 5 htSNPs needed to tag these haplotypes. Haplotype information was also available for the *ICOS* gene from a similar previous study (Haaning Andersen *et al.* 2003). This study listed 16-SNP haplotypes. No genotype or haplotype information

was available for the *CD28* gene on chromosome 2q33, hence individual SNPs (non ht-SNPs) were selected from the NCBI dbSNP database (http://www.ncbi.nlm.nih.gov/SNP). Information available on all candidate genes on chromosome 11q23 and in the *TNFSF14* gene, was in the form of genotype or haplotype information obtained from various databases as outlined in table 2.1.

**Table 2.1 Genotype reference information sources for chromosome 11 genes and *TNFSF14***

| Gene | Genotype Reference Source | Nature of information | Web Address |
|---|---|---|---|
| *IL18* | Innate Immunity PGA | genotype | http://innateimmunity.net/PGAs/InnateImmunity/IL18/ |
| *IL10RA* | SeattleSNPs Variation Discovery Resource | haplotype | http://pga.gs.washington.edu/data/il10ra/ |
| *CD3D* | Gene Genotype Report (NCBI) | genotype | http://www.ncbi.nlm.nih.gov/SNP/GeneGt.cgi |
| *CD3E* | Gene Genotype Report (NCBI) | genotype | http://www.ncbi.nlm.nih.gov/SNP/GeneGt.cgi |
| *CD3G* | Gene Genotype Report (NCBI) | genotype | http://www.ncbi.nlm.nih.gov/SNP/GeneGt.cgi |
| *THY1* | Gene Genotype Report (NCBI) | genotype | http://www.ncbi.nlm.nih.gov/SNP/GeneGt.cgi |
| *TNFSF14* | Gene Genotype Report (NCBI) | genotype | http://www.ncbi.nlm.nih.gov/SNP/GeneGt.cgi |

The genotype information obtained for the SNPs of interest was formatted for entry to the HITAGENE program (http://www.hitagene.com) which was used to generate haplotypes using the expectation maximisation (EM) algorithm. The genotype data was also formatted for entry into the PHASEv2.0.2 program (Stephens and Donnelly 2003), which generates haplotypes using the Stephens-Smith-Donnelly (SDS) method. Haplotypes with a frequency greater than 4 - 5% were considered for further analysis. These haplotypes were then formatted for analysis by SNPtagger (http://www.well.ox.ac.uk/~xiayi/haplotype/), a web-based program used to discover htSNPs from haplotype information. Reconstructed haplotypes with frequencies > 4 -5% were entered into the

program and the resulting output identified htSNPs which were then considered for genotyping analysis in our study.

### 2.7.3 Transcription factor binding site altering SNPs

To investigate if a polymorphism alters a transcription factor (TF) binding site, the Transcription Element Search System (TESS) program (www.cbil.upenn.edu/tess/) was used to identify consensus TF binding sites in DNA sequences. In the use of this program, the "Filtered String-based Search Query" option was utilised. DNA sequence corresponding to 10 base pairs on either side of the polymorphism of interest was entered into the program twice; once with allele 1 of the polymorphism present in the sequence and then with allele 2 present in the input sequence. Comparison of the two outputs of TF binding site patterns allows the identification of any alteration in consensus sequences caused by the polymorphism. The "Factor Filters" option was selected. Under "Factor Attribute 1", the organism species selected was "Human".

### 2.7.4 Non-synonymous SNP selection

Since non-synonymous SNPs alter protein sequence, these are high priority polymorphisms to include in disease association studies. Non-synonymous SNPs were identified on the NCBI dbSNP database (http://www.ncbi.nlm.nih.gov/SNP). Only non-synonymous SNPs with an identified minor allele frequency greater that 5% were included.

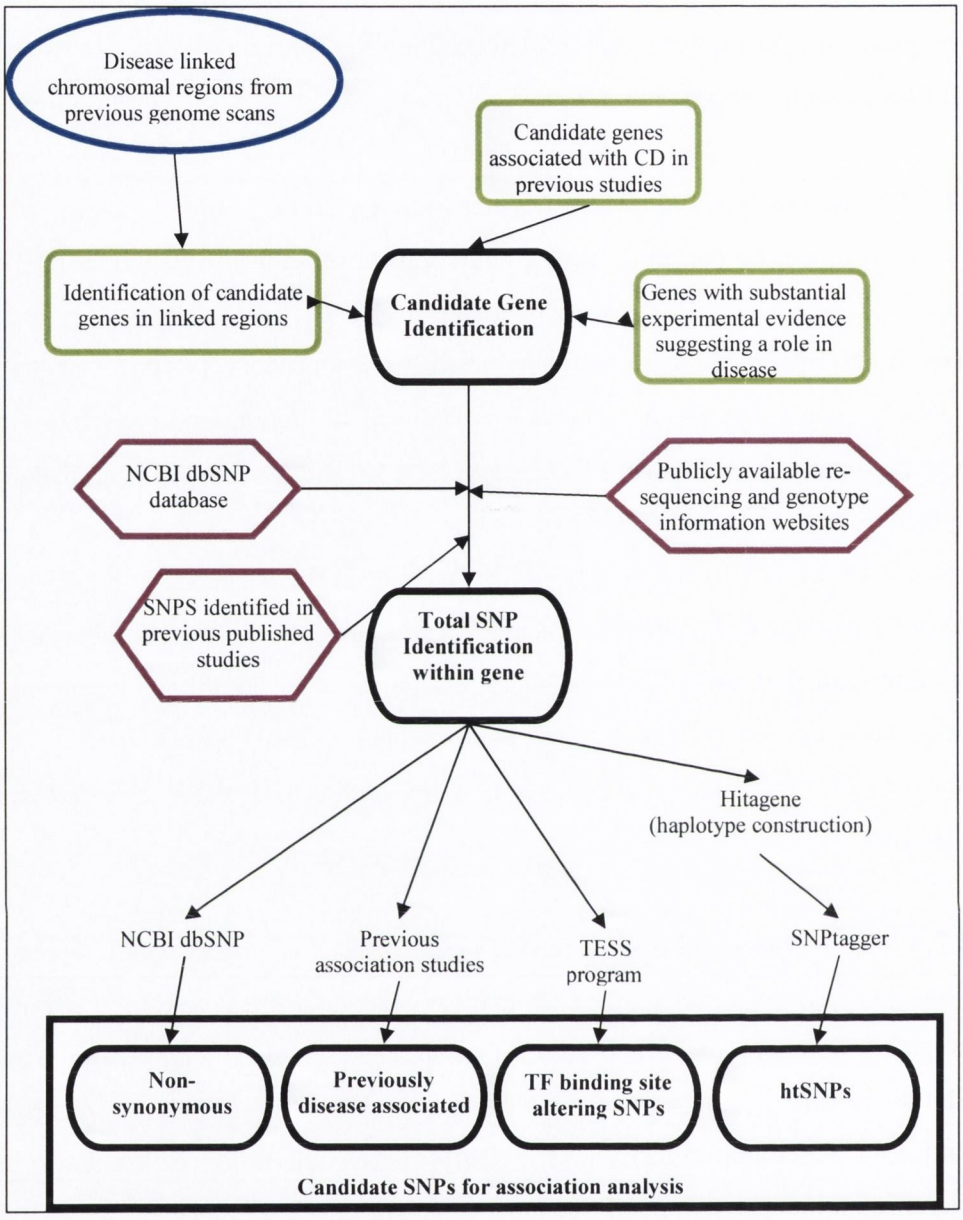



**Figure 2.1 Outline of the SNP Discovery Process (in non-MHC regions)**

**2.8 Identification of candidate polymorphisms in MHC class III region**

Haplotype analysis in the MHC class III region is complicated by a number of reasons which will be discussed in later chapters (Chapter 5), hence a different but well-structured approach had to be implemented for polymorphism selection in this region. This method is outlined in figure 2.2. This region has been previously identified as a candidate region in a number of studies, examples of such studies are McManus *et al.* and Lopez-Vazquez *et al.* (McManus *et al.* 1996; Lopez-Vazquez *et al.* 2002). Selection of SNPs in this region focused on two key areas;

1. Use of haplotype tagging SNPs to tag a haplotype composed of SNPs in the TNF region which have potential functional effects. Eight htSNPs were identified using the htSNP identification method as outlined in section 2.7, which had the ability to tag a haplotype containing 17 SNPs of potential interest in this region. The haplotype structure for this region had been determined in a previous study by Allcock *et al.* (Allcock *et al.* 2004).

2. Selection of non-synonymous SNPs in the MHC class III region between the genes *BAT2* and *DDAH*, which differ between the MHC disease protective haplotype HLA-A3, B7, Cw7, DR15 (DR2) and the MHC disease associated haplotype HLA-A1, B8, Cw7, DR3. The base resource for this investigation was a study by Stewart *et al.* (Stewart *et al.* 2004) which identified more than 18,000 polymorphic differences between these two haplotypes across the entire MHC region.

   To select, identify and verify the polymorphisms to be analysed from the Stewart *et al.* data, the following approach was used;
   Stewart *et al.* used the submitter handle, SI-MHC-SNP to submit SNPs identified in their study to the NCBI database Each MHC class III gene (as listed in a previous published study (Horton *et al.* 2004)) was individually searched for non-synonymous SNPs which have been submitted by the above handler name on the dbSNP database on NCBI. This method, which distinguishes SNPs according to their location

within a gene (i.e. 5'UTR, intron, exon or 3'UTR), and their effects on the coding sequence (i.e. synonymous or non-synonymous) was necessary, as the published paper or associated website did not give a list of these non-synonymous SNPs.

To avoid the inclusion of spurious sites, SNPs identified as non-synonymous polymorphisms submitted by handle SI-MHC-SNP were then double checked by cross-referencing these SNPs to those on the Sanger MHC Haplotype Project website (http://www.sanger.ac.uk/HGP/Chr6/MHC/) (the reference database for the Stewart *et al* study) to ensure that they occur on a clone listed on this site, and hence are really part of their study.

The search for SNPs in promoter regions of the gene followed the same methodology as the search for non-synonymous SNPs (as listed above). Using the sequence positions as listed in dbSNP (NCBI), SNPs were checked to ensure that they were within 1,000 bp upstream of the specific gene transcription start site of interest. To identify if SNPs altered a transcription factor binding site, the TESS program was utilised as outlined in section 2.7.

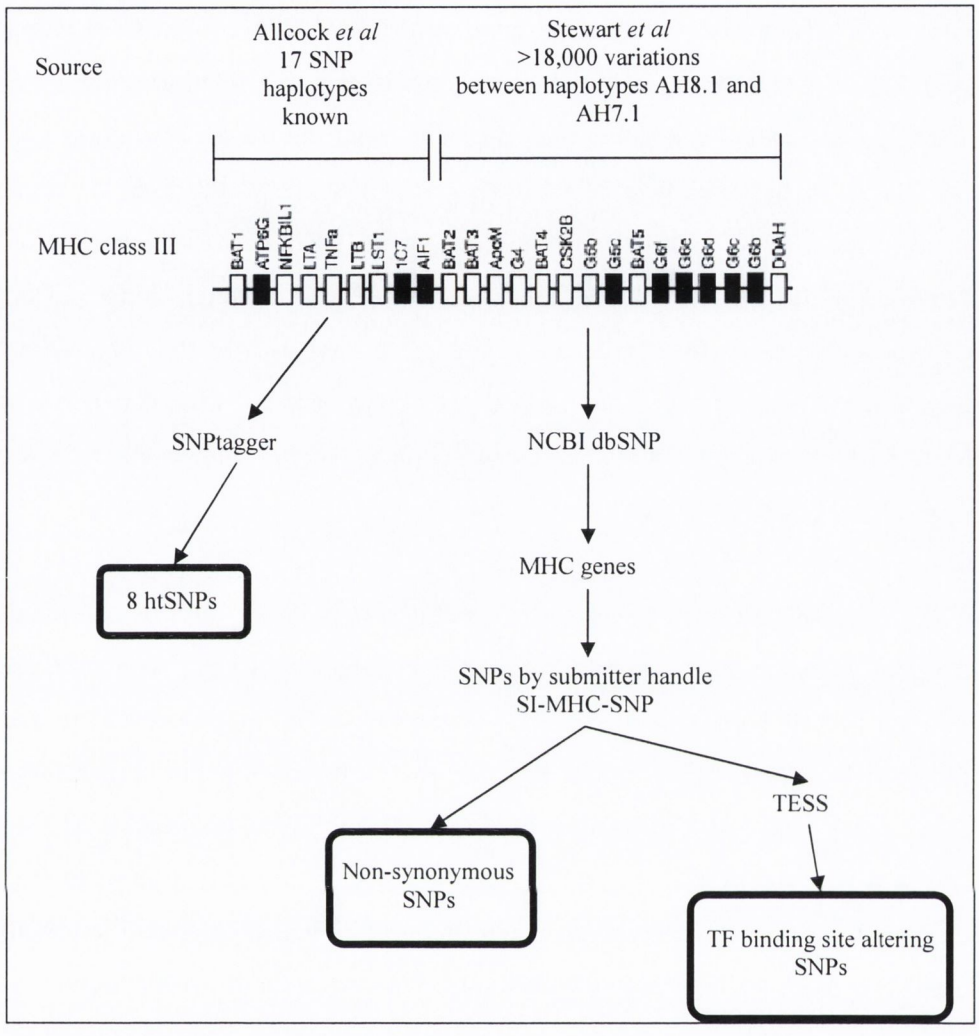



**Figure 2.2 Outline of the SNP Discovery Process for MHC class III region**

## 2.9 Analysis of the accuracy of haplotype reconstruction from haplotype-tagging SNPs

A web-based program called MARKER (http://www.gmap.net/MARKER/) was used to estimate the likely percentage error rates expected to occur at each locus on a haplotype which has been tagged by htSNP data (Forton *et al.* 2005). This program (MARKER) achieves this by simulating the exercise of typing htSNPs in a large sample of individuals and comparing the true and inferred haplotypes.

The user enters the query data into the program. As in the present study, this data includes the haplotypes of interest (including all the locus information present on that haplotype) and their population frequencies. For this study, the

haplotypes of interest are those surrounding CTLA4 (figure 3.2), ICOS (figure 3.3), CD3E, CD3D/G, THY1, IL10RA, IL18 and TNFSF14 (figures 4.1 – 4.6). The user also enters a separate data file containing the names of each specific SNP locus in the haplotype of interest. MARKER allows the user to specify which of the SNP loci will be defined as the htSNPs. (The program will provided a default set of htSNPs which have been derived using Entropy software. The user can then select or change this set of htSNPs). The htSNPs tested in this study have been selected previously using SNPtagger. However MARKER is used here because it can estimate the percentage error rates at each locus on the haplotype.

To do this, the program generates a simulated population of 500 individuals. To this 'virtual' population, the program will randomly assign two of the user defined (input) haplotypes (e.g. CTLA4 haplotypes as in figure 3.2) to each individual in the population according to their population haplotype frequencies as defined by the user (e.g. CTLA4 haplotype frequencies as listed for each haplotype in figure 3.2). HtSNP genotype data is then generated for each individual in the population using the htSNPs which were defined by the user. The program then uses the EM algorithm to generate htSNP haplotypes. Next these htSNP haplotypes are extended to inferred full haplotypes (i.e. haplotypes incorporating both the htSNPs and all non-genotyped loci). These inferred full haplotypes are then compared with the starting haplotypes originally assigned for each individual, to assess the number and location of any errors between inferred and actual full haplotypes.

The program repeats these simulations five times without missing genotype data (blue error bars) and also repeats the simulations five times with up to 20% missing data (red error bars) at each htSNP locus. An overall error rate (mean ± SEM) for each locus on the haplotype is then calculated.

### 2.10 Analysis of the ability of haplotype-tagging SNPs to capture previously unidentified genetic variations on tagged haplotypes

A web-based program called greedy_haplotype (http://www.well.ox.ac.uk/rmott/SNPS/webserver_frame.shtml), which

identifies SNP markers using a Maximum Entropy method was utilised to assess whether the haplotype tagging approach employed in this study captures additional variation on the tagged haplotypes which has not been previously identified (Ackerman *et al.* 2003). The Entropy maximisation method (EMM) identifies the percentage haplotype diversity represented by each SNP in the haplotype. EMM will identify which group of SNP loci is best able to define the total haplotype diversity, or structure, observed. In the present study, this program is useful for the examination of haplotypes composed of htSNPs and additional SNPs which had not been previously known, and hence not included in the original haplotype structure from which the htSNPs were selected. In this case, the EMM method will identify if the group of htSNP loci originally selected can define the total haplotype diversity present or whether the new SNPs add any useful information.

The user enters the haplotypes of interest (including the htSNPs and the previously unidentified SNP loci) and the haplotype frequencies into the program. For this present study, the haplotypes of interest are the TNF block haplotypes, which includes the selected htSNPs and three additional SNPs, TNF-863, TNF-857 and LTA+252, which had not been previously identified or included in the htSNP selection process (Chapter 5). Using this data, the program chooses a subset of markers which best approximate the haplotypic diversity in the population. The program generates an output which identifies the percentage haplotype diversity explained by subsets of SNPs of increasing size (n = 1,2,3,...m, where m = number of markers in haplotype). The program will initially add the SNP which contributes the highest percentage haplotype diversity into the SNP subset. Next, the program will include the SNP which contributes the second highest percentage of haplotype diversity and so on, until all SNPs have been added. This output will identify the smallest subset of SNPs which will explain 100% of haplotype diversity. Addition of extra SNPs to this subset will have no further increase in the percentage of haplotype diversity explained. Hence, this method will identify those SNPs which do not add extra information on haplotype diversity.

## 2.11 Genotyping of polymorphisms

### 2.11.1 Polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) analysis

The CD28int3+17 (rs3116496), CTLA4–658 (rs11571317), CTLA4-318 (rs5742909), CTLA4_CT60 (rs3087243) and CTLA4+49 (rs231775) polymorphisms were analysed using PCR-RFLP analysis. This method uses restriction enzymes to distinguish between allelic variants of a SNP where such a SNP either creates or destroys a restriction site. Digested fragments are separated by electrophoresis and visualised under UV light. The CD28int3+17 and CTLA4–318 polymorphisms were genotyped as previously described (King *et al.* 2002; Wood *et al.* 2002). The CTLA4–658 and CTLA4_CT60 polymorphisms were amplified using mismatch primers designed using PIRA software (Ke *et al.* 2001). The CTLA4+49 polymorphism was analysed using primers designed using the Primer Express program (v2.0, Applied Biosystems, CA, U.S.A.). All primer sequences were supplied by MWG Biotech, Germany and are listed in table A2.1 in Appendix II. For all PCR reactions listed above, 1 µl (approximately 100 ng) of DNA was mixed with 1X PCR buffer (Qiagen, Cologne, Germany), 12.5 µM of each appropriate primer, 200 µM of dNTP mix (Promega, WI, U.S.A.) and 0.5 units of HotStar Taq polymerase (Qiagen) in a 25 µl reaction. PCR reactions were performed in a Perkin-Elmer 2400 thermal cycler (Perkin-Elmer, UK). Thermal cycling conditions for each PCR assay consisted of an initial enzyme activation step at 95°C for 15 minutes, followed by a DNA denaturing step at 94°C for 45 seconds, an annealing step at varying temperatures for different assays for 30 seconds and an extension step at 72°C for 30 seconds. A total of 35 cycles were performed for each assay. The annealing temperatures for each assay are as follows; CTLA4-658 assay at 58°C, the CTLA4_CT60 assay at 58°C and the CTLA4+40 assay at 51°C (Table 2.2).

### 2.11.2 Restriction fragment length polymorphism (RFLP) analysis

For the CTLA4–658 assay, 3 units of *MspI* and 3.5 ul of 10X NE buffer 2 were added to the 25 µl PCR reaction product to a total volume of 35 µl. For CTLA4_CT60, 3 units of *NcoI* and 3.5 µl of 10X NE buffer 4 were added to the 25 µl PCR reaction product to a total volume of 35 µl. For CTLA4+49, 3

units of *BbvI* and 3.5 µl of 10X NE buffer 2 were added to the 25 µl PCR reaction product to a total volume of 35 µl (Table 2.2). (Enzymes and buffers supplied by New England Biolabs, MA, USA). All reactions were digested for 16 hours at 37°C in a Perkin-Elmer 2400 thermal cycler.

### 2.11.3 Agarose gel electrophoresis

PCR products were separated on 3% agarose gels. Gels were produced by adding 6 grams of agarose (Melford laboratories, UK) to 200ml of 1X TAE. This was boiled and allowed to cool slightly before 11 µl of ethidium bromide (Sigma Chemical Co, MO, U.S.A.) was added. The gel was poured into a large gel tray with inserted combs and allowed to polymerise at room temperature. 3 µl of loading dye was added to each PCR product sample and 20 µl of this mix was loaded onto the gel. 10 µl of a 100 bp DNA ladder (New England Biolabs, MA, USA) was also loaded. The gel was then electrophoresed in a 1X TAE buffer for 45 minutes at 120 volts. Gels were then visualised under UV light and genotypes were recorded.

### 2.11.4 Allelic discrimination using Taqman technology

The CTLA4-1661 and CTLA4-1722 SNP assays were genotyped using Taqman technology. Taqman allelic discrimination assays combine PCR amplification and detection with two fluorogenic probes, each with a sequence specific to either allele in a bi-allelic SNP. The assay is performed under competitive conditions where both probes are present in the same reaction tube. During the extension cycle of the PCR reaction, the DNA polymerase cuts the probe so that the quencher is released and the allele specific dye fluoresces. This procedure is outlined in figure 2.3.

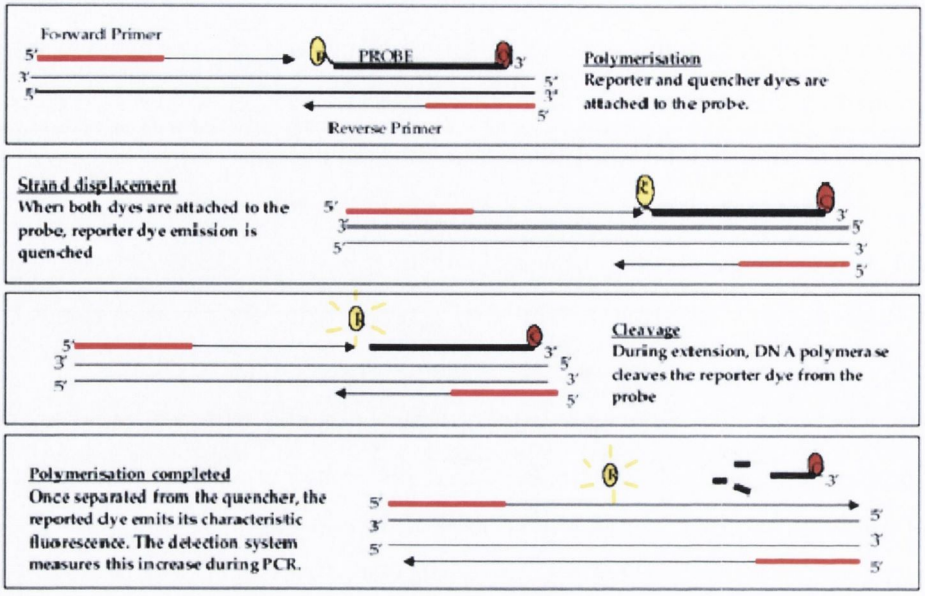



**Figure 2.3 Procedure for a Taqman allelic discrimination assay**

Primers and probes were designed and supplied by the Assay-by-Design service (Applied Biosystems). The primers used for the CTLA4–1661 and CTLA4-1722 SNP assays are listed in table A2.1 in Appendix II. The probes used in the CTLA4-1661 Taqman assay were 5'-TCA CTT TTT AAA AAA CC-3' and 5'-ACT TTT TGA AAA ACC-3'. For the CTLA4–1722 assay, the probes used were 5'-CAG GGA CAG ACA GCT A-3' and 5'-AGG GAC AGG CAG CTA-3'. Each sample assay consisted of approximately 10 ng of DNA, 6.25 µl of 2X Universal MasterMix (Applied Biosystems) and 0.3125 µl of 40X Assay-by-Design mix in a final volume of 12.5 µl. Fragments were amplified on the GeneAmp 9700 (Applied Biosystems) under standard thermal cycling conditions developed for Assay-by-Design assays (Initial Steps; 50°C for 2 minutes followed by 95°C for 10 minutes; a denaturation step at 92°C for 15 seconds and an annealing/extension step at 60°C for 1 minute. A total of 40 cycles were performed) and analysed on the ABI 7000 (Applied Biosystems).

### 2.11.5 Genotyping by Amplifluor

Additional genotyping was carried out commercially by Kbiosciences Ltd, UK. The technique used for genotyping was the Amplifluor SNPs genotyping system. This system is a single closed tube system utilizing two Amplifluor primers and three unlabeled primers. The Amplifluor primers consist of an excited fluorophore which is linked via an oligonucleotide to a complexed acceptor moiety that results in quenching of the fluorescence. A single common reverse primer and two SNP specific primers are designed to amplify across the SNP, each with a 5' tail that corresponds to one of the two Amplifluor Primers. As SNP-specific PCR products are generated, the primer sequence of the Amplifluor Primers hybridizes to the complement of the tail sequence on the newly generated PCR product and, in conjunction with the common reverse primer, amplifies the product further. Depending on which base is present in the target sequence, a FAM or a JOE fluorophore signal is generated. Heterozygotes result in signals from both fluorophores. The Amplifluor Primers and the common reverse primer are used in a 10-fold excess over the SNP specific tailed primers to ensure high yield of the fluorescent product. This procedure is outlined in figure 2.4.

Sequences containing 50 base pairs either side of the polymorphism of interest were identified as listed on the dbSNP website on the NCBI database and sent to the company to develop an assay for genotyping. Primer sequences for the developed assays are listed in table A2.2. As a routine quality control, approximately 9% of samples within each assay were retyped anonymously.

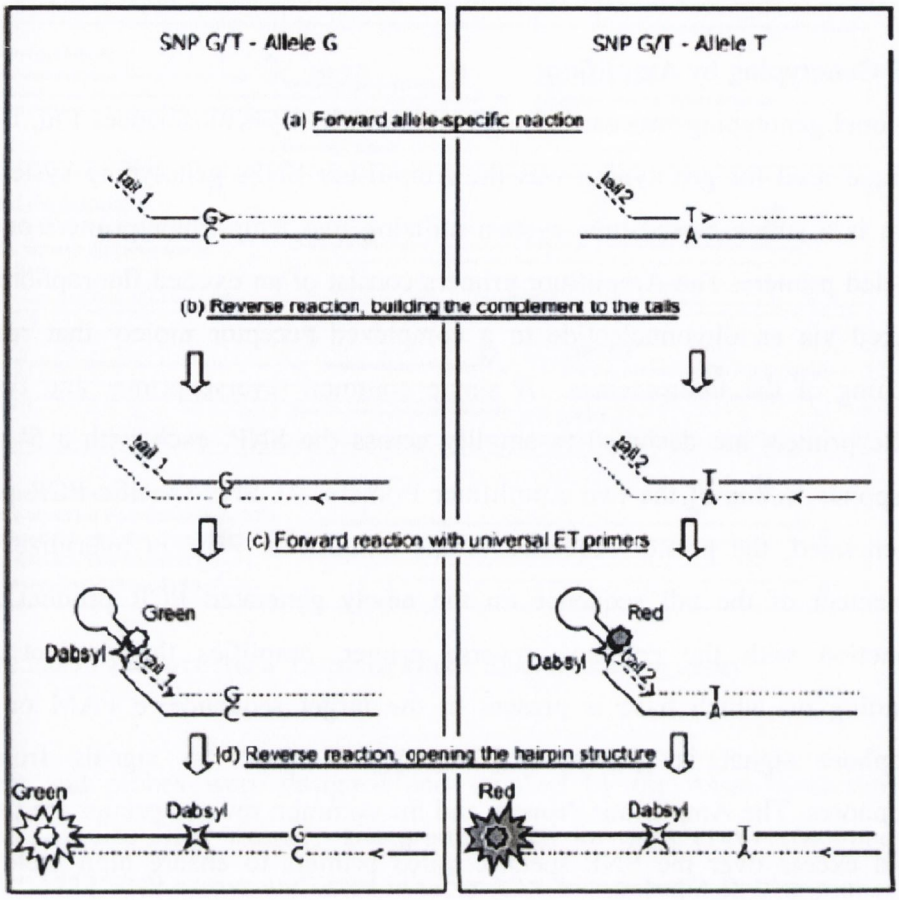



**Figure 2.4. Procedure for Amplifluor SNPs detection of a G/T genotype. (a) Allele-specific primers anneal to the target and are elongated by Taq Polymerase. Each primer has a unique 5' tail sequence that is identical to the priming domain of one of the Amplifluor SNPs Primers. (b) The common reverse primer anneals and is elongated by Taq Polymerase resulting in synthesis of the tail sequence complement. (c) Amplifluor SNPs Primers anneal specifically to the products of the reverse reaction and are elongated by Taq Polymerase. (d) During PCR, the hairpin structures of the Amplifluor SNPs Primers are unfolded and fluorescent signal are generated.** (Source: Chemicon International Company web site (http://www.chemicon.com))

**Table 2.2 PCR cycling conditions and restriction enzyme used in digestion for RFLP genotype assays**

| | IL10RA rs2229113 | ICOS +1323 | IkBL -62 | CTLA4 -658 | CTLA4 _CT60 | CTLA4 +49 |
|---|---|---|---|---|---|---|
| Enzyme activation step (15 min) | 95°C | 95°C | 95°C | 95°C | 95°C | 95°C |
| Denaturing step (45 secs) | 94°C | 94°C | 94°C | 94°C | 94°C | 94°C |
| Annealing step (30 secs) | 64°C | 64°C | 63°C | 58°C | 58°C | 51°C |
| Elongation step (30 secs) | 72°C | 72°C | 72°C | 72°C | 72°C | 72°C |
| Number of cycles | 30 | 30 | 30 | 35 | 35 | 35 |
| Final elongation (10 min) | 72°C | 72°C | 72°C | 72°C | 72°C | 72°C |
| Restriction enzyme (3 Units) | MspI | TaqI | BbvI | MspI | NcoI | BbvI |

### 2.11.6 Confirmation of Ampliflour genotyping results

Approximately 100 samples from three Ampliflour assays were retyped using PCR-RFLP analysis as a routine quality control measure. IL10RA_ rs2229113, ICOS+1323 and IkBL-62 were retyped. Primers used are listed in table A2.1 in Appendix II. All PCR-RFLP assays used standard reagent concentrations as described in 2.11.1. PCR cycling conditions and the restriction enzyme used in digestion are listed in table 2.2.

## 2.12 Data analysis

### 2.12.1 Genetic case-control differentiation analysis of allele, genotype and carrier status frequencies

Allelic, genotypic and carrier status frequencies were obtained using HITAGENE (http://www.hitagene.com) and also by direct counting. Comparison of frequency differences between patient and control groups was assessed using the $\chi^2$ test or Fishers exact test as implemented by HITAGENE and also by calculation on a 2 X 2 or 2 X 3 contingency table using the STATCALC program (Dean 1999).

### 2.12.2 Hardy-Weinberg Equilibrium (HWE) testing

Genotyping results were evaluated for HWE using the GENEPOP (http://wbiomed.curtin.edu.au/genepop/) program.

### 2.12.3 Haplotype differentiation analysis

Haplotype frequencies were estimated for the case and control groups separately using the expectation maximisation (EM) algorithm (Excoffier and Slatkin 1995) as implemented by HITAGENE (http://www.hitagene.com). Here, statistical differences between haplotype frequencies in the case and control groups were examined using p-values obtained by a permutation method as described by Zaykin *et al* (Zaykin *et al.* 2002) and carried out using the "haplotype association test" function in HITAGENE. The Stephens-Smith-Donnelly (SSD) method as implemented by PHASE v2.0.2 (Stephens *et al.* 2001; Stephens and Donnelly 2003) was also used to generate haplotypes. This package produced a global p-value to test for any significant difference between groups in the overall distribution of haplotypes frequencies. To test for any significant association between individual haplotypes

and disease, the frequency of each individual haplotype in the patient group was compared to its equivalent in the control group. The frequency of all other haplotypes were condensed to form a secondary group. Comparison of haplotype frequency differences between groups was then assessed using the $\chi^2$ test on a 2 X 2 contingency table as outlined in section 2.12.1. Results from both methods were compared for consistency.

### 2.12.4 Inference of allele frequencies of non-genotyped SNPs

Using the information available from two separate sources, the allele frequencies at non-genotyped SNPs on the haplotypes tagged by htSNPs were estimated in the patient and control groups. The two sources of information utilised were;

1. The frequencies of each of the haplotypes of interest, composed of the htSNPs in the patient and control populations in the present study.
2. Information on the haplotype structure at each gene/region of interest. This includes all SNPs in the haplotype, both htSNPs and all other non-genotyped SNPs as obtained from previous studies or genotyping databases.

Using the information on the haplotype structure in each gene/region of interest, the htSNP haplotypes along with their associated frequencies, as identified in the patient and control groups in the present study, are extended to inferred full haplotypes (i.e. haplotypes incorporating both the htSNPs and all non-genotyped loci). This is achieved as each particular htSNP combination is unique and representative of only one extended haplotype.

Each extended haplotype variant will contain one allelic variant (Allele 1 or Allele 2) at each non-genotyped SNP locus. To infer an estimated allele frequency at each of these, the frequencies of all haplotypes containing Allele 1 at the particular non-genotyped SNP of interest are grouped together in the patient group and then in the control group. Similarly, the frequencies of all haplotypes containing Allele 2 of the particular non-genotyped SNP of interest are grouped together. This method results in four estimated allele frequencies; Allele 1 and Allele 2 frequencies in the patient group, and, likewise, Allele 1 and Allele 2 frequencies in the control group. The

frequency differential between allele counts in the patient and control groups is tested here using a chi-square test or Fishers exact test on a 2 X 2 contingency table.

### 2.12.5 Gene-gene interaction effects

Gene-gene interaction effect analysis was performed using the crosstabulations method as carried out by SPSS version12.0 (SPSS Inc., IL, U.S.A.). Disease status was defined as the dependent variable and the SNP with the most significant disease association from each of the investigated chromosomal regions, 2q33, 11q23 and the TNFSF14 locus were defined as the variables of interest. Each 2-way interaction between these three variables (SNPs) was assessed for its significance to disease status.

The crosstabulation method is part of the descriptive statistics functions of SPSS. A crosstabulation displays the number of cases in each category defined by two or more grouping variables. This method is used to characterise how often a particular allele of a SNP occurs in combination with a particular allele of another SNP in the patients versus the control groups. Here, we are looking at the carrier status of the particular allele at each SNP which has been shown previously to be associated with disease. For example, if for a particular SNP, allele 1 was found to be associated with disease, when carrier status association testing was carried out then carrier status for allele 1 will be examined in the interaction model. This will include grouping individuals into two categories; Allele 1 carriers (includes individuals with genotypes 1,1 and 1,2) and non-allele 1 carriers (includes individuals with genotypes 2,2). In the crosstabulation function in SPSS, disease status is entered as a column variable, variable 1 (i.e. carrier status information for SNP 1) is entered as a row varaible, and variable 2 (i.e. carrier status information for SNP 2) is entered as a layer variable. SPSS then carries out a number of statistical tests to establish if there is an interaction effect present. The odds ratio for risk of disease for each combination of variables is calculated, and a test for the homogeneity of odds ratios across the groups is carried out, as measured by the Breslow-Day and Tarone's chi-square statistic. If there is disparity in the odds ratios, this suggests that the effect of one variable becomes stronger or weaker depending on the level of the other variable. This is a classic example of interaction between variables.

Logistic regression analysis was also used to examine for the presence of any gene-gene interaction effects. Logistic regression was performed using SPSS version 12 (SPSS Inc., IL, U.S.A). Disease status was used as the dependent variable and each individual SNP data and the SNP-SNP interaction term were used as independent categorical variables. Binary logistic regression was employed using the Enter method. This method builds the prediction equation for disease status and tests whether each term in the equation contributes to disease prediction using the Wald statistic. An associated p-value <0.05 for a variable suggests that this variable is a significant variable in the prediction model and is associated with the disease model. Analysis of the p-value associated with the gene-gene (SNP-SNP) interaction term in the prediction equation will assess if this variable adds significantly to disease outcome prediction. If the interaction term is found to be significant, there is reason to suggest that there is the involvement of gene-gene interaction effects in the disease model.

### 2.12.6 Linkage disequilibrium analysis

Linkage disequilibrium (LD) was analysed using D' values as generated by HAPLOVIEW and HITAGENE.

### 2.12.7 Statistical adjustment for multiple testing

Correction for multiple testing was performed using permutation testing. For each of 5,000 permutations, 394 subjects from the combined case and control populations were randomly labelled 'cases' and the remainder labelled 'controls' thus maintaining the overall genotype structure. Frequency differences in 'cases' and 'controls' were compared for each SNP and the minimum p-values calculated. We then compared the calculated p-values to the lower $5^{th}$ centile of minimum p-values among the permutations. This permutation analysis was carried out using Stata 8.2. by Dr. Anthony Fitzgerald (Haughton Institute, Trinity College; now at UCC).

In some instances, the most stringent adjustment for multiple correction, the Bonferroni correction, was carried out. Using this method, a test will only be deemed as having a statistically significant finding if the associated p-value is less than the original designated level of significance divided by the total number of tests carried out. (i.e. p-value < 0.05/n, where n = total number of tests carried out).

# CHAPTER III

## GENETIC ASSOCIATION ANALYSIS OF THE 2q33 REGION

### 3.1 Introduction

Chromosome region 2q33 (CELIAC3) has been associated with CD and contains the *CD28*, cytotoxic T-lymphocyte-associated antigen 4 (*CTLA4*) and inducible costimulator (*ICOS*) genes, all of which are important regulators of T-cell activity. CD28 and ICOS are involved in T-cell activation while CTLA4 is a negative regulator of activation, as it competes with CD28 for the same receptors, CD80 and CD86 (Carreno and Collins 2002). This implicates these genes as potential factors in an immune mediated disease, and makes this region an interesting candidate area for investigation.

A number of other autoimmune diseases have previously been associated with genetic variation at *CTLA4*, for example, type 1 diabetes (T1D) (Nistico *et al.* 1996), multiple sclerosis (MS) (Harbo *et al.* 1999) and Graves disease (GD) (Vaidya *et al.* 1999). Association between markers within or near the *CTLA4* gene and CD have been observed in a number of different populations (Djilali-Saiah *et al.* 1998; Holopainen *et al.* 1999; Naluai *et al.* 2000; King *et al.* 2002; Popat *et al.* 2002; Popat *et al.* 2002; Mora *et al.* 2003; Amundsen *et al.* 2004; van Belzen *et al.* 2004; Hunt *et al.* 2005), while other studies have failed to find any association between *CTLA4* and disease (Clot *et al.* 1999; Wijmenga *et al.* 2000; King *et al.* 2003; Martin-Pagola *et al.* 2003; Haimila *et al.* 2004). Linkage studies have also shown conflicting results, with evidence for (Holopainen *et al.* 1999; Naluai *et al.* 2000; Amundsen *et al.* 2004), and against (Clot *et al.* 1999; Wijmenga *et al.* 2000), observed in different populations. A large European study (Holopainen *et al.* 2004) recently observed nominally significant linkage and association to disease but were unable to conclusively identify the primary functional gene variant. However, multiple replications of association or linkage findings between the CTLA4 region and disease suggest that this is a genuine disease susceptibility locus.

A recent comprehensive investigation of the 2q33 region, which incorporated 108 SNPs, identified the primary association with a number of autoimmune diseases to be the CT60 (+6230G/A) polymorphism (Ueda *et al.* 2003). A number of studies have investigated CT60 for an association with coeliac disease. Results reveal a borderline significant association in two populations (van Belzen *et al.* 2004; Hunt *et al.* 2005) and no association in three others (King *et al.* 2003; Amundsen *et al.* 2004; Haimila

*et al.* 2004; van Belzen *et al.* 2004; Hunt *et al.* 2005). Few studies have investigated the association between *CD28* and *ICOS* single nucleotide polymorphisms and CD (Haimila *et al.* 2004; Hunt *et al.* 2005). Results have revealed a putative association was observed between four *ICOS* SNPs and disease in one population (Haimila *et al.* 2004).

In this study, we have used a haplotype tagging approach formulated from previous studies of genetic variation in the 2q33 region to investigate whether an association exists between this region and coeliac disease. We selected previously identified haplotype tagging SNPs (htSNPs) in *CTLA4* (Johnson *et al.* 2001), and generated our own htSNPs in the *ICOS* gene using haplotype data previously published (Haaning Andersen *et al.* 2003). We also selected three SNPs in *CD28* based on their location to obtain a haplotype that will incorporate a large section of the gene. Haplotype tagging SNPs were combined with several others in *CTLA4* and *ICOS* which have previously either been associated with disease or have possible functional genetic effects, thus allowing a comprehensive study of the region with the ability to detect all common haplotypes as previously defined for *CTLA4* and *ICOS*.

### 3.2 Objectives

The aims of this chapter were to:

1. Identify genes in the 2q33 region which are potential candidate genes in an autoimmune or immune mediated disease such as coeliac disease.
2. Select SNPs within these candidate genes which are most informative based on the available information, either in their ability to act as haplotype tagging SNPs (htSNPs) or their potential to alter either transcription factor binding sites or the amino acid sequence of the gene.
3. Test the accuracy of the htSNP method to tag all SNPs within a tagged haplotype.
4. Test for a statistically significantly allelic, genotypic, carrier status or haplotype association between polymorphisms in the 2q33 region and coeliac disease.
5. Analyse the structure of linkage disequilibrium (LD) in the region.
6. Compare different methods of haplotype reconstruction.
7. Test for haplotype association with disease.

8. Reconstruct non-genotyped SNPs from haplotype data.
9. To investigate any evolutionary conserved regions within 2q33 and to examine if selected polymorphisms occur within these regions

## 3.3 Results

### 3.3.1 Gene selection on chromosomal region 2q33

To identify all genes surrounding the *CTLA4* locus, previously shown to be associated with CD, 100 kb of sequence was searched for genes either side of the *CD28/CTLA4/ICOS* region using the Human ContigView option in the Ensembl (v32) genome browser. Results are shown in figure 3.1 and table 3.1. *CD28*, *CTLA4* and *ICOS* were the only known genes identified within this region, though there are some unidentified novel sequences which have also been identified in the region (as outlined in table 3.1). However, *CD28*, *CTLA4* and *ICOS* are the only genes present in the region which have a clear immunological role and hence, these three genes were selected as candidates in this study.

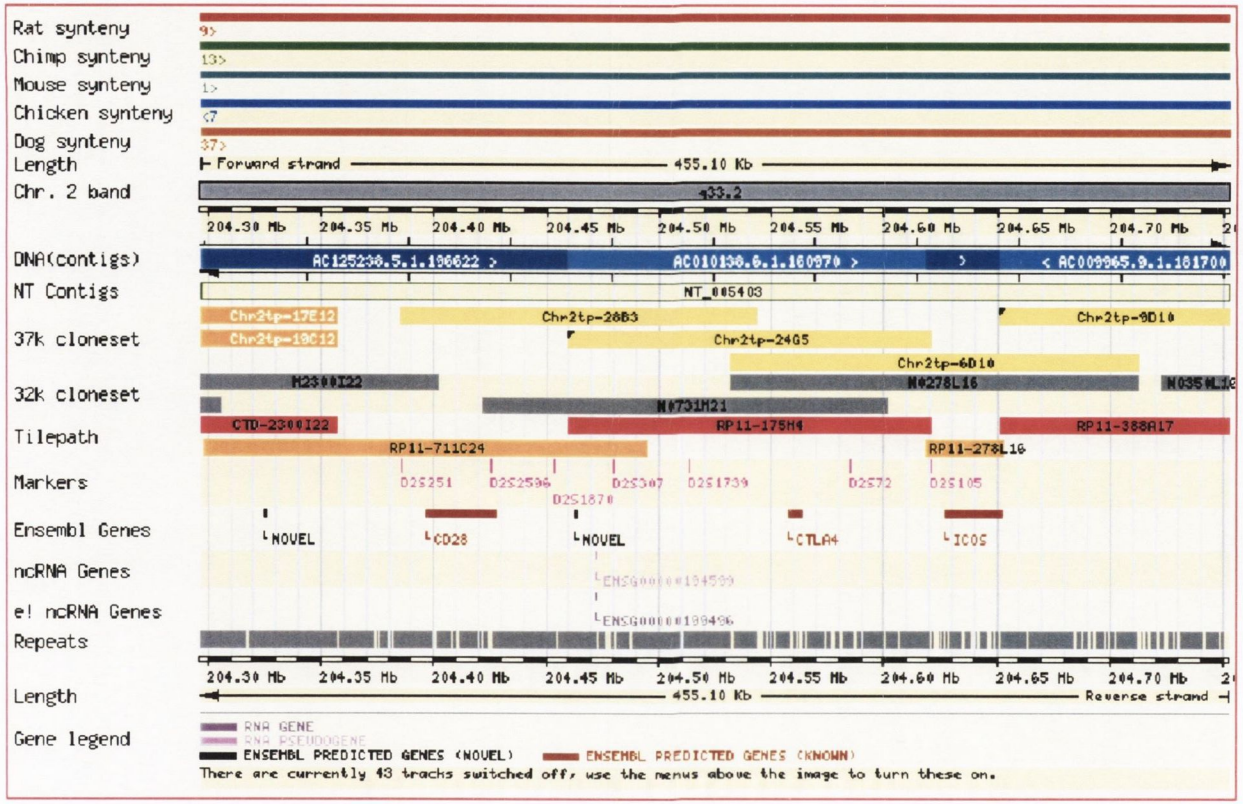



**Figure 3.1 Chromosome 2q33 information and genes as identified by Human ContigView in the Ensembl (v32) genome browser**

(This figure provides a graphical presentation of all the genomic information present in the Ensembl browser for the 2q33 region. Of interest is the row entitled Ensembl genes, which indicates that *CD28*, *CTLA4* and *ICOS* are the only known genes in this region).

**Table 3.1 Genes located 100 kb from the *CD28/CTLA4/ICOS* locus as identified by Human ContigView in Ensembl (v32)**

| Chromosome Name | Start Position (bp) | Band | External Gene ID | Description | Biotype | Confidence |
|---|---|---|---|---|---|---|
| 2 | 204324806 | q33.2 | | | protein_coding | NOVEL |
| 2 | 204396704 | q33.2 | CD28 | T-cell-specific surface glycoprotein CD28 precursor (TP44). [Source:Uniprot/SWISSPROT;Acc:P10747] | protein_coding | KNOWN |
| 2 | 204472266 | q33.2 | | | snRNA_pseudogene | NOVEL |
| 2 | 204472266 | q33.2 | | | snRNA | NOVEL |
| 2 | 204558017 | q33.2 | CTLA4 | Cytotoxic T-lymphocyte protein 4 precursor (Cytotoxic T-lymphocyte- associated antigen 4) (CTLA-4) (CD152 antigen). [Source:Uniprot/SWISSPROT;Acc:P16410] | protein_coding | KNOWN |
| 2 | 204626977 | q33.2 | ICOS | Inducible T-cell co-stimulator precursor (Activation- inducible lymphocyte immuno-mediatory molecule) (CD278 antigen). [Source:Uniprot/SWISSPROT;Acc:Q9Y6W8] | protein_coding | KNOWN |

### 3.3.2 SNP selection results

#### *3.3.2.1* CTLA4

Five SNPs were selected in the *CTLA4* gene as they had been previously identified by Johnson *et al.* as haplotype tagging SNPs (Johnson *et al.* 2001). Theses five SNPs are CTLA4-1722, CTLA4-1661, CTLA4-658, CTLA4-318, CTLA4+49 (The identification numbers (rs numbers) as outlined on the dbSNP database (http://www.ncbi.nlm.nih.gov) for all SNPs discussed in this study are listed in Appendix II). Figure 3.2 identifies the htSNPs (shaded in grey) and the haplotypes (frequencies >5%) which they tag. Other SNPs identified and located on the tagged haplotypes are also listed and their relative positions within *CTLA4* are shown in the diagram of the gene. CTLA4+6230 (commonly designated CTLA4_CT60) was also chosen as a candidate SNPs as it has been previously shown to be associated with a number of autoimmune disorders and has been implicated to have a functional role in splicing of the gene.

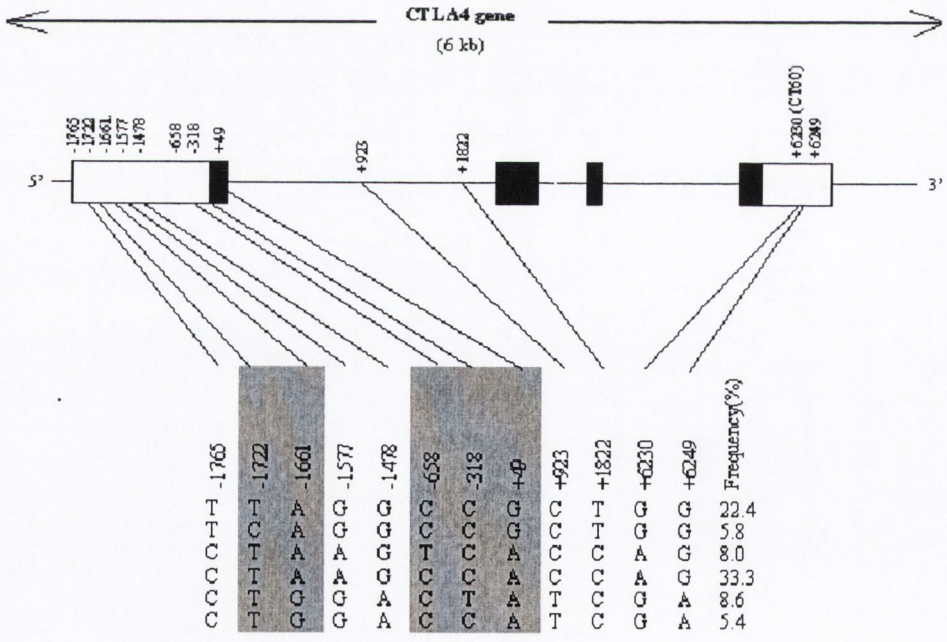



**Figure 3.2 CTLA4 haplotype tagging SNP selection as identified by Johnson *et al.* (2001). HtSNPs identified, are shaded in grey, and describe all haplotypes with a frequency >5%. All SNPs contained within the haplotypes are listed within their haplotype structure in the above table and their relative positions within the *CTLA4* gene are shown in the above diagram. The *CTLA4* gene is shown with exons identified by shaded boxes.**

### *3.3.2.2* ICOS

SNPs were also identified in the *ICOS* gene based on their ability to tag haplotypes within the gene. 16-SNP *ICOS* haplotypes had previously been identified by Haaning Andersen *et al.* (Haaning Andersen *et al.* 2003), and this information was used to identify htSNPs through the use of the SNPtagger program. 6 SNPs were identified to tag the 16-SNP haplotype; ICOS+1323, ICOS+2003, ICOS+2414, ICOS+3990, ICOS+4031, ICOS+5291. These htSNPs, their relative positions within the *ICOS* gene and the haplotypes (frequencies >5%) which they tag are shown in figure 3.3. Also included for analysis in this study were two polymorphisms identified in the same study (Haaning Andersen *et al.* 2003), which are located in the 5'UTR and which have been shown to reside in putative transcription factor binding sites; ICOS-2119 occurs in a consensus NFkB site and ICOS-2394 is located in a consensus Sp1 site.

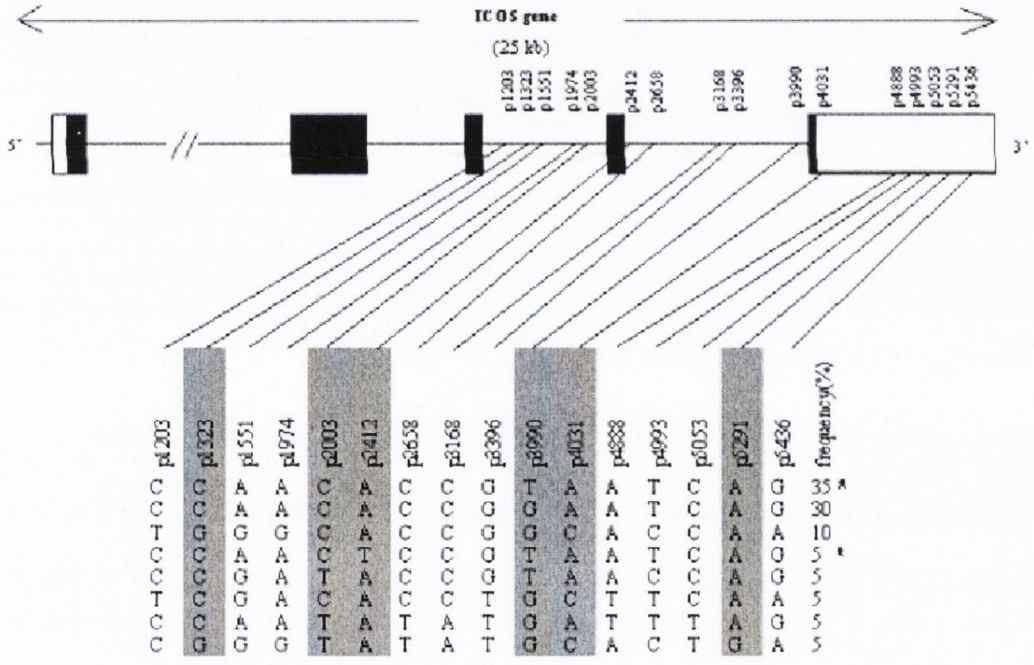



**Figure 3.3 ICOS haplotype tagging SNP selection as identified by Haaning *et al.* HtSNPs identified, are shaded in grey, and describe all haplotypes with a frequency >5%. All SNPs contained within the haplotypes are listed within their haplotype structure in the above table and their relative positions within the *ICOS* gene are shown in the above diagram. The *ICOS* gene is shown with exons identified by shaded boxes.**

#### *3.3.2.3* CD28

No haplotype or genotype information was available for the *CD28* gene at the time of this study. A previous study had examined the association between a polymorphism in the gene and type 1 diabetes, and this polymorphism (CD28int2+17) was included in this study. Examination of the dbSNP listings on the NCBI website identified a number of SNPs, all of which, with a minor allele frequency >5%, were intronic. Two SNPs; CD28_rs3769683 and CD28_rs3181113, were selected to assay the gene on the basis of relative separation of the three sites.

### 3.3.3 Haplotype-tagging SNP (htSNP) selection and accuracy testing.

#### *3.3.3.1* CTLA4

To identify the percentage error expected to occur in the reconstruction of total haplotypes from htSNP data, we employed the MARKER software package as described in section 2.9. HtSNPs were selected for the *CTLA4* gene based on a previous study by Johnson *et al.* (Johnson *et al.* 2001). We examined the percentage error rate attached when haplotypes were tagged with these 5 htSNPs (-1772, -1661, -658, -318, +49). The percentage error rates are shown in figure 3.4. Results (mean ± SEM) from five simulations with 0% missing data are shown in blue and results with up to 20% missing data are shown in red. Shaded SNPs indicate the htSNPs selected. With up to 20% missing data, error rates for all SNPs are 2.5% or less. 0% error rates are expected if no missing data occurs.

The influence of the addition of an extra SNP as an htSNP on the percentage (%) error rates of SNP frequency estimations was then examined to investigate if such an addition reduced the error rates by an appreciable amount. The CTLA4_CT60 SNP (also called +6230) was chosen as an htSNP. Results are shown in figure 3.5. With up to 20% missing data, all % errors are below 1%. 0% error rates for all SNPs are expected if no missing data occurs. This small reduction in error rates did not warrant the addition of an extra htSNP for haplotype tagging proposes. (In fact, CTLA4_CT60 is included as a candidate SNP in the present study. However, this SNP was chosen initially based on the possible functional effects associated with the polymorphism).

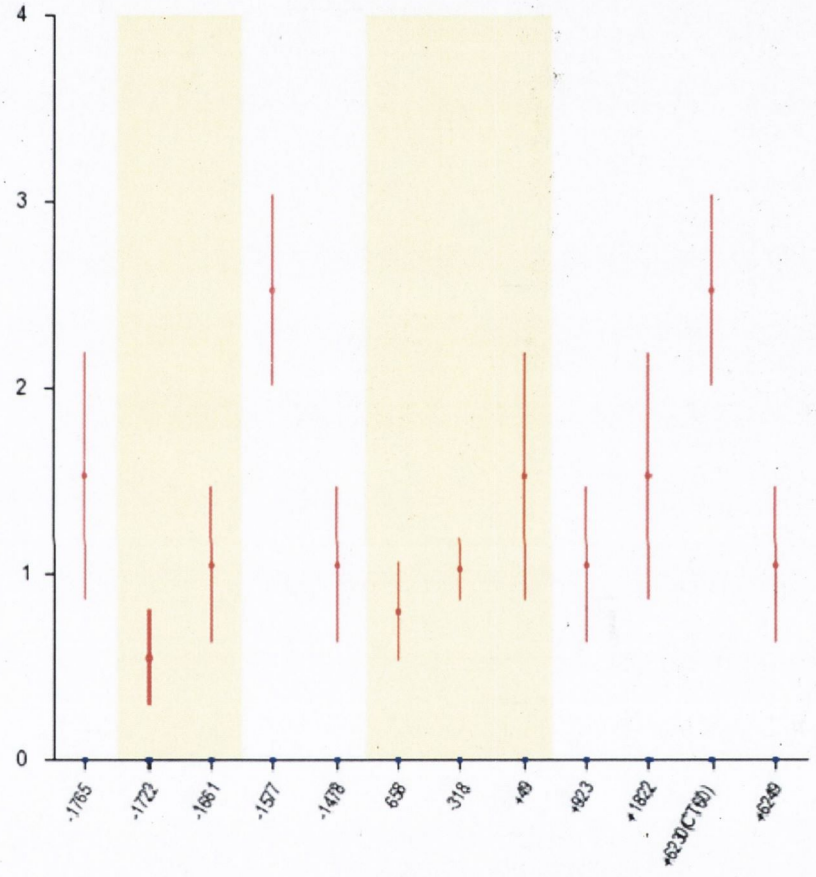



**Figure 3.4 Percentage error rates associated with reconstructing total haplotype information in the *CTLA4* gene using 5 original htSNPS (-1772, -1661, -658, -318, +49), (SNP loci on x-axis, % error rates on y-axis) Blue data points represent 0% missing data, red data points represent 20% missing data**

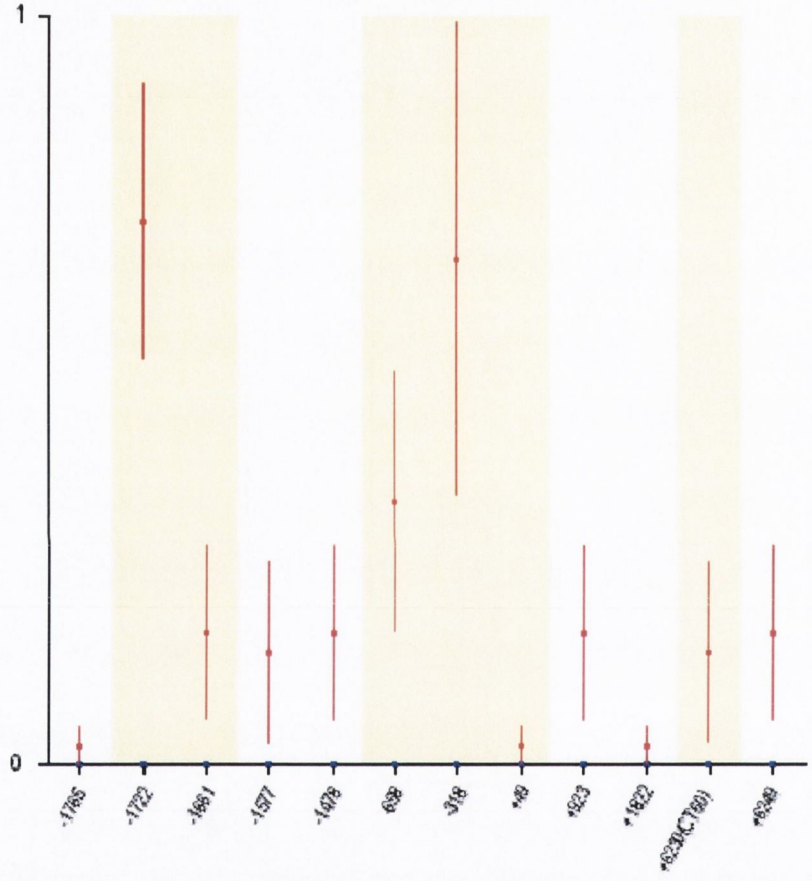



**Figure 3.5 Percentage error rates associated with reconstructing total haplotype information in the *CTLA4* gene using 5 original htSNPS (-1772, -1661, -658, -318, +49) and CTLA4_CT60 (Blue data points represent 0% missing data, red data points represent 20% missing data)**

To further investigate the reproducibility of the haplotype tagging method, haplotype frequencies for the *CTLA4* haplotypes were compared between the previous Johnson *et al.* study and the present study. Johnson *et al.* had generated haplotypes for the gene using 384 European samples and these were compared to the 354 control samples in the present study. Results are listed in table 3.2. In general, haplotype frequencies follow the same trend in both populations, however haplotype -1772T, -1661A, -658C, -318C, +49G, CT60G shows a significant difference between groups (Previous study 22.4%; present study 35.5% p<0.0001). This may be due in part to the high occurrence of the CTLA4+49 G allele in this population sample.

**Table 3.2 Comparison of haplotype frequencies between Johnson *et al* study and present control population**

| -1722 | -1661 | -658 | -318 | +49 | CT60 | Johnson *et al* freq n (%) | Present study freq n (%) | P-value | OR (95%CI) |
|---|---|---|---|---|---|---|---|---|---|
| T | A | C | C | A | A | 256 (33.3) | 247 (34.9) | 0.5292 | 1.07(0.86,1.34) |
| T | A | C | C | G | G | 172 (22.4) | 251 (35.5) | <0.0001 | 1.90(1.50,2.41) |
| T | G | C | T | A | G | 66 (8.6) | 43 (6.1) | 0.0644 | 0.69(0.45,1.04) |
| T | A | T | C | A | A | 61 (8.0) | 39 (5.5) | 0.0630 | 0.68(0.44,1.04) |
| C | A | C | C | G | G | 45 (5.8) | 60 (8.5) | 0.0509 | 1.49(0.98,2.26) |
| T | G | C | C | A | G | 41 (5.4) | 52 (7.4) | 0.1130 | 1.41(0.90,2.19) |

n = number of chromosomes; OR = odds ratio; 95%CI = 95% confidence intervals

### *3.3.3.2* ICOS

To identify the percentage errors expected to occur in the reconstruction of total haplotypes from htSNP data in the *ICOS* gene (haplotype and htSNP data as outlined in figure 3.3), the MARKER program was again utilised. Analysis was carried out in a similar manner to that carried out on the *CTLA4* data (section 3.3.3.1). No substantial reduction in error rates, which would warrant the addition of an extra htSNP was observed. The P4993 SNP was chosen as the extra SNP added to the htSNP set in this analysis (data not shown).

*ICOS* haplotype frequencies were also compared between the previous Haaning Andersen *et al* (2003) study and the present study. Results are shown in table 3.3. The ICOS+2414 SNP assay failed the genotyping process, hence, it was not possible to distinguish between haplotype +1323C, +2003C, +2414A, +3990T, +4031A, +5291A and haplotype +1323C, +2003C, +2414T, +3990T, +4031A, +5291A. These

two haplotype frequencies were grouped together and compared to the frequencies as described in the previous study. Haplotype frequencies follow a similar pattern in both studies, however, there is significant differences between the two groups. The previous study generated haplotypes based on 20 individuals, hence, the confidence intervals attendant to the haplotype frequency estimations are likely to be quite large. We have used up to 421 individuals to generate haplotype frequencies in this study, therefore, our estimates are likely to more accurately reflect the actual frequency in the sample population.

**Table 3.3 Comparison of haplotype frequencies between Haaning Andersen *et al* study and present control population**

| p1323 | p2003 | p2412 | p3990 | p4031 | p5291 | Haaning *et al* freq n (%) | Present Study freq n (%) | P-value | OR (95%CI) |
|---|---|---|---|---|---|---|---|---|---|
| C | C | A | T | A | A | 16(35) | 242(34.2) | 0.4513 | 0.78 (0.39,1.57) |
| C | C | T | T | A | A | 5 | | | |
| C | C | A | G | A | A | 12(30) | 297(41.9) | 0.1354 | 1.69 (0.81,3.57) |
| G | C | A | G | C | A | 4(10) | 113(15.9) | 0.3127 | 1.71 (0.59,6.74)* |
| C | T | A | T | A | A | 2(5) | 0(0) | 0.0031 | 0.00 (0.00,0.31)* |
| C | C | A | G | C | A | 2(5) | 0(0) | 0.0031 | 0.00 (0.00,0.31)* |
| C | T | A | G | A | A | 2(5) | 0(0) | 0.0031 | 0.00 (0.00,0.31)* |
| G | T | A | G | C | G | 2(5) | 56(7.9) | 0.7613 | 1.63 (0.40,14.31)* |

n = number of chromosomes; OR = odds ratio; 95%CI = 95% confidence intervals; * = Fishers exact test and exact confidence intervals used

### 3.3.4 Power calculations

Power calculations were carried out to assess the power of the present study to detect a disease effect at loci with varying genotype relative risks (1.3 to 2.0). Results are shown in table 3.4. The frequency of the disease allele was set at 10%, CD prevalence was set at 1% and the required level of statistical significance, $\alpha$, was 0.05. Power was calculated for the 394 cases available in this study with a control:cases ratio of 1. These results show that the present study has greater than 75% power to detect a disease locus with a relative risk of 1.6 or greater. (This program states that genotypic odds ratios for the 'Aa' and 'AA' genotypes will be very similar to the genotypic relative risks for rare diseases. In this study, genotype odds ratios as calculated by the Genetic Power Calculator were identical to the genotype relative risks for CD corrected to two decimal places (data not shown)).

The number of cases necessary to provide 80% power to successfully detect a disease effect at $\alpha = 0.05$ was also calculated for the various genotype relative risks (table 3.4). To detect a disease locus with a relative risk (or odds ratio) of 1.3, using the conditions specified above, 1441 cases and controls would be necessary, while a disease locus with a relative risk (or odds ratio) of 2.0 would be detectable using only 192 cases and controls. It should be noted that these calculations are strongly influenced by allele frequencies which can not be predicted *a priori*.

**Table 3.4 Power calculations of various genotype relative risks in the present study**

| Relative Risk* | Power+ % | Number of cases for 80% power |
|---|---|---|
| 1.3 | 31.1 | 1441 |
| 1.4 | 47.3 | 863 |
| 1.5 | 63.2 | 587 |
| 1.6 | 76.3 | 432 |
| 1.7 | 85.9 | 335 |
| 1.8 | 92.2 | 271 |
| 1.9 | 95.9 | 226 |
| 2.0 | 98.0 | 192 |

* relative risk of genotypes AA and Aa relative to baseline genotype aa

+ % power for 394 cases and controls

### 3.3.5 Hardy-Weinberg Equilibrium (HWE) Testing

All 16 SNPs in this region were tested for compliance with HWE using GENEPOP. Results are listed in table 3.5. After correction for multiple testing, only the CTLA4-658 locus, in patients, is shown to be out of HWE.

**Table 3.5 Hardy Weinberg Equilibrium Analysis results for all SNPs in 2q33**

| SNP | Patient group p-value | Control group p-value |
|---|---|---|
| CD28_rs3769683 | 0.0184 | 1.0000 |
| CD28int3+17 | 0.0677 | 1.0000 |
| CD28_rs3181113 | 1.0000 | 1.0000 |
| CTLA4-1722 | 1.0000 | 0.7050 |
| CTLA4-1661 | 0.5720 | 0.3887 |
| CTLA4-658 | <0.0001 | 1.0000 |
| CTLA4-318 | 0.6939 | 0.3736 |
| CTLA4+49 | 0.7606 | 0.7376 |
| CTLA4_CT60 | 1.0000 | 0.0546 |
| ICOS-2394 | 0.8749 | 1.0000 |
| ICOS-2119 | 0.7156 | 0.6263 |
| ICOS+1323 | 0.7086 | 0.5363 |
| ICOS+2003 | 0.7126 | 1.0000 |
| ICOS+3990 | 0.1368 | 0.6324 |
| ICOS+4031 | 1.0000 | 0.8826 |
| ICOS+5291 | 0.1492 | 0.2388 |

### 3.3.6 Allele, Genotype and Carrier Status frequencies and disease association testing in genotyped polymorphisms

All polymorphic sites in *CTLA4*, *ICOS* and *CD28* were analysed for allelic, genotypic and carrier status association with disease individually and results are listed in tables 3.6 (genotype and allele associations) and 3.7 (carrier status associations). The p-values and odds ratios for the allele frequency differentials between patient and control groups are displayed graphically in figure 3.6. Four polymorphisms (CTLA4-1661, CTLA4-658, CTLA4_CT60 and CD28_rs3769683) showed an association with disease either at allele frequency, genotype frequency or carrier status level prior to correction for multiple testing. CTLA4-1661 displayed an association with disease (G allele: patients 16.6% vs controls 12.3%; p=0.0315). This polymorphism also displayed nominal significance at the carrier status level (A allele carrier status and conversely homozygote status for the minor G allele) (patients 3.1% vs controls 0.7%; p=0.0384). CTLA4-658 was also associated with disease, at the level of allele frequency (patients 8.5% vs controls 5.4%; p=0.0263), genotype frequency (p=0.0066) and carrier status (C allele carrier status and/or homozygote status for the T allele; patients 3.8% vs controls 0.03%; p=0.0016).

CTLA4_CT60 also displayed significant association, at the allele (p=0.0259),

genotype (p=0.0312) and carrier status level (p=0.0085). Here, analysis of carrier status indicated that, disease association lies with a dominant A allele or recessive G allele model (patients 41.5% vs controls 31.7%). CD28_rs3769683 showed a significant association with carrier status for the G allele and recessive status for the A allele (patients 4.2% vs controls 1.4%; p=0.0262). A permutation analysis, using the minimum p-value across all SNPs as the test statistic, was used to allow for multiple testing and to assess the overall significance of the association between SNPs and disease (described in section 2.12.7). There was a significant association between carrier status and coeliac disease; only 58 of the 5,000 simulations (1.16%) yielded a minimum p-value less than that for the observed data. The association between genotype and disease was also statistically significant with 227 of the 5,000 simulations (4.52%) yielding a test statistic less than the observed data test statistic. The CTLA4–658 C allele carrier status was the only polymorphism with a p-value below the corresponding $5^{th}$ percentile for the permutations (0.0069 for carrier status and 0.0034 for genotypes). CTLA4+49 was not found to be independently associated in this study.

**Figure 3.6 P-values and odds ratios for allele frequency differences between patient and control groups for SNP loci at 2q33**

Actual allele frequencies are listed in table 3.6

**Table 3.6 Genotype and allele frequencies in the coeliac disease patient and control populations**

| | | Genotype n (%) | | | | Allele n (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | p | $A_1$ | $A_2$ | p | OR | 95% CI |
| CTLA4 -1722T/C | Patient | 258 (83.2) | 50 (16.1) | 2 (0.6) | 1.0000 | 566 (91.3) | 54 (8.7) | 0.9542 | 0.99 | (0.64,1.52) |
| | Control | 222 (83.1) | 44 (16.5) | 1 (0.3) | | 488 (91.4) | 46 (8.6) | | | |
| CTLA4 -1661A/G | Patient | 267 (69.9) | 103 (27.0) | 12 (3.1) | 0.0511 | 637 (83.4) | 127 (16.6) | **0.0315** | 0.70 | (0.50,0.98) |
| | Control | 204 (76.1) | 62 (23.1) | 2 (0.7) | | 470 (87.7) | 66 (12.3) | | | |
| CTLA4 -658C/T | Patient | 316 (86.8) | 34 (9.3) | 14 (3.8) | **0.0066** | 666 (91.5) | 62 (8.5) | **0.0263** | 0.62 | (0.39,0.97) |
| | Control | 288 (89.4) | 33 (10.2) | 1 (0.4) | | 609 (94.6) | 35 (5.4) | | | |
| CTLA4 -318C/T | Patient | 250 (85.0) | 42 (14.3) | 2 (0.7) | 0.7260 | 542 (92.2) | 46 (7.8) | 0.7224 | 0.92 | (0.57,1.48) |
| | Control | 218 (85.5) | 37(14.5) | 0 (0) | | 473 (92.7) | 37 (7.3) | | | |
| CTLA4 +49A/G | Patient | 106 (28.2) | 191 (50.8) | 79 (21.0) | 0.7320 | 403 (53.6) | 349 (46.4) | 0.4355 | 0.92 | (0.74,1.14) |
| | Control | 105 (30.3) | 175 (50.6) | 66 (19.1) | | 385 (55.6) | 307 (44.4) | | | |
| CTLA4 _CT60G/A | Patient | 140 (41.5) | 155 (46.0) | 42 (12.5) | **0.0312** | 435 (64.5) | 239 (35.5) | **0.0259** | 1.29 | (1.02,1.61) |
| | Control | 105 (31.7) | 178 (53.8) | 48 (14.5) | | 388 (58.6) | 274 (41.4) | | | |
| ICOS -2394T/C | Patient | 243 (63.0) | 126 (32.6) | 17 (4.4) | 0.6064 | 612 (79.3) | 160 (20.7) | 0.3200 | 0.88 | (0.67,1.14) |
| | Control | 232 (66.1) | 107 (30.5) | 12 (3.4) | | 571 (81.3) | 131 (18.7) | | | |
| ICOS -2119C/T | Patient | 314 (84.4) | 57 (15.3) | 1 (0.3) | 0.2026 | 685 (92.1) | 59 (7.9) | 0.1875 | 0.73 | (0.47,1.12) |
| | Control | 300 (87.7) | 42 (12.3) | 0 (0) | | 642 (93.9) | 42 (6.2) | | | |

| | | Genotype n (%) | | | | Allele n (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | p | $A_1$ | $A_2$ | p | OR | 95% CI |
| ICOS +1323C/G | Patient | 254 (69.6) | 100 (27.4) | 11 (3.0) | 0.5640 | 608 (83.3) | 122 (16.7) | 0.5480 | 0.92 | (0.68,1.23) |
| | Control | 239 (70.7) | 93 (27.5) | 6 (1.8) | | 571 (84.5) | 105 (15.5) | | | |
| ICOS +2003C/T | Patient | 320 (85.1) | 55 (14.6) | 1 (0.3) | 0.8500 | 695 (92.4) | 57 (7.6) | 0.8185 | 1.05 | (0.70,1.57) |
| | Control | 295 (84.8) | 51 (14.7) | 2 (0.6) | | 641 (92.1) | 55 (7.9) | | | |
| ICOS +3990G/T | Patient | 176 (45.8) | 158 (41.1) | 50 (13.0) | 0.7772 | 510 (66.4) | 258 (33.6) | 0.8352 | 1.02 | (0.82,1.28) |
| | Control | 151 (44.0) | 150 (43.7) | 42 (12.2) | | 452 (65.9) | 234 (34.1) | | | |
| ICOS +4031A/C | Patient | 206 (54.4) | 147 (38.8) | 26 (6.9) | 0.5971 | 559 (73.7) | 199 (26.3) | 0.3214 | 0.89 | (0.69,1.13) |
| | Control | 200 (57.5) | 129 (37.1) | 19 (5.5) | | 529 (76.0) | 167 (24.0) | | | |
| ICOS +5291A/G | Patient | 329 (84.8) | 59 (15.2) | 0 (0) | 0.7263 | 717 (92.4) | 59 (7.6) | 0.7371 | 0.93 | (0.62,1.41) |
| | Control | 294 (85.7) | 49 (14.3) | 0 (0) | | 637 (92.9) | 49 (7.1) | | | |
| CD28 -1G/A | Patient | 280 (72.7) | 89 (23.1) | 16 (4.2) | 0.0747 | 649 (84.3) | 121 (15.7) | 0.0810 | 0.77 | (0.57,1.04) |
| | Control | 268 (76.4) | 78 (22.2) | 5 (1.4) | | 614 (87.5) | 88 (12.5) | | | |
| CD28 -2T/C | Patient | 277 (76.1) | 76 (20.9) | 11 (3.0) | 0.2312 | 630 (86.5) | 98 (13.5) | 0.3128 | 1.19 | (0.84,1.67) |
| | Control | 162 (71.1) | 61 (26.8) | 5 (2.2) | | 385 (84.4) | 71 (15.6) | | | |
| CD28 -3G/T | Patient | 342 (92.2) | 29 (7.8) | 0 (0) | 0.6605 | 713 (96.1) | 29 (3.0) | 0.6666 | 0.89 | (0.49,1.59) |
| | Control | 321 (93.0) | 24 (7.0) | 0 (0) | | 666 (96.5) | 24 (3.5) | | | |

$A_1$ = Allele 1; $A_2$ = Allele 2; n = number of individuals; p = p-value, CI = confidence interval, CD28-1 = CD28_rs3769683; CD28-2 = CD28int3+17; CD28-3 = CD28_rs3181113

**Table 3.7 Carrier status frequencies in the coeliac disease patient and control populations**

| | | Allele $A_1$ Carrier Status | | | | | Allele $A_2$ Carrier Status | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1/A_1A_2$ | $A_2A_2$ | p | OR[(a)] | 95% CI | $A_1A_1$ | $A_1A_2/A_2A_2$ | p | OR[(b)] | 95% CI |
| CTLA4 -1722 | Patient | 308 (99.4) | 2 (0.6) | 0.5956 | 0.42 | (0.01,8.13)* | 258 (83.2) | 52 (16.8) | 0.9796 | 0.99 | (0.63,1.58) |
| | Control | 366 (99.6) | 1 (0.4) | | | | 222 (83.1) | 45 (16.9) | | | |
| CTLA4 -1661 | Patient | 370 (96.9) | 12 (3.1) | **0.0384** | 0.23 | (0.02,1.06) | 267 (69.9) | 115 (30.1) | 0.0804 | 1.37 | (0.95,1.99) |
| | Control | 266 (99.3) | 2 (0.7) | | | | 204 (76.1) | 64 (23.9) | | | |
| CTLA4 -658 | Patient | 350 (96.2) | 14 (3.8) | **0.0016** | 0.08 | (0.00,0.52) | 316 (86.8) | 48 (13.2) | 0.2897 | 1.29 | (0.79,2.11) |
| | Control | 321 (99.7) | 1 (0.3) | | | | 288 (89.4) | 34 (10.3) | | | |
| CTLA4 -318 | Patient | 292 (99.3) | 2 (0.7) | 0.5016 | 0.00 | (0.00,6.14)* | 250 (85.0) | 44 (15.0) | 0.8805 | 1.04 | (0.63,1.71) |
| | Control | 255 (100.0) | 0 (0.0) | | | | 218 (85.56) | 37 (14.5) | | | |
| CTLA4 +49 | Patient | 297 (79.0) | 79 (21.0) | 0.5164 | 0.89 | (0.60,1.30) | 106 (28.2) | 270 (71.8) | 0.5245 | 1.11 | (0.79,1.55) |
| | Control | 280 (81.0) | 66 (19.0) | | | | 105 (30.3) | 241 (69.7) | | | |
| CTLA4 _CT60 | Patient | 295 (87.5) | 42 (12.5) | 0.4404 | 1.19 | (0.75,1.90) | 140 (41.5) | 197 (58.5) | **0.0085** | 0.65 | (0.47,0.91) |
| | Control | 283 (85.5) | 48 (14.5) | | | | 105 (31.7) | 226 (68.3) | | | |
| ICOS -2394 | Patient | 369 (95.6) | 17 (4.4) | 0.4920 | 0.77 | (0.34,1.72) | 243 (63.0) | 143 (37.0) | 0.3732 | 1.15 | (0.84,1.57) |
| | Control | 339 (96.6) | 12 (3.4) | | | | 232 (66.1) | 119 (33.9) | | | |
| ICOS -2119 | Patient | 371 (99.7) | 1 (0.3) | 1.0000 | 0.00 | (0.00,42.42)* | 314 (84.4) | 58 (15.6) | 0.2029 | 1.32 | (0.84,2.07) |
| | Control | 342 (100.0) | 0 (0.0) | | | | 300 (87.7) | 42 (12.3) | | | |

| | | Allele $A_1$ Carrier Status | | | | | Allele $A_2$ Carrier Status | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1/A_1A_2$ | $A_2A_2$ | p | OR[(a)] | 95% CI | $A_1A_1$ | $A_1A_2/A_2A_2$ | p | OR[(b)] | 95% CI |
| ICOS +1323 | Patient | 354 (97.0) | 11 (3.0) | 0.2855 | 0.58 | (0.17,1.74)* | 254 (69.6) | 111 (30.4) | 0.7456 | 1.05 | (0.75,1.48) |
| | Control | 332 (98.2) | 6 (1.8) | | | | 239 (70.7) | 99 (29.3) | | | |
| ICOS +2003 | Patient | 375 (99.7) | 1 (0.3) | 0.6105 | 2.17 | (0.11,128.22)* | 320 (85.1) | 56 (14.9) | 0.8994 | 0.97 | (0.64,1.49) |
| | Control | 346 (99.4) | 2 (0.6) | | | | 295 (84.8) | 53 (15.2) | | | |
| ICOS +3990 | Patient | 334 (87.0) | 50 (13.0) | 0.7534 | 0.93 | (0.59,1.48) | 176 (45.8) | 208 (54.2) | 0.6243 | 0.93 | (0.69,1.26) |
| | Control | 301 (87.8) | 42 (12.2) | | | | 151 (44.0) | 192 (56.0) | | | |
| ICOS +4031 | Patient | 353 (93.1) | 26 (6.9) | 0.4338 | 0.78 | (0.41,1.50) | 206 (54.4) | 173 (45.6) | 0.3977 | 1.13 | (0.84,1.57) |
| | Control | 329 (94.5) | 19 (5.5) | | | | 200 (57.5) | 148 (42.5) | | | |
| ICOS +5291 | Patient | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| | Control | N/A | N/A | | | | N/A | N/A | | | |
| CD28 -1 | Patient | 369 (95.8) | 16 (4.2) | 0.0262 | 0.33 | (0.09,0.97) | 280 (72.7) | 105 (27.3) | 0.2599 | 1.21 | (0.86,1.71) |
| | Control | 346 (98.6) | 5 (1.4) | | | | 268 (76.4) | 83 (23.6) | | | |
| CD28 -2 | Patient | 353 (97.0) | 11 (3.0) | 0.5450 | 0.72 | (0.19,2.28)* | 277 (76.1) | 87 (23.9) | 0.1723 | 0.77 | (0.52,1.14) |
| | Control | 223 (97.8) | 5 (2.2) | | | | 162 (71.1) | 66 (28.9) | | | |
| CD28 -3 | Patient | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| | Control | N/A | N/A | | | | N/A | N/A | N/A | N/A | |

a; Odds ratio for comparison of $A_1A_1$/ $A_1A_2$ to $A_2A_2$ , b; Odds ratio for $A_1A_1$/ $A_1A_2$ compared to $A_1A_1$; * indicates that Fisher's exact test and exact confidence intervals were used.

### 3.3.7 Linkage Disequilibrium Analysis

Linkage disequilibrium (LD) analysis was carried out in the 2q33 region incorporated by the 16 SNP markers. This area shows a block of strongly linked SNPs in the *ICOS* gene, while the remainder of the region appears to form a haplotype extending from the *CD28* gene up to and including the *ICOS* promoter SNPs which is characterised by lower levels of inter SNP LD, an effect which may be maximized by the selection of haplotype tagging markers. Thus a clear LD boundary exists between the ICOS-2119 and ICOS+1323 loci. This is outlined in figure 3.7.

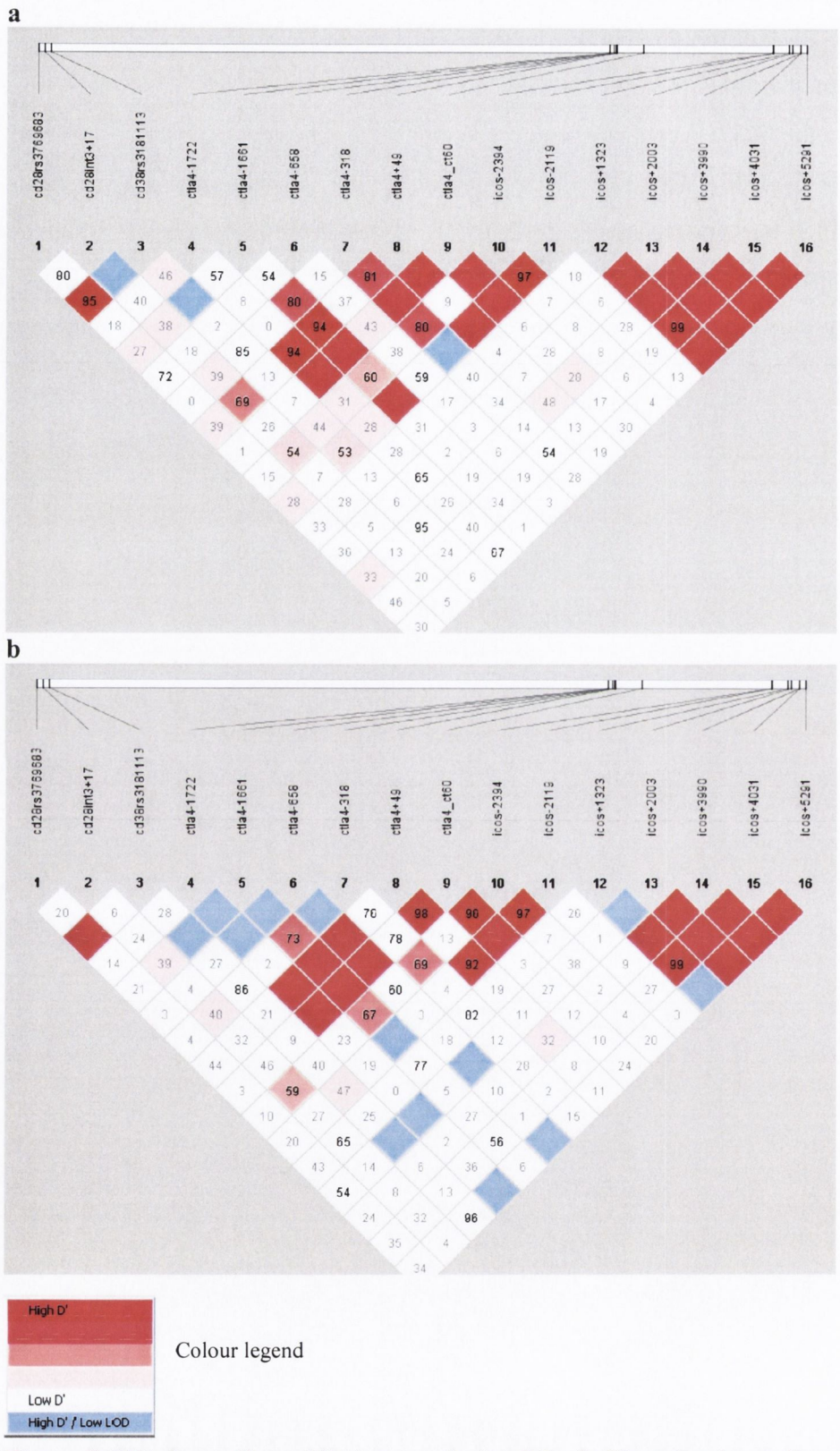



**Figure 3.7 Linkage disequilibrium analysis of the 2q33 region in coeliac disease cases (a) and controls (b)**

### 3.3.8 Haplotype Analysis of the 2q33 region

Haplotype analysis was centered on a gene-based, functional (promoter) or haplotype-tagging SNP approach. Haplotypes were generated using two different methods for haplotype reconstruction; the expectation maximization (EM) algorithm as implemented by the HITAGENE software and the Stephen-Smith-Donnelly (SSD) method as implemented by PHASEv2.0.2. 95% confidence intervals for the estimated haplotype frequencies generated from both methods were compared and were not found to be different for the haplotypes analysed. For reasons of clarity, the results presented here are from one method only, the EM method as implemented by HITAGENE. Haplotypes reconstructed from all sixteen SNP loci spanning the three genes of interest did not reveal any association with disease. Interestingly, only a comparison of haplotypes constructed from *CD28-CTLA4-5'ICOS* markers approached statistical significance for haplotypic frequency differentiation between coeliac cases and controls. Two haplotypes constructed from these markers had frequencies above 5%, one of which, (Hap1; GTGTACCAATC) has a significant frequency differential between cases and controls (0.2405 vs. 0.3154 respectively; p=0.033, table 3.8). Haplotype differentiation analysis using the PHASEv2.0.2 program also revealed a significant difference between the patient and control populations for this haplotype (p-value = 0.0061). Analysis of the different regions within this haplotype involved analysis of haplotypes containing all *CTLA4* SNPs, ht*CTLA4* SNPs, promoter *CTLA4* SNPs, all *ICOS* SNPs, ht*ICOS* SNPs, promoter *ICOS* SNPs, and *CD28* SNPs. None of these smaller haplotypes was significantly associated with disease status at the 5% level. Results are shown in tables 3.8 to 3.15. Haplotypes consisting of the CTLA4+49 and CTLA4_CT60 SNPs only were also investigated, as these have both been widely investigated and proposed to have a functional effect on disease susceptibility. This haplotype was not associated with disease status.

**Table 3.8 LD block haplotype association testing and frequency estimations (>5%) for patient and control populations**

| cd28rs3769683 | cd28int3+17 | cd28rs3181113 | ctla4-1722 | ctla4-1661 | ctla4-658 | ctla4-318 | ctla4+49 | ctla4_ct60 | icos-2394 | icos-2119 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | Patient | Control |
| G | T | G | T | A | C | C | G | G | T | C | 0.4066 | 0.3716 | 0.2901 | 0.2718 |
| G | T | G | T | A | C | C | A | A | T | C | 0.0453 | 0.0330 | 0.2405 | 0.3153 |

**Table 3.9 *CTLA4* haplotype association testing and frequency estimations (>5%) for patient and control populations**

| ctla4-1722 | ctla4-1661 | ctla4-658 | ctla4-318 | ctla4+49 | ctla4_ct60 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Patient | Control |
| T | A | C | C | G | G | 0.6755 | 0.6451 | 0.3510 | 0.3552 |
| T | A | C | C | A | A | 0.0855 | 0.0627 | 0.3090 | 0.3488 |
| T | G | C | C | A | G | 0.1599 | 0.1537 | 0.0898 | 0.0735 |
| C | A | C | C | G | G | 0.5186 | 0.5237 | 0.0758 | 0.0852 |
| T | G | C | T | A | G | 0.4620 | 0.4692 | 0.0657 | 0.0608 |
| T | A | T | C | A | A | 0.6929 | 0.6815 | 0.0522 | 0.0549 |

**Table 3.10 Ht*CTLA4* haplotype association testing and frequency estimations (>5%) for patient and control populations**

| ctla4-1722 | ctla4-1661 | ctla4-658 | ctla4-318 | ctla4+49 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Patient | Control |
| T | A | C | C | G | 0.7385 | 0.7163 | 0.3480 | 0.3473 |
| T | A | C | C | A | 0.1056 | 0.0857 | 0.3189 | 0.3676 |
| T | G | C | C | A | 0.1409 | 0.1370 | 0.0903 | 0.0700 |
| C | A | C | C | G | 0.4868 | 0.4841 | 0.0744 | 0.0850 |
| T | G | C | T | A | 0.6818 | 0.6721 | 0.0637 | 0.0570 |
| T | A | T | C | A | 0.8093 | 0.8112 | 0.0536 | 0.0556 |

**Table 3.11 Promoter *CTLA4* haplotype association testing and frequency estimations (>5%) for patient and control populations**

| ctla4-1722 | ctla4-1661 | ctla4-658 | ctla4-318 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|---|
| | | | | | | Patient | Control |
| T | A | C | C | 0.2721 | 0.1344 | 0.6731 | 0.7212 |
| T | G | C | C | 0.1685 | 0.1620 | 0.0912 | 0.0685 |
| C | A | C | C | 0.4613 | 0.4584 | 0.0719 | 0.0826 |
| T | A | T | C | 0.2796 | 0.2722 | 0.0693 | 0.0517 |
| T | G | C | T | 0.6091 | 0.6075 | 0.0614 | 0.0546 |

**Table 3.12 *ICOS* haplotype association testing and frequency estimations (>5%) for patient and control populations**

| icos-2394 | icos-2119 | icos+1323 | icos+2003 | icos+3990 | icos+4031 | icos+5291 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Patient | Control |
| T | C | C | C | G | A | A | 0.3465 | 0.3080 | 0.3260 | 0.3534 |
| T | C | C | C | T | A | A | 0.9746 | 0.9750 | 0.2676 | 0.2713 |
| T | C | G | C | G | C | A | 0.4248 | 0.4000 | 0.1323 | 0.1177 |
| T | C | C | T | G | C | G | 0.5327 | 0.5490 | 0.0614 | 0.0698 |
| C | C | C | C | G | A | A | 0.5236 | 0.5010 | 0.0505 | 0.0629 |

**Table 3.13 Ht*ICOS* haplotype association testing and frequency estimations (>5%) for patient and control populations**

| icos+1323 | icos+2003 | icos+3990 | icos+4031 | icos+5291 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Patient | Control |
| C | C | G | A | A | 0.5939 | 0.5284 | 0.4049 | 0.4190 |
| C | C | T | A | A | 0.9003 | 0.8822 | 0.3328 | 0.3420 |
| G | C | G | C | A | 0.3414 | 0.3160 | 0.1780 | 0.1585 |
| C | T | G | C | G | 0.9930 | 0.9928 | 0.0791 | 0.0790 |

**Table 3.14 Promoter *ICOS* haplotype association testing and frequency estimations (>5%) for patient and control populations**

| icos-2394 | icos-2119 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|
| | | | | Patient | Control |
| T | C | 0.6094 | 0.3782 | 0.8120 | 0.7872 |
| C | C | 0.7949 | 0.7908 | 0.1257 | 0.1307 |
| C | T | 0.2839 | 0.2762 | 0.0606 | 0.0779 |

**Table 3.15 *CD28* haplotype association testing and frequency estimations (>5%) for patient and control populations**

| cd28rs3769683 | cd28int3+17 | cd28rs3181113 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|
| | | | | | Patient | Control |
| G | T | G | 0.6019 | 0.4356 | 0.7115 | 0.7355 |
| G | C | G | 0.4369 | 0.4128 | 0.1291 | 0.1393 |
| A | T | G | 0.0730 | 0.0670 | 0.1187 | 0.0826 |

### 3.3.9 Inference of non-genotyped SNP frequencies using observed haplotype frequencies

In section 3.8, haplotype frequencies were estimated and these frequencies are now used to infer allele frequencies of non-genotyped SNPs. Using the 5 htSNPs and CTLA4_CT60, as these SNPs were shown to produce lower error rates in haplotype reconstruction (section 3.3.3), approximately 95% of *CTLA4* haplotypes have been tagged. To deduce the frequency of a particular allele of a non-genotyped SNP, all frequencies of haplotypes which contain that allele are added and compared between patients and controls. Results are shown in tables A3.16 to A3.21 in Appendix IV. No non-genotyped SNP showed a significant association with disease when allele frequencies were examined. A similar approach was undertaken for inferral of the allele frequences of non-genotyped *ICOS* SNPs which were found to occur on the haplotypes tagged by the *ICOS* htSNPs, as outlined in figure 3.3. Results are shown in tables A3.22 and A3.23 in Appendix IV. No SNP was found to be significantly associated with disease.

### 3.3.10 Evolutionary Conserved Regions Analysis

The 2q33 region was analysed using the ECR browser software (http://ecrbrowser.dcode.org/). The ECR Browser is a program which identifies Evolutionary Conserved Regions (ECRs) in genomes of sequenced species including human, mouse, rat and *Fugu*. Here the 2q33 region has been aligned to dog, mouse, chicken and frog sequences. The conservation profile created by aligning the human sequence (the "base") with all others in a pairwise fashion is graphically displayed in figure 3.8. ECRs are identified as regions of high sequence identity against a neutrally evolving background. By scanning an alignment, the program detects and highlights sequence elements greater then 100 bp that are conserved above a specified level of sequence identity between the two genomes (70%) and highlights them as ECRs. Visually ECRs are represented as coloured peaks on a graph, with the x-axis representing positions in the base genome and the y-axis representing % identity between the base and aligned genomes at that specified position.

To position ECRs in relation to protein coding features, annotated genes are identified as a horizontal blue line above the graph, with transcriptional orientation indicated by the inclined vertical lines. Blue boxes indicate positions of coding

exons, while yellow boxes correspond to UTRs. Peaks within the conservation profile that correspond to these exons are also coloured blue. Peaks within the conservation profile that do not correspond to transcribed sequences are highlighted in red if they are intergenic or pink if they lie within an intron. Green bars on the bottom axis of the plot show the position of repetitive elements in the base genome and this annotation is shaded to the top of the plot in grey.

No ECR element which was conserved between all species was identified within any of the three genes analysed and, hence, no SNPs selected in this project could be identified as occurring in ECRs across the species analysed. SNP rs3180418 was the only SNP found to occur within an ECR.

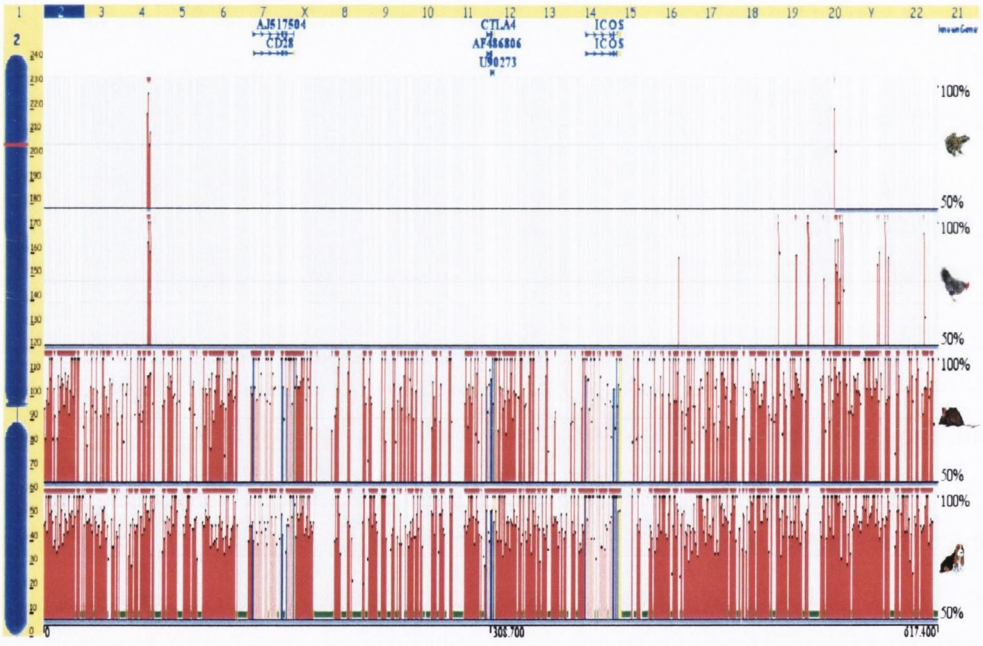



**Figure 3.8 Graphical summary of evolutionary conserved region (ECR) analysis in the 2q33 region using the ECR browser software**

### 3.4 Discussion

The *CTLA4* gene has been the focus of much analysis in the effort to map the causative factors of coeliac disease, and autoimmune diseases in general in the chromosomal region 2q33. This region, which also contains the immunologically important genes, *CD28* and *ICOS*, has long been associated with autoimmune disease, but the exact causal genetic sequence variation has yet to be established in CD.

Our results show the association of several SNPs in the *CTLA4* gene with CD, among them, the functional SNP, CT60, however after correction for multiple testing, only CTLA4-658 remains significant. Furthermore, haplotypes constructed in the region show a larger frequency differential between cases and controls than any individual SNP, with the strongest association defining a haplotype block extending from the 5' end of *ICOS* through *CTLA4* and including the most distal gene *CD28*.

Previous studies have resulted in conflicting findings, with no genetic association being consistently reproducible. The conflicting results between the studies investigating the *CTLA4* region and CD may be due to a number of factors. For example, there may be genetic heterogeneity between population groups studied, while some studies may not have had sufficient power to detect a disease locus in the region, as this locus appears to confer only a low genotypic risk to disease. One study (Popat *et al.* 2002) has estimated the disease locus in the 2q33 gene region as being associated with, at most, a 1.8-fold increased sibling relative risk of CD, but stated that the real figure is probably closer to 1.3 or less. With a 1.8-fold increased sibling relative risk, a study of 100 affected sib-pairs will have less than 35% power to detect a susceptibility locus. The choice of genetic markers in association studies is also a critical factor (Zondervan and Cardon 2004).

Of the 16 SNPs we analysed in this region, only four show independent association with disease status; these are CTLA4-658, CTLA4-1661, CTLA4_CT60 and CD28_rs3769683. CTLA4-658 was the only polymorphism found to deviate from HWE in the patient population. Such deviation may also suggest a role for the locus in disease susceptibility (Nielsen *et al.* 1998). Notably, Ueda *et al* (Ueda *et al.* 2003) have comprehensively studied the *CD28-CTLA4-ICOS* region, identifying 108 SNPs

and determined that the highest level of association with a number of autoimmune disorders (CD was not studied) was at CT60 (+6230A/G). They suggested that the CT60 polymorphism influences alternate RNA splicing of *CTLA4*, resulting in differing ratios of a full-length form, fl*CTLA4*, and a soluble form, s*CTLA4*, of the protein. Individuals who carry the G allele have a lower proportion of s*CTLA4* when compared with those carrying the A allele. s*CTLA4* can bind to the CD80/CD86 receptors, competing with *CD28* and can inhibit T-cell proliferation, hence individuals deficient in the soluble form of the protein may experience abnormal T-cell proliferation similar to that seen in autoimmune disease. CT60 is the only variant among those showing a trend towards significance which has been shown experimentally to have a functional role in gene regulation or protein function.

Most earlier studies of this region have focused on the characterized +49, -318 and 3'$(AT)_n$ polymorphisms in *CTLA4*, with inconsistent results. Several studies have linked the CTLA4+49A allele with susceptibility. We do not find an independent association with this SNP to CD and, in fact, the A allele is over-represented on haplotypes showing a protective or neutral effect in the data here presented. Recently, studies have expanded this search to incorporate neighbouring *CD28* and *ICOS* markers (Holopainen *et al.* 2004), but results have again been inconsistent. For example, Amundsen *et al.* have found the most likely association at CTLA4-1147 (Amundsen *et al.* 2004) while Haimila *et al.* have implicated the *ICOS* gene as the susceptibility locus (Haimila *et al.* 2004).

In this study, we have employed a haplotype tagging approach. We have utilised the haplotype tagging approach for the *CTLA4* gene outlined by Johnson *et al* (Johnson *et al.* 2001), and the approach outlined in Haaning Andersen *et al.* for the *ICOS* gene (Haaning Andersen *et al.* 2003). We have investigated the accuracy of this method to tag haplotypes using newly available software and have found the method to be reliable, and have been able to reproduce haplotype frequency patterns from the original studies. Any deviations observed may be explained by population differentiation, and differing patterns of LD between populations.

The haplotype tagging method used for the *CTLA4* gene had previously identified five htSNPs to tag a 12 SNP haplotype (Johnson *et al.* 2001). We genotyped all 5

htSNPs in our population. We also genotyped CT60, given its potential functional importance. Similarly, our work on the *ICOS* gene involved the genotyping of htSNPs describing haplotypes as previously defined in a European population (Haaning Andersen *et al.* 2003). Using haplotypes constructed from 16 *ICOS* SNPs, we identified 6 htSNPs to tag this variation.

In addition to the utility of the htSNP method to capture the haplotype structure in the region of interest, this method can also be used to capture information at non-genotyped loci. We have implemented this method in the present study to infer the allele frequencies of non-genotyped SNPs in the *CTLA4* and *ICOS* genes which have been captured within the haplotypes tagged by the htSNPs. It has to be noted that this is a crude method which does not account for error estimations intrinsic in the computational reconstruction of the haplotypes. It also does not take into account the presence of rare haplotypes. However, this method is useful for identifying SNPs which may potentially exhibit a significant frequency differential between patient and control groups. Any such identified frequency differential would need to be experimentally verified. In this study, no non-genotyped SNP in the *CTLA4* or *ICOS* genes was found to have a statistically significant inferred frequency differential between patient and control group.

Due to genotyping difficulties, we were unable to type ICOS+2412, so we could not distinguish between two separate haplotypes with a combined predicted frequency of 40% (35% and 5%). However, none of the identified haplotypes, including this combined pair, showed results even approaching a significant difference between the groups in our population sample, so it is unlikely that this combined haplotype is masking any significant association. We also included two SNPs in the *ICOS* promoter region in addition to the htSNPs. These were included as they may have functional effects on gene expression and they have the added benefit of further defining the *ICOS* haplotypes. These are the -2394 polymorphism situated in a consensus Sp1 site and the -2119 SNP situated in a consensus NF-κB site. No significant association was observed with any *ICOS* haplotype or SNP. Comparing these results to a recent study of *ICOS* (Haimila *et al.* 2004) which found an association with CD, we conclude that our results do not confirm this association. In

this previous analysis, eight polymorphisms were genotyped, of which two polymorphisms overlap with those studied here (+4031 and +5291). They found evidence for association with four polymorphic sites including +4031 (identified as ICOSc.602A/C in their study). We found no evidence for association at this locus. This investigation also found the haplotype containing all four associated SNPs to be more highly associated with disease than the individual SNPs. Using our data and the haplotype information given in their study, we are able to tag this haplotype using the +5291 SNP (ICOSc.1862 in previous study) alone, as the associated haplotype is the only one to contain the G allele at this position and thus we conclude that this haplotype is not associated with disease in our population.

Our haplotype analysis of *CD28* involved three markers selected from dbSNP, chosen to give coverage of the gene. No conclusive association with disease was observed and this finding is consistent with an earlier pan European study (Holopainen *et al.* 2004).

Analysis of the entire region reveals that the most strongly associated haplotype is coterminous with the previously reported "block" structure of the region (Haimila *et al.* 2004) which we confirm by pairwise linkage disequilibrium analysis incorporates *CD28*, *CTLA4* and the 5' region of *ICOS*. Haplotypes consisting of *CTLA4* markers only are less strongly associated with disease status. The greater magnitude of association when the disease associated haplotypes spanning all three genes are considered, suggests a number of possible scenarios. For example, other unassayed SNPs which are on, or in LD with these haplotypes could be the susceptibility variants in this population. Udea *et al.* (Ueda *et al.* 2003) have actually tested for disease association with 108 SNPs, identified through sequencing of the three genes, and they did not identify any polymorphism other than CTLA4_CT60 which was significantly associated with a number of autoimmune disease. However, it should be noted that CD was not included amongst the diseases investigated in this study.

It is also possible that a specific combination of known or unknown SNPs in the vicinity may be required for susceptibility. *CD28*, *CTLA4* and *ICOS* interact to regulate T-cell activity. The fact that all or part of these three genes are found on the associated haplotypes allows for the possibility that small levels of genetic variation

at any combination of these genes may have an additive or epistatic effect in perturbing this demonstrably important system.

In conclusion, the association of the *CTLA4* region with CD has now been replicated several times. The fact that some studies have failed to detect association may in part be due to the low magnitude of the risk conferred by this region. Heterogeneity in study design compounds the difficulty in making direct comparisons which should be minimised by the use of haplotype tagging. However, alleles that are associated in some populations (e.g. CTLA4+49A) are clearly not in ours, suggesting that population heterogeneity may be a contributing factor in the conflicting results presented in the literature. Our results suggest that the disease association in this population lies in LD with a specific extended haplotype around *CTLA4*, strengthening the evidence for the involvement of chromosome 2q33 in CD risk modification.

# CHAPTER IV

## GENETIC ASSOCIATION ANALYSIS OF THE 11q23 REGION AND THE *TNFSF14* GENE

## 4.1 Introduction

### 4.1.1 Coeliac disease and chromosomal region 11q23

Chromosomal region 11q23 is of interest in the study of the genetics of coeliac disease as it has been shown to be linked to CD in a number of genome scans. Three studies have independently replicated linkage to this region. Initial linkage to 11q23 was identified in an Italian sample (Greco *et al.* 1998). Following this, Holopainen *et al.* (Holopainen *et al.* 2001) have found linkage at 11q23, with five markers, in a 5.6 cM region between markers D11S4111 and CD3D, showing statistically significant (point-wise significance $p<0.05$) linkage with disease. The highest level of association was observed at marker D11S4142. Linkage in the region was confirmed through the use of both the transmission disequilibrium test (TDT) and association studies. The 11q23 region was also demonstrated to be linked to CD by Naluai *et al.* (Naluai *et al.* 2001). Here, a genome scan was carried out with a mean inter-marker spacing of 10cM. The highest level of linkage to disease observed in this study, after the well-defined linkage at 6p21 (HLA region) was at marker D11S4464. Such independent replications of a linkage region would suggest that this area is indeed a true linkage region in CD and warrants further investigation.

### 4.1.2 Candidate genes in the 11q23 region

Analysis of the 11q23 linkage region identified several genes which may represent good candidates for a role in the pathogenesis of a T-cell mediated disease such as coeliac disease, due to their T-cell or immunological function. The CD3 genes, CD3-epsilon (*CD3E*), CD3-delta (*CD3D*) and CD3-gamma (*CD3G*), lie within 50 kb of each other in this region. These genes form part of the T cell-receptor (TCR) complex. This complex consists of either alpha and beta or gamma and delta variant chains, in association with the invariant chains CD3E, CD3D, CD3G and CD3-zeta (CD3Z). During development, this CD3 protein complex plays an important role in the transition of thymocytes from immature precursors to the final mature CD4+ or CD8+ single-positive T cell. Studies have shown that the CD3 components are essential during the early stages of human thymopoiesis and deficiencies in these genes have been linked with severe combined immunodeficiency (Dadi *et al.* 2003; de Saint Basile 2004). One preliminary study has been carried out to date on the association between CD3 and autoimmune disease (Ghabanbasani *et al.* 1994). This

study identified a significant association between *CD3D* and type 1 diabetes with the use of microsatellites.

Also located on 11q23 is the *THY1* gene. This gene is a major cell surface glycoprotein characteristic of T cells and is a member of the immunoglobulin supergene family. Although the role of this protein is unclear, its position as a cell surface molecule on T cells may suggest an involvement in cell-cell interactions and hence *THY1* may be considered as a candidate gene for coeliac disease pathogenesis.

Also located in this region is the gene encoding the interleukin 10 receptor alpha chain (*IL10RA*). This molecule is the receptor for IL10, an anti-inflammatory cytokine produced by a subset of activated T cells, B cells and macrophages which has a role in a variety of immunoregulatory functions (Moore KW *et al.* 2001). A polymorphism in *IL10* (-1087) has been linked to a number of autoimmune diseases including inflammatory bowel disease, rheumatoid arthritis (RA) and systemic lupus erythematosus (Turner *et al.* 1997). A significant association between the IL10-1087 polymorphism and coeliac disease has been shown (Hahn-Zoric *et al.* 2003), however, this association failed to be replicated in subsequent work (Lio *et al.* 2005). The IL10-1087 polymorphism has also been shown to be associated with differing levels of IgA anti-endomysium and anti-tissue transglutaminase antibodies in CD patients (Hahn-Zoric *et al.* 2003), while recombinant human interleukin 10 has been shown to suppress gliadin dependent T cell activation in *ex vivo* cultured coeliac intestinal mucosa (Salvati *et al.* 2005). These observations suggest that *IL10* may be a factor in the pathogenesis of CD. However, as *IL10* signals through the specific receptor, *IL10RA*, the possible role for this receptor molecule in disease, mediated through the *IL10* pathway should also be considered. Indeed it has be shown that *IL10RA* plays an important role in IL10-mediated immune regulation (Ding *et al.* 2001). To date, no genetic association studies have been carried out to test for association between polymorphisms in the *IL10RA* gene and coeliac disease.

Also included as a candidate gene in this region is *IL18*. Though this gene, located at chromosomal position 11q22.2-22.3 (111.5 MB), is some distance from the major linkage peaks in the region, it may still fall within the region of linkage as it is located between a marker shown to be linked to disease (D11S4111) at position

115.8 MB and the next tested marker (D11S898) at 100.6 MB which was not linked to disease (Holopainen *et al.* 2001). *IL18* is a proinflamatory cytokine, and in synergy with *IL12*, promotes development of the Th1 lymphocyte response by induction of γ-interferon (IFN-γ) production. It also modulates activity of NK cells, increases TNF-α and IL1 production by macrophages, upregulates the expression of adhesion molecules, and induces production of molecules destructive to tissue such as nitric oxide. *IL18* is known to play an important role in inflammatory, autoimmune and infectious diseases (Gracie *et al.* 2005), and increased serum levels of IL18 have been identified in patients with autoimmune diseases such as RA (Gracie *et al.* 2005) and acute asthma (Furugaki *et al.* 2004).

Two polymorphisms in the promoter of *IL18* have shown evidence of altering *IL18* protein expression. One polymorphism located at position -607 has been found to disrupt a potential cAMP-response element protein-binding site, while the other at position -137 has been shown to alter a consensus H4TF-1 nuclear factor-binding site. Multiple sclerosis patients homozygous for the -607C and -137G alleles have higher levels of *IL18* mRNA compared to other haplotypes, suggesting that these polymorphisms do indeed regulate activity of the gene (Gierdraitis *et al.* 2001). Numerous genetic association studies have been carried out on these SNPs and others in the gene, to investigate if any association exists with various autoimmune diseases. Several studies have shown a significant association between *IL18* with disease, for example, type 1 diabetes (Kretowski *et al.* 2002), Crohn's disease (Tamura *et al.* 2002), atopic eczema (Novak *et al.* 2005), inflammatory bowel disease (Aizawa *et al.* 2005), and asthma (Higa *et al.* 2003). Other studies have failed to find a disease association between *IL18* and several diseases, among which are studies on coeliac disease (Rueda *et al.* 2005), RA (Rueda *et al.* 2005), type 1 diabetes (Hovdenak *et al.* 1999; Martin *et al.* 2005; Novota *et al.* 2005), and periodontitis (Folwaczny *et al.* 2005).

### 4.1.3 LIGHT (TNFSF14)

An interesting candidate for inclusion in this study was the lymphotoxin-like inducible protein that competes with glycoprotein D for binding herpesvirus entry mediator on T cells (known as *LIGHT* or *TNFSF14*) gene. This gene is located on chromosomal region 19p13.3 and at this position is located outside the linkage region

for CD on chromosome 19 at position 19p13.1 (Van Belzen *et al.* 2003). However, it is located within a region of linkage to another inflammatory disorder, Crohn's disease (Rioux *et al.* 2000), and recent evidence relating to its function in T cells as outlined below, suggests that *LIGHT* is an excellent candidate gene for coeliac disease pathogenesis.

*LIGHT* is a member of the TNF superfamily, a group of molecules which play an important costimulatory role in T cell activation and inflammation. The *LIGHT* molecule is expressed in the lymphoid compartment by activated T cells, and also by monocytes. *LIGHT* signalling is via one or two pathways; the herpesvirus entry mediator (HVEM, TNFRSF14) and the lymphotoxin (LT)βR (TNFRSF3) receptors. *LIGHT* is thought to play an immunomodulatory role controlling stimulatory T cell-T cell interactions via the HVEM pathway, as HVEM is a costimulatory signal, enhancing proinflammatory cytokine production and T cell proliferation (Harrop *et al.* 1998). Recently, this gene has become of interest in inflammatory diseases of the gut as experiments on mouse models have shown that constitutive T cell expression of *LIGHT* resulted in mucosal inflammation and this provided the initial evidence for a role for *LIGHT* in mucosal immune regulation (Shaikh *et al.* 2001; Wang *et al.* 2001). In addition to this, inhibition of the LTβR signalling pathway with a LTβR-Fc chimera decoy receptor diminished inflammatory symptoms in the $CD4^+$ $CD45RB^{high}$ T cell transfer model of colitis, suggesting a role for *LIGHT* in this $CD4^+$ T cell-mediated pathology (Morrissey and Charrier 1994). Moreover, recent experiments (Cohavy *et al.* 2004; Cohavy *et al.* 2005), which have shown elevated constitutive *LIGHT* expression on human intestinal T cells and a requirement of *LIGHT* for interferon-gamma (IFN-γ) production, suggest a mechanism by which the mucosal specificity of *LIGHT*-mediated inflammation could have significant pathological implications for inflammatory diseases of the gut such as coeliac disease. Hence, this gene was included as a candidate for analysis in this study.

### 4.2 Objectives

The aims of the work outlined in this chapter were to:

1. Identify genes in the previously linked 11q23 region which are potential candidate genes in an autoimmune disease such as coeliac disease.

2. Select SNPs within these candidate genes which are most informative based on the available information, either in their ability to act as haplotype tagging SNPs (htSNPs) or their potential to alter protein sequence or transcription factor binding sites.
3. Test the accuracy of the htSNP method to tag all SNPs within a tagged haplotype.
4. Test for a statistically significantly allelic, genotypic, carrier status or haplotype association between polymorphisms in the 11q23 region and coeliac disease.
5. Analyse the LD structure in the region.
6. Test for haplotype association with disease.
7. Infer non-genotyped SNP frequencies from haplotype data.
8. Analysis the LIGHT gene using the methods outlined in steps 1-7.

## 4.3 Results

### 4.3.1 Gene selection on chromosomal region 11q23.3

As discussed, chromosomal region 11q23.3 has been shown in a number of independent genome scans to be linked to coeliac disease. Association of 11q markers with coeliac disease have been demonstrated by Greco *et al.* (Greco *et al.* 1998) Naluai *et al.* (Naluai *et al.* 2001) and Holopainen *et al.* (Holopainen *et al.* 2001). Markers from 115.36 Mb (D11S4111) to 123.66 Mb (D11S4464) have shown linkage, with peak associations at D11S4142 (115.38 Mb, Holopainen *et al*) and D11S4464 (123.66 Mb, Naluai *et al*). Hence, a gene search was carried out in the region demarked by D11S4111 and D11S4464. Also 100 kb of sequence was searched either side of the region using the Human ContigView option in the Ensembl (v32) genome browser. All genes identified within this region are listed in Appendix III. Five genes with demonstrated immune function were identified in the region, (*IL10RA*, *CD3D*, *CD3E*, *CD3G* and *THY1*) were selected for further analysis. *IL18* is located between marker D11S4111, which was linked to disease and the subsequent marker (D11S898), analysed by Holopainen *et al.*, which was found not to associated with disease. *IL18* was considered to be a good candidate for analysis, given that it plays an important role in immune and inflammatory processes and because SNPs with reportedly functional regulatory effects have been already described for *IL18*.

### 4.3.2 SNP selection results

#### *4.3.2.1* CD3E

Each database, as listed in Appendix I, was searched for genotype/haplotype information for *CD3E*. The gene genotype report feature on the NCBI database (http://www.ncbi.nlm.nih.gov) was the only resource found to contain information on this gene. Seven polymorphisms were listed in the Perlegen European dataset, of which 2 polymorphisms (rs17122060 and rs17122067) were monomorphic. The genotype information on the remaining 5 SNPs (listed in figure 4.1), as obtained from the 24 individuals listed in this database, were used to generate haplotypes using PHASEv2.0.2. These haplotypes, along with their associated frequencies are listed in figure 4.1. The SNPtagger program was used to identify htSNPs to tag all identified genetic variation in the region. All SNPs identified were intronic or immediately surround the *CD3E* gene. (Note: transcription of all genes is shown from left to right in diagrams).

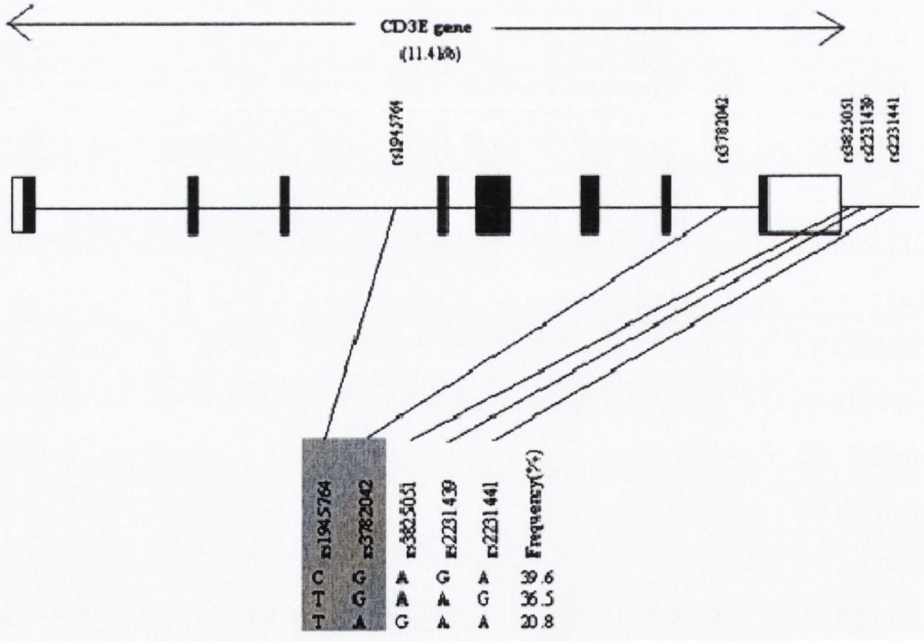



**Figure 4.1 CD3E haplotype tagging SNP selection. HtSNPs identified, are shaded in grey, and describe all haplotypes with a frequency >5%. All SNPs contained within the haplotypes are listed within their haplotype structure in the above table and their relative positions within the *CD3E* gene are shown in the diagram. The *CD3E* gene is shown with exons identified by shaded boxes.**

### *4.3.2.2 CD3D and CD3G*

The *CD3D* and *CD3G* genes are located only 1.8 kb apart on chromosome 11, hence it was possible to tag polymorphisms in both genes simultaneously. Again, the gene genotype report feature on the NCBI database was the only resource found to contain information on this gene. 11 polymorphisms were identified (listed in figure 4.2). The genotype information on these 11 SNPs, as obtained from the 24 individuals listed in this database, were used to generate haplotypes using PHASEv2.0.2. These haplotypes, along with their associated frequencies are listed in figure 4.2. The SNPtagger program was again used to identify htSNPs. As with the *CD3E* gene, the NCBI database did not contain any non-synonymous SNPs with identified minor allele frequencies.

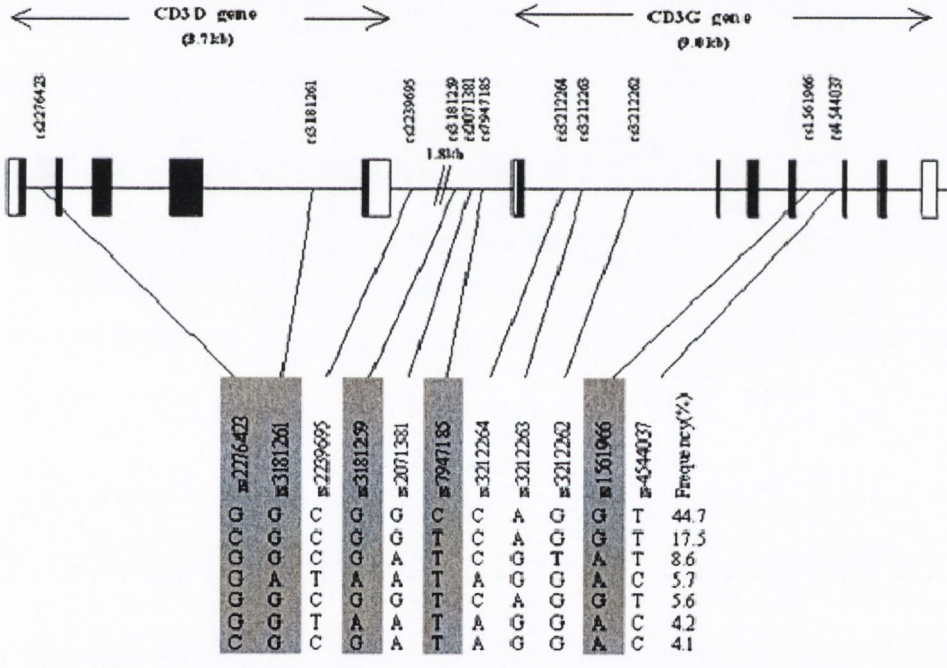



**Figure 4.2 CD3D/CD3G haplotype tagging SNP selection. HtSNPs identified, are shaded in grey, and describe all haplotypes with a frequency >4%. All SNPs contained within the haplotypes are listed within their haplotype structure in the above table and their relative positions within the genes are shown in the diagram. The *CD3D/CD3G* gene is shown with exons identified by shaded boxes.**

### *4.3.2.3* THY1

Again for the *THY1* gene, the gene genotype report feature on the NCBI dbSNP database was the only resource found to contain information on this gene. Nine *THY1* polymorphisms have been genotyped in the Perlegen European sample. One SNP (rs17122908) was found to be monomorphic in this population. The genotype information on the remaining 8 SNPs (listed in figure 4.3) was used to generate haplotypes, as carried out previously for genes *CD3E* and *CD3D/CD3G*. HtSNPs were again identified using SNPtagger and these are highlighted in figure 4.3. The NCBI SNP database did not list any non-synonymous SNPs with a minor allele frequency greater >5% which may be of interest for the purposes of this study.

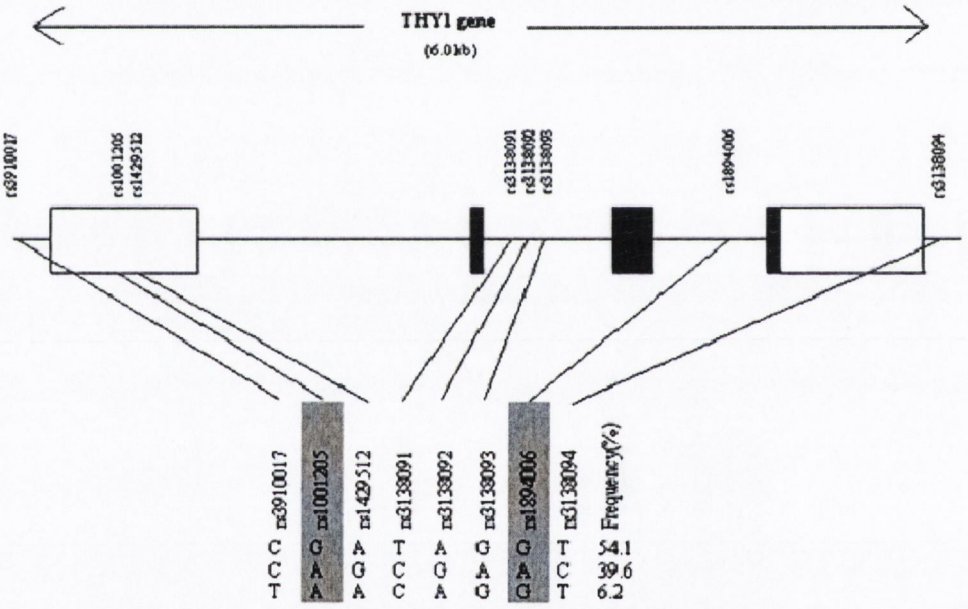



**Figure 4.3 THY1 haplotype tagging SNP selection. HtSNPs are shown in grey, and describe all haplotypes with a frequency >5%. Exons are identified by shaded boxes.**

### *4.3.2.4* IL10RA

Extensive haplotype information was available for the *IL10RA* gene from the Seattle SNPs website (http://pga.gs.washington.edu/data/il10ra/). This database listed haplotypes reconstructed in a European population, consisting of 51 loci. However, 20 of these loci were monophorphic in the population when haplotypes with a frequency >4% were analysed, hence these SNPs were not included in this analysis.

The remaining 31 polymorphic loci are listed in figure 4.4 below. HtSNPs were identified as before. HtSNPs rs4252272, rs2229113 and rs2229114 are non-synonymous changes. HtSNP rs2256111 is a synonymous SNP, rs4252314, is located in intron 6 and the remaining htSNP rs9610 is found in the untranslated region of the gene. All other SNPs identified in the haplotype are intronic or in the area surrounding the gene, with the exception of rs4252301 (synonymous SNP in exon 1) and rs4252250 (nonsynonymous SNP in exon 2).

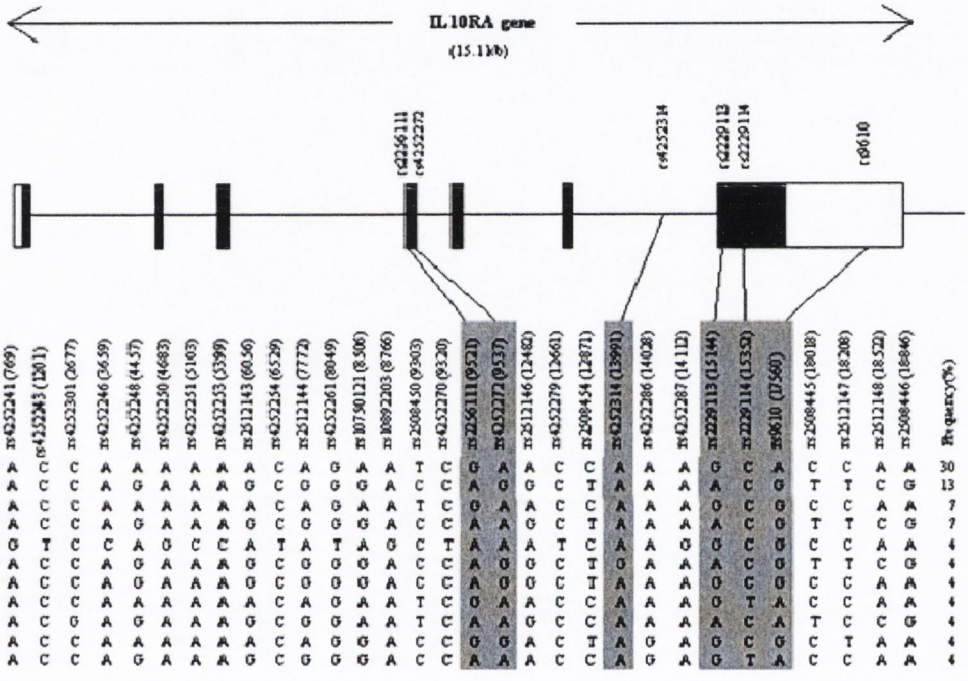



**Figure 4.4 IL10RA haplotype tagging SNP selection. HtSNPs identified, are shaded in grey, and describe all haplotypes with a frequency >4%**

### *4.3.2.5* IL18

Extensive genotype information was available for the *IL18* gene from the IIPGA (Innate Immunity) website (http://innateimmunity.net/PGAs/InnateImmunity/IL18/). This database consisted of 38 polymorphic loci identified in a European population. However, 11 loci had a minor allele frequency <5% and were not included for further analysis. The genotype information on the remaining 27 SNPs (listed in figure 4.5), which was available on the IIPGA site, was then used to construct haplotype data using PHASEv2.0.2, following which, htSNPs were selected using SNPtagger. Results are shown in figure 4.5. HtSNPs rs5744241 and rs187238 are located in the promoter region of the gene and correspond to positions -607 and -137 respectively.

The other htSNPs, rs2043055 and rs5744241 are intronic. All remaining SNPs occur in the intronic or untranslated region of the gene.

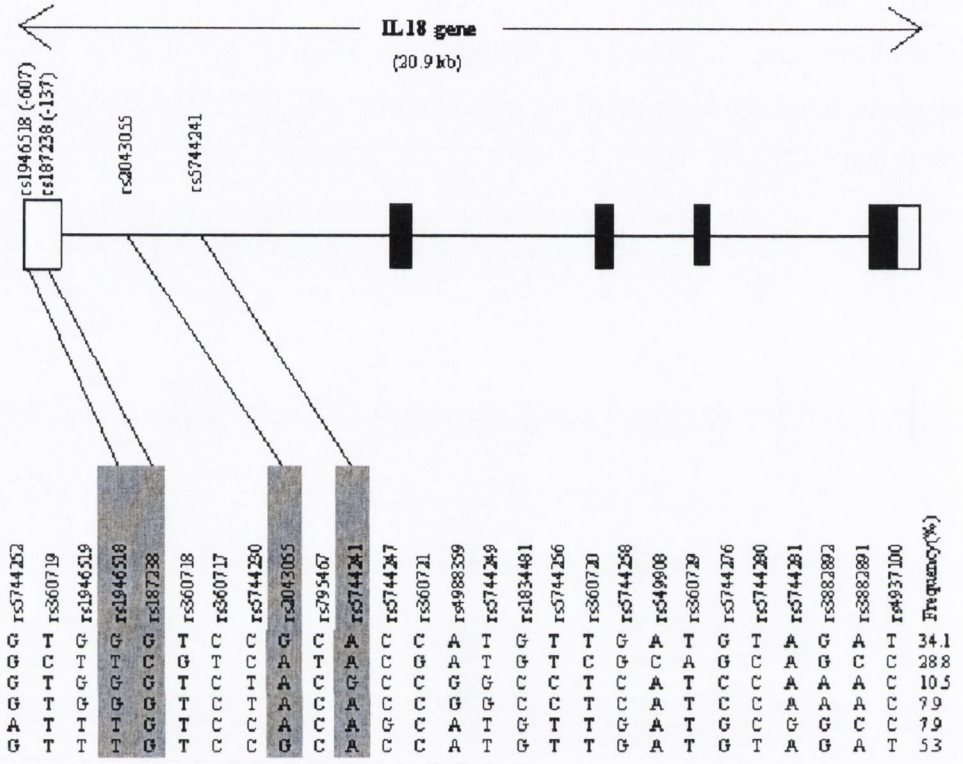



**Figure 4.5 IL18 haplotype tagging SNP selection. HtSNPs identified, are shaded in grey, and describe all haplotypes with a frequency >5%**

### *4.3.2.6 LIGHT (TNFSF14)*

The largest body of genotype/haplotype information for the *TNFSF14* gene was found in the Perlegen European sample of the NCBI database where five polymorphisms have been genotyped. One SNP (rs2291667) was found to be monomorphic in this population. Haplotypes and htSNPs were identified using genotype information on the remaining SNPs (listed in figure 4.6) as before (figure 4.6). HtSNP rs2291668 is a synonymous SNP, htSNP rs3760746 occurs in the untranslated region of the gene and htSNP rs11878563 occurs in close proximity to the gene. However, the non-synonymous SNP rs344560, as identified on the NCBI dbSNP site was also chosen for inclusion in this study as was SNP rs2279627. The rs2279627 SNP was identified on the HapMap project (http://www.hapmap.org) and is located 3' of the TNFSF14 gene. The rs344560 SNP was the only non-synonymous SNP on the NCBI dbSNP database with a minor allele frequency >5%.

Both of these SNPs were selected based on their potential functional effects either as an amino acid changing SNP (rs344560) or potential gene regulatory SNP (rs2279627).

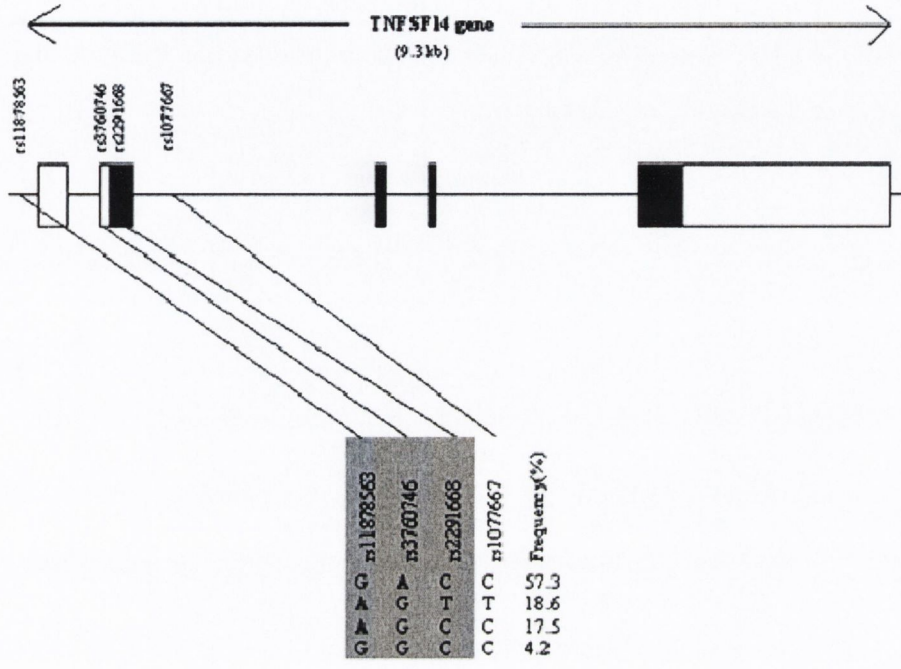



**Figure 4.6 TNFSF14 haplotype tagging SNP selection. HtSNPs identified, are shaded in grey, and describe all haplotypes with a frequency >4%**

### 4.3.3 Haplotype-tagging SNP (htSNP) selection and accuracy testing

As a means of assessing the estimated accuracy of our haplotype tagging approach, we used the MARKER program (section 2.9) to estimate the percentage error associated with haplotype inference when using each of the htSNPs sets for the genes in the 11q23 region (as outlined in 4.3.2). The percentage error rates associated with the use of the htSNP sets alone were estimated. Also the effect of the addition of an extra SNP to the htSNP set on the percentage error rates was also considered. This method also allows us to assess how good haplotype inference was likely to be under different conditions, for example, when significant amounts of genotype data is missing due to genotyping dropout. The MARKER program evaluates the data under two difference situations; with 0% missing data and 20% missing data.

Results were analysed for the CD3D, CD3E, CD3G, IL10RA, THY1 and IL18 genes as was carried out for the *CTLA4* gene in section 3.3.3.1. No gene displayed an

appreciable reduction in error rates upon addition of an extra SNP to the htSNP data set, which would warrant the inclusion of this extra SNP (Data not shown).

The accuracy of the haplotype tagging method undertaken in the *TNFSF14* (*LIGHT*) gene could not be assessed as the MARKER program does not facilitate analysis of haplotypes containing less than 5 markers.

### 4.3.4 Hardy-Weinberg Equilibrium (HWE) Testing

All SNPs in this region were tested for compliance with HWE using GENEPOP. Results are listed in tables 4.1 and 4.2. No SNP violated HWE after correction for multiple testing.

**Table 4.1 Hardy Weinberg Equilibrium Analysis results for all SNPs in candidate genes on chromosome 11**

| SNP | Patient group p-value | Control group p-value |
|---|---|---|
| CD3E_ht1 | 0.5835 | 0.9085 |
| CD3E_ht2 | 0.7882 | 0.0447 |
| CD3DG_ht1 | 0.8883 | 1.0000 |
| CD3DG_ht3 | 1.0000 | 0.5567 |
| CD3DG_ht4 | 0.9028 | 0.1544 |
| CD3DG_ht5 | 0.5937 | 0.4943 |
| IL18_ht1 | 0.6767 | 0.0299 |
| IL18_ht2 | 0.0884 | 0.0741 |
| IL18_ht3 | 1.0000 | 0.6841 |
| IL18_ht4 | 0.0423 | 0.5041 |
| Thy1_htSNP2 | 0.4914 | 0.1944 |
| Thy1_htSNP7 | 0.9080 | 0.9049 |
| IL10RA_ rs2256111 | 0.5341 | 0.0612 |
| IL10RA_ rs4252272 | 0.0420 | 0.1593 |
| IL10RA_ rs2229113 | 0.7236 | 0.0287 |
| IL10RA_ rs2229114 | 0.5269 | 0.1612 |

**Table 4.2 Hardy Weinberg Equilibrium Analysis results for all SNPs in TNFSF14**

| SNP | Patient group p-value | Control group p-value |
|---|---|---|
| TNFSF14_rs344560 | 1.0000 | 1.0000 |
| TNFSF14_rs2279627 | 0.1519 | 1.0000 |
| TNFSF14_ht1rs2291668 | 0.2014 | 0.4736 |
| TNFSF14_ht3rs11878563 | 0.4485 | 0.1580 |

### 4.3.5 Allele, Genotype and Carrier Status frequencies and disease association testing of genotyped polymorphisms in chromosome 11 candidate genes

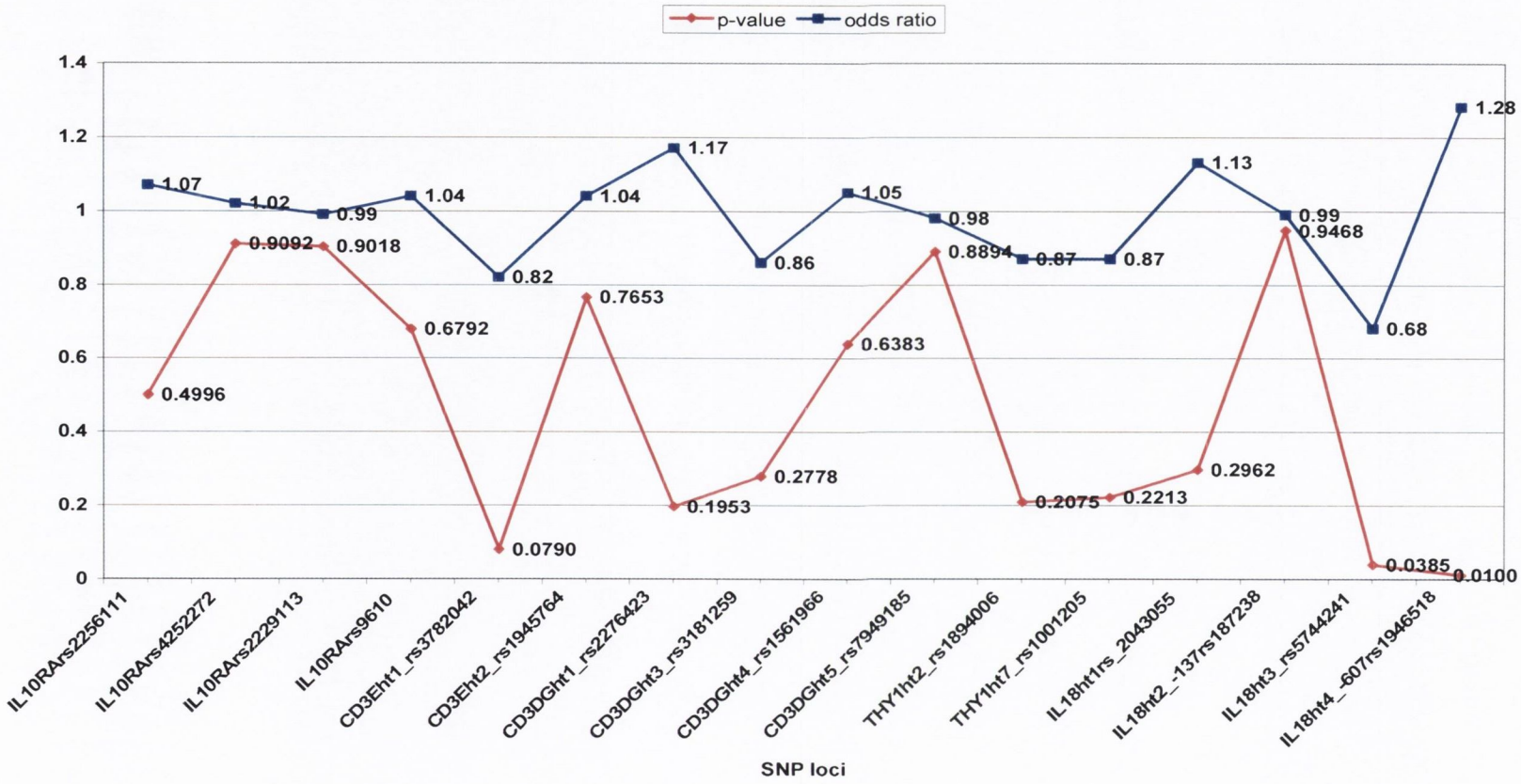
All polymorphic sites in *IL10RA*, *CD3D*, *CD3E*, *CD3G*, *THY1* and *IL18* were analysed for allelic, genotypic and carrier status association with disease individually and results are listed in tables 4.3 (genotype and allele associations) and 4.4 (carrier status associations). The p-values and odds ratios for the allele frequency differentials between patient and control groups are displayed graphically in figure 4.7. No polymorphism within the *IL10RA*, *CD3D*, *CD3E*, *CD3G* or *THY1* genes were found to be significantly associated with disease in this study. Three polymorphisms in the *IL18* gene (IL18_ht2_-137rs187238, IL18_ht3_rs5744241 and IL18_ht4_-607rs1946518) showed a significant association with disease either at allele frequency, genotype frequency or carrier status level prior to correction for multiple testing. IL18-137 was significantly associated with disease when genotype frequencies were analysed (p=0.0380). IL18_ht3 was significantly associated with disease when allele frequencies (patients 87.0%, controls 90.8%, p=0.0385) and carrier status (G allele carrier status / homozygosity for the A allele; patients 75.7%, controls 82.3%, p=0.0490) were examined. IL18-607 also showed a significant association with disease status at the allele frequency level (patients 62.4%, controls 56.5%, p=0.0100), the genotype level (p=0.0265), and also at the level of allele carrier status. G allele carrier status and conversely homozygous status for the T allele was significantly associated with disease (patients 87.1%, controls 81.2%, p=0.0159).

### 4.3.6 Allele, Genotype and Carrier Status frequencies and disease association testing in genotyped polymorphisms in *TNFSF14*

As above, *TNFSF14* SNPs were analysed for allelic, genotypic and carrier status association with disease individually and results are listed in tables 4.5 (genotype and allele associations) and 4.8 (carrier status associations). The p-values and odds ratios for the allele frequency differentials between patient and control groups are displayed graphically in figure 4.7. A significant association was observed between three polymorphisms and disease status. SNP TNFSF14_rs2279627 was associated with disease at the allelic level (patients 27.8%, controls 21.3%, p-value = 0.0052), at the genotypic level (p-value = 0.0096), and at the level of carrier status where an association was seen with carrier status of the G allele (patients 49.5%, controls 37.9,

p-value = 0.0023). A significant association between disease and TNFSF14_ ht3_rs11878563 at the allelic level (patients 41.4%, controls 48.5%, p-value = 0.0092), genotypic level (p-value = 0.0106) and at the level of carrier status where an association was seen with carrier status of the G allele (patients 83.9%, controls 74.5%, p-value = 0.0027). A disease association was also observed with TNFSF14_ ht2_rs3760746 when carrier status for the T allele was tested (p-value = 0.0202).

**Figure 4.7 P-values and odds ratios for allele frequency differences between patient and control groups for SNP loci at 11q23**




Actual allele frequencies listed in table 4.3

**Table 4.3 Genotype and allele frequencies for candidate genes on chromosome 11 in the coeliac disease patient and control populations**

| SNP | | Genotype n (%) | | | | Allele n (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | p | $A_1$ | $A_2$ | p | OR | 95% CI |
| IL10RA | Patient | 86 (23.1) | 192 (51.8) | 93 (25.1) | 0.5138 | 364 (49.1) | 378 (50.9) | 0.4996 | 1.07 | (0.87,1.33) |
| rs2256111 A/G | Control | 69 (19.8) | 192 (55.0) | 88 (25.2) | | 330 (47.3) | 368 (52.7) | | | |
| IL10RA | Patient | 255 (68.0) | 101 (26.9) | 19 (5.1) | 0.0683 | 611 (81.5) | 139 (18.5) | 0.9092 | 1.02 | (0.77,1.33) |
| rs4252272 A/G | Control | 226 (64.8) | 115 (33.0) | 8 (2.3) | | 567 (81.2) | 131 (18.8) | | | |
| IL10RA | Patient | 171 (45.4) | 169 (44.8) | 37 (9.8) | 0.4264 | 511 (67.8) | 243 (32.2) | 0.9018 | 0.99 | (0.79,1.24) |
| rs2229113 G/A | Control | 150 (43.7) | 167 (48.7) | 26 (7.6) | | 467 (68.1) | 219 (31.9) | | | |
| IL10RA | Patient | 113 (30.3) | 191 (51.2) | 69 (18.5) | 0.7672 | 417 (55.9) | 329 (44.1) | 0.6792 | 1.04 | (0.84,1.29) |
| rs9610 G/A | Control | 96 (28.0) | 184 (53.6) | 63 (18.4) | | 376 (54.8) | 310 (45.2) | | | |
| CD3E | Patient | 118 (33.3) | 178 (50.3) | 58 (16.4) | 0.1959 | 414 (58.5) | 294 (41.5) | 0.0790 | 0.82 | (0.66,1.03) |
| ht1_rs3782042 G/A | Control | 133 (39.6) | 158 (47.0) | 45 (13.4) | | 424 (63.1) | 248 (36.9) | | | |
| CD3E | Patient | 158 (49.7) | 131 (41.2) | 29 (9.1) | 0.2227 | 447 (70.3) | 189 (29.7) | 0.7653 | 1.04 | (0.81, 1.32) |
| ht2_rs1945764 T/C | Control | 155 (45.9) | 160 (47.3) | 23 (6.8) | | 470 (69.5) | 206 (30.5) | | | |
| CD3DG | Patient | 193 (54.4) | 137 (38.6) | 25 (7.0) | 0.4343 | 523 (73.7) | 187 (26.3) | 0.1953 | 1.17 | (0.92, 1.49) |
| ht1_rs2276423 G/C | Control | 171 (49.9) | 142 (41.4) | 30 (8.8) | | 484 (70.6) | 202 (29.5) | | | |
| CD3DG | Patient | 240 (66.5) | 109 (30.2) | 12 (3.3) | 0.4698 | 589 (81.6) | 133 (18.4) | 0.2778 | 0.86 | (0.64, 1.14) |
| ht3_rs3181259 G/A | Control | 238 (69.6) | 97 (28.4) | 7 (2.1) | | 573 (83.8) | 111 (16.2) | | | |

| SNP | | Genotype n (%) | | | | Allele n (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | p | $A_1$ | $A_2$ | p | OR | 95% CI |
| CD3DG | Patient | 147 (41.9) | 159 (45.3) | 45 (12.8) | 0.4220 | 453 (64.5) | 249 (35.5) | 0.6383 | 1.05 | (0.84, 1.32) |
| ht4_rs1561966 G/A | Control | 129 (38.2) | 170 (50.3) | 39 (11.5) | | 428 (63.3) | 248 (36.7) | | | |
| CD3DG | Patient | 187 (52.5) | 139 (39.0) | 30 (8.4) | 0.9825 | 513 (72.1) | 199 (28.0) | 0.8894 | 0.98 | (0.77, 1.25) |
| ht5_rs7949185 T/C | Control | 183 (53.2) | 132 (38.4) | 29 (8.4) | | 498 (72.4) | 190 (27.6) | | | |
| THY1 | Patient | 161 (43.5) | 171 (46.2) | 38 (10.3) | 0.3796 | 493 (66.6) | 247 (33.4) | 0.2075 | 0.87 | (0.69,1.09) |
| ht2_rs1894006 G/A | Control | 161 (47.1) | 155 (45.3) | 26 (7.6) | | 477 (69.7) | 207 (30.3) | | | |
| THY1 | Patient | 122 (34.3) | 173 (48.6) | 61 (17.1) | 0.4673 | 417 (58.6) | 295 (41.4) | 0.2213 | 0.87 | (0.70,1.09) |
| ht7_rs1001205 G/A | Control | 127 (37.9) | 160 (47.8) | 48 (14.3) | | 414 (61.8) | 256 (38.2) | | | |
| IL18 | Patient | 145 (44.2) | 147 (44.8) | 36 (11.0) | 0.2444 | 437 (66.6) | 219 (33.4) | 0.2962 | 1.13 | (0.90,1.43) |
| ht1rs_2043055 A/G | Control | 116 (38.2) | 156 (51.3) | 32 (10.5) | | 388 (63.8) | 220 (36.2) | | | |
| IL18 | Patient | 154 (50.3) | 135 (44.1) | 17 (5.6) | **0.0380** | 443 (72.4) | 169 (27.6) | 0.9468 | 0.99 | (0.77,1.28) |
| ht2_-137rs187238 G/C | Control | 184 (54.6) | 121 (35.9) | 32 (9.5) | | 489 (72.6) | 185 (27.5) | | | |
| IL18 | Patient | 227 (75.7) | 68 (22.7) | 5 (1.7) | 0.1140 | 522 (87.0) | 78 (13.0) | **0.0385** | 0.68 | (0.47,0.98) |
| ht3_rs5744241 A/G | Control | 237 (82.3) | 49 (17.0) | 2 (0.7) | | 523 (90.8) | 53 (9.2) | | | |
| IL18 | Patient | 166 (37.6) | 218 (49.4) | 57 (12.9) | **0.0265** | 550 (62.4) | 332 (37.6) | **0.0100** | 1.28 | (1.06,1.54) |
| ht4_-607rs1946518 G/T | Control | 152 (31.7) | 238 (49.6) | 90 (18.8) | | 542 (56.5) | 418 (43.5) | | | |

n = number of individuals; p = p-value, CI = confidence interval

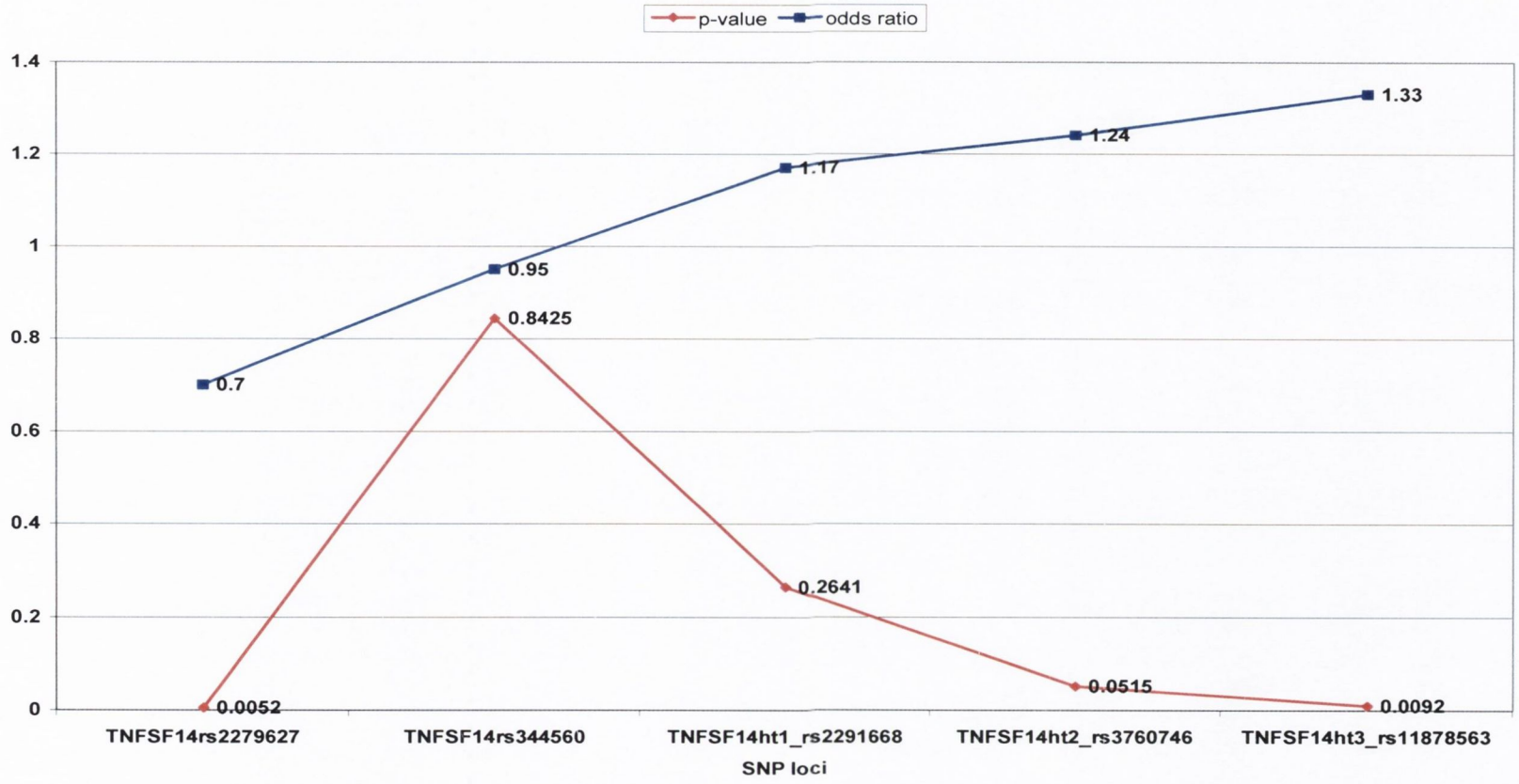
**Table 4.4 Carrier status frequencies for candidate genes on chromosome 11 in the coeliac disease patient and control populations**

| SNP | | Allele $A_1$ Carrier Status | | | | | Allele $A_2$ Carrier Status | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1/A_1A_2$ | $A_2A_2$ | p | OR[(a)] | 95% CI | $A_1A_1$ | $A_1A_2/A_2A_2$ | p | OR[(b)] | 95% CI |
| IL10RA | Patient | 278 (75.0) | 93 (25.0) | 0.9636 | 1.01 | (0.71,1.44) | 86 (23.1) | 285 (76.9) | 0.2659 | 0.82 | (0.56,1.19) |
| rs2256111 A/G | Control | 261 (74.8) | 88 (25.2) | | | | 69 (19.8) | 280 (80.2) | | | |
| IL10RA | Patient | 356 (95.0) | 19 (5.0) | 0.0500 | 0.44 | (0.17,1.08) | 255 (68.0) | 120 (32.0) | 0.3557 | 0.86 | (0.63,1.19) |
| rs4252272 A/G | Control | 341 (97.7) | 8 (2.3) | | | | 226 (64.8) | 123 (35.2) | | | |
| IL10RA | Patient | 340 (90.2) | 37 (9.8) | 0.2893 | 0.75 | (0.43,1.31) | 171 (45.4) | 206 (54.6) | 0.6610 | 0.94 | (0.69,1.27) |
| rs2229113 G/A | Control | 317 (92.4) | 26 (7.6) | | | | 150 (43.7) | 193 (56.3) | | | |
| IL10RA | Patient | 304 (81.5) | 69 (18.5) | 0.9639 | 0.99 | (0.67,1.47) | 113 (30.3) | 260 (69.7) | 0.4977 | 0.89 | (0.64,1.25) |
| rs9610 G/A | Control | 280 (81.6) | 63 (18.4) | | | | 96 (28.0) | 247 (72.0) | | | |
| CD3E | Patient | 296 (83.6) | 58 (16.4) | 0.2704 | 0.79 | (0.51, 1.23) | 118 (33.3) | 236 (66.7) | 0.0881 | 1.31 | (0.95, 1.81) |
| ht1 G/A | Control | 291 (86.6) | 45 (13.4) | | | | 133 (39.6) | 203 (60.4) | | | |
| CD3E | Patient | 289 (90.9) | 29 (9.1) | 0.2727 | 0.73 | (0.40,1.33) | 158 (49.7) | 160 (50.3) | 0.3266 | 0.86 | (0.62,1.18) |
| ht2 T/C | Control | 315 (93.2) | 23 (6.8) | | | | 155 (45.9) | 183 (54.l) | | | |
| CD3DG | Patient | 330 (93.0) | 25 (7.0) | 0.4035 | 1.27 | (0.70, 2.28) | 193 (54.4) | 162 (45.6) | 0.2329 | 0.83 | (0.61,1.14) |
| ht1 G/C | Control | 313 (91.3) | 30 (8.8) | | | | 171 (49.9) | 172 (50.2) | | | |
| CD3DG | Patient | 349 (96.7) | 12 (3.3) | 0.2966 | 0.61 | (0.21, 1.68) | 240 (66.5) | 121 (33.5) | 0.3772 | 1.15 | (0.83,1.61) |
| ht3 G/A | Control | 335 (98.0) | 7 (2.1) | | | | 238 (69.6) | 104 (30.4) | | | |

| SNP | | Allele $A_1$ Carrier Status | | | | | Allele $A_2$ Carrier Status | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1/A_1A_2$ | $A_2A_2$ | p | OR[(a)] | 95% CI | $A_1A_1$ | $A_1A_2/A_2A_2$ | p | OR[(b)] | 95% CI |
| CD3DG ht4 G/A | Patient | 306 (87.2) | 45 (12.8) | 0.6071 | 0.89 | (0.55, 1.44) | 147 (41.9) | 204 (58.1) | 0.3199 | 0.86 | (0.62,1.18) |
| | Control | 299 (88.5) | 39 (11.5) | | | | 129 (38.2) | 209 (61.8) | | | |
| CD3DG ht5 T/C | Patient | 326 (91.6) | 30 (8.4) | 0.9988 | 1.00 | (0.57, 1.76) | 187 (52.5) | 169 (47.5) | 0.8592 | 1.03 | (0.75,1.40) |
| | Control | 315 (91.6) | 29 (8.4) | | | | 183 (53.2) | 161 (46.8) | | | |
| THY1 ht2 G/A | Patient | 332 (89.7) | 38 (10.3) | 0.2137 | 0.72 | (0.41,1.25) | 161 (43.5) | 171 (51.5) | 0.5321 | 1.10 | (0.80,1.52) |
| | Control | 316 (92.4) | 26 (7.6) | | | | 161 (51.0) | 155 (49.1) | | | |
| THY1 ht7 G/A | Patient | 295 (82.9) | 61 (17.1) | 0.3118 | 0.81 | (0.52,1.25) | 122 (34.3) | 234 (65.7) | 0.3191 | 1.17 | (0.85,1.62) |
| | Control | 287 (85.7) | 48 (14.3) | | | | 127 (37.9) | 208 (62.1) | | | |
| IL18 ht1 A/G | Patient | 292 (89.0) | 36 (11.0) | 0.8555 | 0.95 | (0.58,1.58) | 145 (44.2) | 183 (55.8) | 0.1228 | 0.78 | (0.56,1.08) |
| | Control | 272 (89.5) | 32 (10.5) | | | | 116 (38.2) | 188 (61.8) | | | |
| IL18 ht2 G/C | Patient | 289 (94.4) | 17 (5.6) | 0.0600 | 1.78 | (0.93,3.43) | 154 (50.3) | 152 (49.7) | 0.2785 | 1.19 | (0.86,1.64) |
| | Control | 305 (90.5) | 32 (9.5) | | | | 184 (54.6) | 153 (45.4) | | | |
| IL18 ht3 A/G | Patient | 295 (98.3) | 5 (1.7) | 0.4510 | 0.41 | (0.04,2.55)* | 227 (75.7) | 73 (24.3) | **0.0490** | 1.49 | (0.98,2.28) |
| | Control | 286 (99.3) | 2 (0.7) | | | | 237 (82.3) | 51 (17.7) | | | |
| IL18 ht4 G/T | Patient | 384 (87.1) | 57 (12.9) | **0.0159** | 1.55 | (1.08,2.23) | 166 (37.6) | 218 (56.8) | 0.0568 | 0.77 | (0.58,1.02) |
| | Control | 390 (81.7) | 90 (18.8) | | | | 152 (39.0) | 238 (61.0) | | | |

a; Odds ratio for comparision of $A_1A_1$/ $A_1A_2$ to $A_2A_2$ , b; Odds ratio for $A_1A_1$/ $A_1A_2$ compared to $A_1A_1$; * = Fisher's exact test and exact confidence limits used

**Figure 4.8 P-values and odds ratios for allele frequency differences between patient and control groups for SNP loci at TNFSF14**




Actual allele frequencies listed in table 4.5

**Table 4.5 Genotype and allele frequencies in the *TNFSF14* gene in the coeliac disease patient and control populations**

| | | Genotype n (%) | | | | Allele n (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | p | $A_1$ | $A_2$ | p | OR | 95% CI |
| TNFSF14 | Patient | 184 (50.6) | 158 (43.4) | 22 (6.0) | **0.0096** | 526 (72.3) | 202 (27.8) | **0.0052** | 0.70 | (0.54,0.91) |
| rs2279627C/G | Control | 203 (62.1) | 109 (33.3) | 15 (4.6) | | 515 (78.8) | 139 (21.3) | | | |
| TNFSF14 | Patient | 332 (89.3) | 39 (10.5) | 1 (0.3) | 0.6396 | 703 (94.5) | 41 (5.5) | 0.8425 | 0.95 | (0.59,1.55) |
| rs344560G/A | Control | 297 (89.5) | 35 (10.5) | 0 (0.0) | | 629 (94.7) | 35 (5.3) | | | |
| TNFSF14 | Patient | 251 (67.5) | 114 (30.7) | 7 (1.9) | 0.1611 | 616 (82.8) | 128 (17.2) | 0.2641 | 1.17 | (0.88,1.55) |
| ht1_rs2291668C/T | Control | 209 (65.3) | 97 (30.3) | 14 (4.4) | | 515 (80.5) | 125 (19.5) | | | |
| TNFSF14 | Patient | 115 (31.7) | 183 (50.4) | 65 (17.9) | 0.0669 | 413 (56.9) | 313 (43.1) | 0.0515 | 1.24 | (1.00,1.53) |
| ht2_rs3760746T/C | Control | 89 (28.4) | 145 (46.3) | 79 (25.2) | | 323 (51.6) | 303 (48.4) | | | |
| TNFSF14 | Patient | 122 (33.3) | 185 (50.6) | 59 (16.1) | **0.0106** | 429 (58.6) | 303 (41.4) | **0.0092** | 1.33 | (1.07,1.67) |
| ht3_rs11878563G/A | Control | 86 (28.5) | 139 (46.0) | 77 (25.5) | | 311 (51.5) | 293 (48.5) | | | |

n = number of individuals; p = p-value, CI = confidence interval

**Table 4.6 Carrier status frequencies in the *TNFSF14* gene in the coeliac disease patient and control populations**

| | | Allele $A_1$ Carrier Status | | | | | Allele $A_2$ Carrier Status | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1/A_1A_2$ | $A_2A_2$ | p | OR[(a)] | 95% CI | $A_1A_1$ | $A_1A_2/A_2A_2$ | p | OR[(b)] | 95% CI |
| TNFSF14 | Patient | 342 (94.0) | 22 (6.0) | 0.3957 | 0.75 | (0.36,1.53) | 184 (50.6) | 180 (49.5) | **0.0023** | 1.60 | (1.17,2.20) |
| rs2279627 | Control | 312 (95.4) | 15 (4.6) | | | | 203 (62.1) | 124 (37.9) | | | |
| TNFSF14 | Patient | 371 (99.7) | 1 (0.3) | 1.0000 | 0.00 | (0.00,43.7) | 332 (89.3) | 40 (10.8) | 0.9280 | 1.02 | (0.62,1.70) |
| rs344560 | Control | 332 (1.00) | 0 (0.0) | | | | 297 (89.5) | 35 (10.5) | | | |
| TNFSF14 | Patient | 365 (98.1) | 7 (1.9) | 0.0566 | 2.39 | (0.89,6.61) | 251 (68.8) | 114 (31.2) | 0.5483 | 0.91 | (0.65,1.26) |
| ht1_rs2291668 | Control | 306 (95.6) | 14 (4.4) | | | | 209 (65.3) | 111 (34.7) | | | |
| TNFSF14 | Patient | 298 (82.1) | 65 (17.9) | **0.0202** | 1.55 | (1.07,2.24) | 115 (31.7) | 248 (68.3) | 0.3593 | 0.86 | (0.61,1.21) |
| ht2_rs3760746 | Control | 234 (74.8) | 79 (25.2) | | | | 89 (28.4) | 224 (71.6) | | | |
| TNFSF14 | Patient | 307 (83.9) | 59 (16.1) | **0.0027** | 1.78 | (1.20,2.65) | 122 (39.7) | 185 (60.3) | 0.1773 | 0.80 | (0.56,1.12) |
| ht3_rs11878563 | Control | 225 (74.5) | 77 (25.5) | | | | 86 (38.2) | 139 (61.8) | | | |

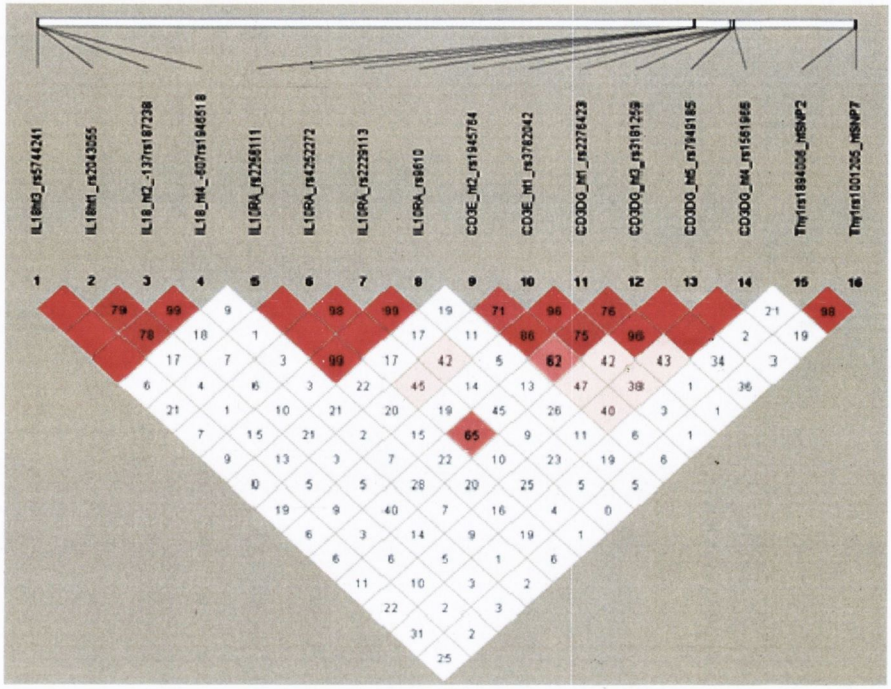
a; Odds ratio for comparision of $A_1A_1/ A_1A_2$ to $A_2A_2$ , b; Odds ratio for $A_1A_1/ A_1A_2$ compared to $A_1A_1$

### 4.3.7 Linkage Disequilibrium Analysis

Linkage disequilibrium (LD) analysis was carried out on the candidate genes in chromosome 11 using Haploview (Barrett *et al.* 2004). Levels of LD observed between SNPs in the patient population are shown in figure 4.9a and the control population results are shown in figure 4.9b. These figures show pairwise LD levels between each pair of SNPs as measured by the D' statistic computed from the genotype data. The D' value for each pair of SNPs is shown in the corresponding diamond. Diamonds not displaying a number indicate a D' value of 1. Diamonds with a blue shading represents D' = 1 and LOD < 2, while a bright red shading represents D' = 1 and LOD > 2. White diamonds represent D' < 1 and LOD < 2, and various shades of pink and red represent D' <1 and LOD >2. The darker the shading of red, the higher the associated D' value. This standard colour scheme used by the Haploview program is outlined in the legend accompanying the LD figure below (figure 4.9).

These results show similar patterns of LD for the patient and control groups. Every gene in both populations shows high levels of intra-gene LD (generally, D'>.7). All *IL18* SNPs analysed are in strong LD, a similar pattern is seen for *IL10RA*, the *CD3* genes and *THY1*. There is little inter-gene LD. The *IL10RA* and *THY1* genes are not in LD with any other gene analysed. There is some LD between the three *CD3* genes. The *CD3E* gene SNPs are in stronger LD with the proximal *CD3D/G* SNPs (CD3DG_ht1_rs2276423 and CD3DG_ ht3_rs3181259), however, this LD pattern does not extend to the distal end of this gene cluster, although there is some LD present between all three genes. The pattern of LD observed in this region suggests that it would be sensible to take a gene based haplotype analysis approach (including an analysis examining all three *CD3* genes simultaneously) when testing for a haplotype association with disease. LD analysis was also carried out in the *TNFSF14* (*LIGHT*) gene (figure 4.10). High levels of LD were observed for all SNPs except rs344560 where lower levels of LD were observed between this SNP and the remainder, a finding that was especially pronounced in the control population.

a




b

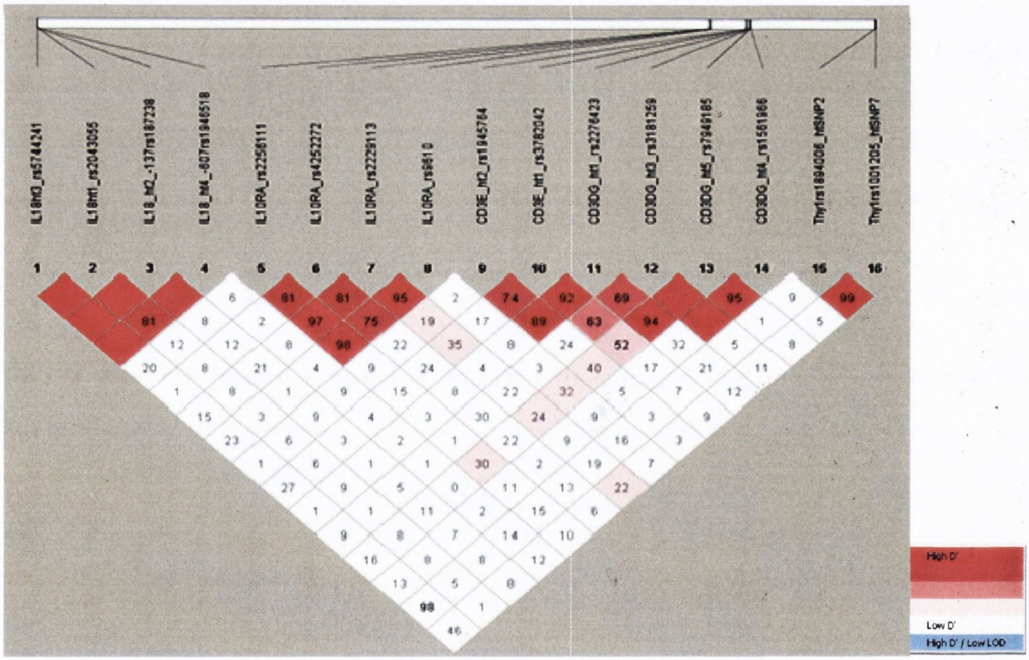



**Figure 4.9 Linkage disequilibrium analysis of the 11q23 region in coeliac disease cases (a) and controls (b)**

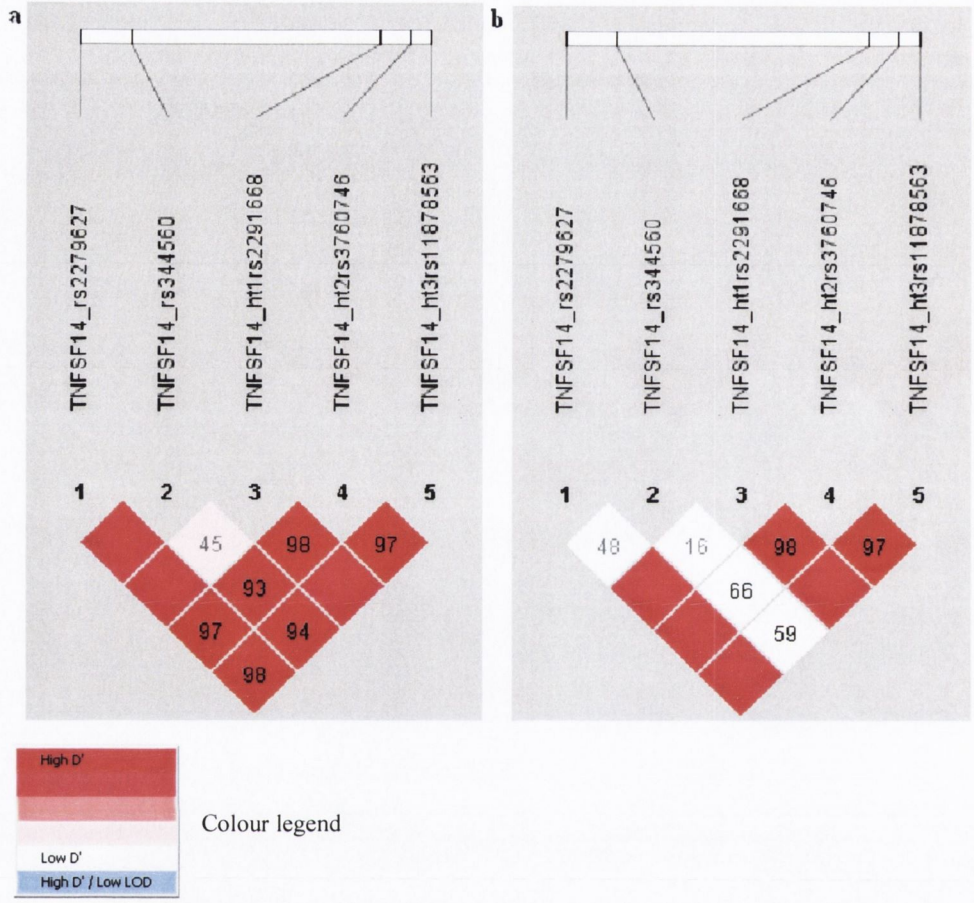



**Figure 4.10 Linkage disequilibrium analysis of the *TNFSF14* (*LIGHT*) gene in coeliac disease cases (a) and controls (b)**

### 4.3.8 Haplotype analysis of the candidate genes on chromosome 11

Haplotype reconstruction was carried out for each of the candidate genes using the EM algorithim as outlined in section 2.12.3 to establish if an overall gene (i.e. haplotype) effect was associated with disease. The frequency of the *CD3E* haplotype, consisting of two htSNPs, was not found to differ significantly between the patient and control groups. Results are shown in figure 4.11 and table 4.7. Genes, *CD3D* and *CD3G*, are located 1.8 kb apart on chromosome 11q23.3, hence a haplotype consisting of both genes was considered in association testing. Due to genotyping assay failure, genotype data for CD3DG_ht2_rs3181261 was not available. Hence, haplotype 4; G<u>A</u>ATA (Perlegen CEU sample frequency; 5.7%) and haplotype 6; G<u>G</u>ATA (Perlegen CEU sample frequency; 4.2%) consisting of the 5 htSNPs, as identified in figure 4.2, are

indistinguishable from each other as they both condense to a GATA haplotype, upon removal of the CD3DG_ht2_rs3181261 (underlined) polymorphism. However, as this haplotype shows no apparent frequency difference between the groups, it was not considered necessary to differentiate between these two haplotypes. Also, as all other loci on the two condensed haplotypes share identical alleles (i.e. this is the only SNP site that differs between these two haplotypes), it is not necessary to differentiate between them to predict the allele present at other non-genotyped loci. No *CD3D/CD3G* haplotype was found to be associated with disease. Results are shown in 4.8. A haplotype constructed of all three *CD3* genes, *CD3E*, *CD3D* and *CD3G* also failed to show a significant association with disease as shown in table 4.9.

Haplotype analysis of the *THY1* gene consisted of haplotype reconstruction using the two htSNPs as outlined in figure 4.3. No significant association with disease was observed as detailed in table 4.10.

For the *IL10RA* gene, six htSNPs were identified to tag a 31 SNP haplotype. However, htSNP IL10RA_ rs4252314 had a minor allele frequency <3% as reported on the NCBI dbSNP database, hence, this SNP was not assayed. Also htSNP IL10RA_ rs2229114 failed to amplify and therefore, genotype information was not available for this polymorphism. Upon removal of these two htSNPs, IL10RA_rs4252314 and IL10RA_rs2229114, haplotype 1; GAAGCA and haplotype 8; GAAGTA constructed from the 6 htSNPs as outlined in figure 4.4, merge to form haplotype GAGA, and haplotype 2 (AGAACG) and haplotype 6 (AGGACG) merge to form haplotype AGAG. As frequencies for all haplotypes observed in the *IL10RA* gene are very similar between patient and control groups, and no significant association is observed between any haplotype and disease status, and indeed between individual gene analysis and disease (tables 4.3 and 4.4), this gene does not show evidence of involvement in coeliac disease and it was considered unlikely that further investigation of this gene through the differentiation of these haplotypes would be worthwhile. Results for *IL10RA* haplotypes are shown in table 4.11.

Haplotype analysis of the *IL18* gene consisted of the analysis of two components of the gene. Total gene haplotype analysis was carried out, constructing haplotypes using all

four SNPs genotyped in this gene. A small but significant association was observed between disease and haplotype 4; TGAA (patients 9.0%, controls 13.6%, p-value = 0.0379) as outlined in table 4.12. A separate analysis of the promoter region of the gene was then carried out using htSNPs located in the promoter at positions -607 and -137 as shown in table 4.13. A significant association was observed between disease and two promoter haplotypes (haplotype 1; GG) (patients 61.8%, controls 55.3%, p-value = 0.0153), (haplotype 3; TG) (patients 10.7%, controls 16.2%, p-value = 0.0053).

### 4.3.9 Haplotype analysis of the *TNFSF14* gene

Haplotype reconstruction was carried out for the *TNFSF14* gene to establish if an overall gene (i.e. haplotype) effect was associated with disease. This analysis revealed a significant association between disease status and the htSNP haplotype composed of SNPs TNFSF14_ht1rs2291668, TNFSF14_ht2rs3760746 and TNFSF14_ht3rs11878563 (table 4.14). Haplotype 1; CTG was significantly associated with disease (patients 55.7%, controls 49.7%, p-value = 0.0221), as was haplotype 2; CCA (patients 23.1%, controls 28.1%, p-value = 0 .0309). Analysis of haplotypes composed of all SNPs in the gene also revealed a significant association with disease (table 4.15). Haplotype 2; GGCTG was significantly associated with disease (patients 27.1%, controls 21.1%, p-value = 0.0090), as was haplotype 3; CGCTA (patients 20.9%, controls 26.6%, p-value = 0.0054).

**Table 4.7 CD3E haplotype frequency estimations (for haplotypes with predicted frequency >5%) and haplotype association test results for patient and control populations**

| CD3E_ht2_rs1945764 | CD3E_ht1_rs3782042 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|
| | | | | Patient | Control |
| T | A | 0.19853500 | 0.1541 | 0.3774 | 0.3429 |
| T | G | 0.34420100 | 0.2968 | 0.3229 | 0.3499 |
| C | G | 0.52076298 | 0.4967 | 0.2608 | 0.2778 |

**Table 4.8 CD3D/CD3G haplotype frequency estimations (for haplotypes with predicted frequency >5%) and haplotype association test results for patient and control populations**

| CD3DG_ht1_rs2276423 | CD3DG_ht3_rs3181259 | CD3DG_ht5_rs7949185 | CD3DG_ht4_rs1561966 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|---|
| | | | | | | Patient | Control |
| G | G | C | G | 0.6997 | 0.6694 | 0.2770 | 0.2670 |
| C | G | T | G | 0.8459 | 0.8396 | 0.2125 | 0.2137 |
| G | A | T | A | 0.1090 | 0.0947 | 0.1761 | 0.1471 |
| G | G | T | G | 0.7939 | 0.7882 | 0.1502 | 0.1449 |
| G | G | T | A | 0.5487 | 0.5393 | 0.1284 | 0.1402 |

**Table 4.9 All CD3 genes haplotype frequency estimations (for haplotypes with predicted frequency >5%) and haplotype association test results for patient and control populations**

| CD3E_ht1 | CD3E_ht2 | CD3DG_ht1 | CD3DG_ht3 | CD3DG_ht4 | CD3DG_ht5 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Patient | Control |
| G | T | C | G | G | T | 0.9763 | 0.9800 | 0.1939 | 0.1861 |
| G | C | G | G | G | C | 0.5699 | 0.5550 | 0.1797 | 0.1600 |
| A | T | G | A | A | T | 0.1201 | 0.1050 | 0.1512 | 0.1232 |
| A | T | G | G | G | T | 0.6553 | 0.6320 | 0.1013 | 0.1073 |
| A | T | G | G | A | T | 0.8204 | 0.8330 | 0.0607 | 0.0557 |

**Table 4.10 THY1 haplotype frequency estimations (for haplotypes with predicted frequency >5%) and haplotype association test results for patient and control populations**

| Thy1rs1001205_htSNP7 | Thy1rs1894006_htSNP2 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|
| | | | | Patient | Control |
| G | G | 0.4854 | 0.3605 | 0.5874 | 0.6113 |
| A | A | 0.1897 | 0.1437 | 0.3332 | 0.2999 |
| A | G | 0.4173 | 0.4089 | 0.0752 | 0.0870 |

**Table 4.11 IL10RA haplotype frequency estimations (for haplotypes with predicted frequency >5%) and haplotype association test results for patient and control populations**

| IL10RA _rs2256111 | IL10RA _rs4252272 | IL10RA _rs2229113 | IL10RA _rs9610 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|---|
| | | | | | | Patient | Control |
| G | A | G | A | 0.9526 | 0.9466 | 0.4377 | 0.4395 |
| A | G | A | G | 0.9819 | 0.9806 | 0.1831 | 0.1827 |
| A | A | G | G | 0.2572 | 0.2385 | 0.1667 | 0.1469 |
| A | A | A | G | 0.9192 | 0.9155 | 0.1379 | 0.1347 |
| G | A | G | G | 0.4021 | 0.3899 | 0.0716 | 0.0827 |

**Table 4.12 IL18 htSNP haplotype frequency estimations (for haplotypes with predicted frequency >5%) and haplotype association test results for patient and control populations**

| IL18_ht4_ -607rs1946518 | IL18_ht2_ -137rs187238 | IL18_ht1_ rs2043055 | IL18_ht3_ rs5744241 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|---|
| | | | | | | Patient | Control |
| G | G | G | A | 0.6146 | 0.5879 | 0.3136 | 0.3039 |
| T | C | A | A | 0.1683 | 0.1367 | 0.2541 | 0.2986 |
| G | G | A | A | 0.1264 | 0.1109 | 0.1785 | 0.1351 |
| T | G | A | A | 0.0429 | 0.0379 | 0.0899 | 0.1357 |
| G | G | A | G | 0.2018 | 0.1919 | 0.1228 | 0.0957 |
| G | C | G | A | 0.6146 | 0.5879 | 0.3136 | 0.3039 |

**Table 4.13 IL18 promoter haplotype frequency estimations (for haplotypes with predicted frequency >5%) and haplotype association test results for patient and control populations**

| IL18_ht4_-607rs1946518 | IL18_ht2_-137rs187238 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|
| | | | | Patient | Control |
| G | G | 0.0654 | 0.0153 | 0.6181 | 0.5525 |
| T | C | 0.5889 | 0.5617 | 0.2728 | 0.2850 |
| T | G | 0.0060 | 0.0053 | 0.1071 | 0.1624 |

**Table 4.14 TNFSF14 htSNP haplotype frequency estimations (for haplotypes with predicted frequency >5%) and haplotype association test results for patient and control populations**

| TNFSF14_ht1 rs2291668 | TNFSF14_ht2 rs3760746 | TNFSF14_ht3 rs11878563 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|
| | | | | | Patient | Control |
| C | T | G | 0.06863800 | 0.02210000 | 0.5573 | 0.4969 |
| C | C | A | 0.04359300 | 0.03090000 | 0.2310 | 0.2811 |
| T | C | A | 0.37803599 | 0.35929999 | 0.1745 | 0.1931 |

**Table 4.15 TNFSF14 haplotype frequency estimations (for haplotypes with predicted frequency >5%) and haplotype association test results for patient and control populations**

| TNFSF14_rs2279627 | TNFSF14_rs344560 | TNFSF14_ht1 rs2291668 | TNFSF14_ht2 rs3760746 | tnfsf14_ht3 rs11878563 | p-value (F-test) | p-value (permutation) | Frequencies | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Patient | Control |
| C | G | C | T | G | 0.9646 | 0.9613 | 0.2839 | 0.2847 |
| G | G | C | T | G | 0.0149 | 0.0090 | 0.2707 | 0.2114 |
| C | G | C | C | A | 0.0089 | 0.0054 | 0.2087 | 0.2662 |
| C | G | T | C | A | 0.3711 | 0.3575 | 0.1410 | 0.1585 |

### 4.3.10 Inference of non-genotyped SNP frequencies using observed haplotype frequencies

In section 4.3.8 and 4.3.9, haplotype frequencies were estimated and these frequencies are now used to infer allele frequencies of non-genotyped SNPs using the method for the inference of allele frequencies of non-genotyped SNPs as outlined in section 2.12.4. All the non-genotyped SNPs included for analysis in this section are those shown within their respective haplotype structures in figures 4.1 to 4.6. A significant association with disease was observed with the IL18 SNPs as shown in tables 4.16 and 4.17 below. All SNPs in table 4.16 had an inferred patient frequency = 65.76% and inferred control frequency = 73.82%, p-value = 0.0137. These frequencies were inferred from haplotypes constructed using only fully genotyped individuals to ensure reduced error rates in inferring genotypes at non-genotyped loci. Also shown to have a significant association with disease was SNP IL18_rs3882891 (table 4.17), with inferred patient freq = 61.49% and inferred control freq = 53.47%, p-value = 0.0080. All other SNPs in the 11q23 region and in the *TNFSF14* gene, investigated using this method, were not found to have a significant association with disease (data not shown).

**Table 4.16 Inferred allele frequencies and association testing for IL18_rs5744230C/T, IL18_rs4988359A/G, IL18_rs5744249T/G, IL18_rs1834481G/C, IL18_rs5744258G/C, IL18_rs5744276G/C, IL18_rs3882892G/A.**

| Allele | Hap | Pat chr % | Pat Total n (%) | Con chr % | Con Total n (%) | p-value | OR(95%CI) |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 31.36 | | 30.39 | | | |
| | 2 | 25.41 | | 29.86 | | | |
| | 4 | 8.99 | | 13.57 | | | |
| | | | 292(65.76) | | 319(73.82) | | |
| 2 | 3 | 17.85 | | 13.51 | | | |
| | 5 | 12.28 | | 9.57 | | | |
| | | | 134(30.12) | | 100(23.08) | 0.0137 | 0.68(0.50,0.94) |

**Total number of patient and control chromosomes used to infer allele frequencies in patients and controls are n = 444 and n = 432, respectively; Haplotype frequencies were obtained from table 4.12; Allele 1 in table represents the first allelic variant in SNP title, allele 2 represents the second allelic variant in SNP title; Full haplotype structure obtained from figure 4.5**

**Table 4.17 Inferred Allele frequencies and association testing for IL18_rs3882891**

| Allele | Hap | Pat chr % | Pat Total n (%) | Con chr % | Con Total n (%) | p-value | OR(95%CI) |
|---|---|---|---|---|---|---|---|
| A | 1 | 31.36 | | 30.39 | | | |
| | 3 | 17.85 | | 13.51 | | | |
| | 5 | 12.28 | | 9.57 | | | |
| | | | 273(61.49) | | 231(53.47) | | |
| C | 2 | 25.41 | | 29.86 | | | |
| | 4 | 8.99 | | 13.57 | | | |
| | | | 153(34.40) | | 188(43.43) | 0.0080 | 1.45(1.09,1.93) |

**Total number of patient and control chromosomes used to infer allele frequencies in patients and controls are n = 444 and n = 432, respectively; Haplotype frequencies were obtained from table 4.12; Allele 1 in table represents the first allelic variant in SNP title, allele 2 represents the second allelic variant in SNP title; Full haplotype structure obtained from figure 4.5**

### 4.4 Gene-gene interaction effects in coeliac disease susceptibility genes

To test if there were gene-gene interaction effects present among those genes examined in this study, we selected the most statistically significant gene polymorphism from each of the regions to test if there were interactions between SNPs at different loci. The SNPs selected were the TNFSF14rs 2279627 SNP on chromosome 19, SNP IL18ht4_-607rs1946518 on chromosome 11 and SNP CTLA4_CT60 on chromosome 2, all of which have shown independent significant association with coeliac disease. While CTLA4-658 displayed the most highly significant association with disease on chromosome 2q33, this gene had so few individuals homozygous for the minor allele (15 individuals in total), that it was not included in this analysis. As all SNPs in each separate region are in high LD with each other (and thus not independent of each other), it is not possible to test for interactions between SNPs within the same region, as independence of variables (or a lack of correlation between variables) is a requirement necessary for this analysis. Hence, we attempt only to investigate interactions using SNPs in different regions.

A simple method to investigate interaction effects between variables such as the SNPs presented here, is to perform a crosstabulation of the data (as outlined in section 2.12.5). A crosstabulation was carried out to test if there was an interaction effect between SNPs, IL18ht4-607 and TNFSF14rs2279627, which have both been previously shown to be individually associated with disease. Interaction between the carrier statuses of the disease associated alleles at each SNP was tested. This included the G allele carrier status in both cases. The number of cases in each group are shown in table 4.18. Examination of an interaction effect between TNFSF14rs2279627 and IL18ht4-607, found the IL18ht4-607 G allele carrier status to have an odds ratio of 2.129 when controlling for the G allele carrier status at TNFSF14rs2279627 and an odds ratio of 1.825 when controlling for the C allele carrier status at TNFSF14rs2279627. (Table 4.19) The test for homogeneity of the odd ratios, testing for equality of the odds ratios across the groups found that there was no statistically significant disparity between the odds ratios in any group ($p = 0.721$), which is indicative that there is no interaction effect between these SNPs in disease outcome (Table 4.20).

**Table 4.18 Crosstabulation of disease status and IL18ht4-607, conditioning for TNFSF14rs2279627**

| TNFSF14 rs2279627 | | | | Status | | Total |
|---|---|---|---|---|---|---|
| | | | | cases | controls | |
| [a]G carrier status | IL18ht4_607 | G carrier status | Count | 155 | 91 | 246 |
| | | | % within Status | 88.6% | 78.4% | 84.5% |
| | | non-G carrier status | Count | 20 | 25 | 45 |
| | | | % within Status | 11.4% | 21.6% | 15.5% |
| | Total | | Count | 175 | 116 | 291 |
| | | | % within Status | 100.% | 100.% | 100% |
| [b]non-G carrier status | IL18ht4_607 | G carrier status | Count | 157 | 147 | 304 |
| | | | % within Status | 86.7% | 78.2% | 82.4% |
| | | non-G carrier status | Count | 24 | 41 | 65 |
| | | | % within Status | 13.3% | 21.8% | 17.6% |
| | Total | | Count | 181 | 188 | 369 |
| | | | % within Status | 100.% | 100.% | 100.% |

Section (a) reports the number and percentage of IL18ht4-607 G allele carriers and non-G allele carriers in patients and controls who also carry the G allele at TNFSF14rs2279627; Section (b) reports the number and percentage of IL18ht4-607 G allele carriers and non-G allele carriers in patients and controls who do not carry the G allele at TNFSF14rs2279627

**Table 4.19 Risk estimates for the crosstabulation of disease status and IL18ht4-607, conditioning for TNFSF14rs2279627**

| TNFSF14 rs2279627 | | Value | 95% Confidence Interval | |
|---|---|---|---|---|
| | | | Lower | Upper |
| G carrier status | Odds Ratio for IL18ht4_607 (G carrier status / non-G carrier status) | 2.129[c] | 1.120 | 4.048 |
| | For cohort Status = cases | 1.418[a] | 1.009 | 1.993 |
| | For cohort Status = controls | .666[b] | .489 | .906 |
| non-G carrier status | Odds Ratio for IL18ht4_607 (G carrier status / non-G carrier status) | 1.825[c] | 1.051 | 3.168 |
| | For cohort Status = cases | 1.399[a] | 1.000 | 1.957 |
| | For cohort Status = controls | .767[b] | .616 | .955 |

(a) Risk ratio for cases with the associated allele carrier statuses; (b) Risk ratio for controls with associated allele carrier statuses; (c) Odds ratio for the associated allele carrier statuses (odds ratio = risk ratio for cases/risk ratio for controls)

**Table 4.20 Tests of Homogeneity of the Odds Ratio for crosstabulation of disease status and IL18ht4-607, conditioning for TNFSF14rs2279627**

| | Chi-Squared | df | Asymp. Sig. (2-sided) (p-value) |
|---|---|---|---|
| Breslow-Day | .128 | 1 | .721 |
| Tarone's | .128 | 1 | .721 |

The interaction effect between TNFSF14rs2279627 and the CTLA4_CT60 SNP at chromosomal position 2q33 was also tested, as above. Carrier status for the A allele of CTLA4_CT60 has been shown previously in this study to be associated with a protective or neutral effect against disease, hence here it is postulated that interaction between non-carriage for the A allele at CTLA4_CT60 (i.e. GG genotype individuals) and G-allele carrier status at SNP TNFSF14rs227927 may led to increased disease association. Examination of an interaction effect between the carrier statuses for these SNPs, found the CTLA4_CT60 non-A allele carrier status to have an odds ratio of 1.161 when controlling for the G allele carrier status at TNFSF14rs2279627 and an odds ratio of 1.764 when controlling for the non-G allele carrier status at TNFSF14rs2279627 (table 4.22). The test for homogeneity of the odd ratios did not find any statistically significant disparity between the odds ratios in any group ($p = 0.218$) (table 4.23), which is indicative that there is no interaction effect between these SNPs in disease outcome.

**Table 4.21 Crosstabulation of disease status and CTLA4_CT60, conditioning for TNFSF14rs2279627**

| TNFSF14 rs2279627 | | | | Status | | Total |
|---|---|---|---|---|---|---|
| | | | | cases | controls | |
| [a]G carrier status | CTLA4 _CT60 | non-A carrier status | Count | 56 | 38 | 94 |
| | | | % within Status | 36.1% | 32.8% | 34.7% |
| | | A carrier status | Count | 99 | 78 | 177 |
| | | | % within Status | 63.9% | 67.2% | 65.3% |
| | Total | | Count | 155 | 116 | 271 |
| | | | % within Status | 100% | 100% | 100% |
| [b]non-G carrier status | CTLA4 _CT60 | non-A carrier status | Count | 75 | 61 | 136 |
| | | | % within Status | 44.9% | 31.6% | 37.8% |
| | | A carrier status | Count | 92 | 132 | 224 |
| | | | % within Status | 55.1% | 68.4% | 62.2% |
| | Total | | Count | 167 | 193 | 360 |
| | | | % within Status | 100% | 100% | 100% |

Section (a) reports the number and percentage CTLA4_CT60 non-A allele carriers and A allele carriers in patients and controls who also carry the G allele at TNFSF14rs2279627; Section (b) reports the number and percentage CTLA4_CT60 non-A allele carriers and A allele carriers in patients and controls who do not carry the G allele at TNFSF14rs2279627

**Table 4.22 Risk estimates for the crosstabulation of disease status and CTLA4_CT60, conditioning for TNFSF14rs2279627**

| TNFSF14rs2279627 | | Value | 95% Confidence Interval | |
|---|---|---|---|---|
| | | | Lower | Upper |
| G carrier status | Odds Ratio for CTLA4_CT60 (non-A carrier status / A carrier status) | 1.161[c] | .699 | 1.929 |
| | For cohort Status = cases | 1.065[a] | .862 | 1.316 |
| | For cohort Status = controls | .917[b] | .682 | 1.234 |
| non-G carrier status | Odds Ratio for CTLA4_CT60 (non-A carrier status / A carrier status) | 1.764[c] | 1.147 | 2.712 |
| | For cohort Status = cases | 1.343[a] | 1.080 | 1.670 |
| | For cohort Status = controls | .761[b] | .613 | .945 |

(a) Risk ratio for cases with the associated allele carrier statuses; (b) Risk ratio for controls with associated allele carrier statuses; (c) Odds ratio for the associated allele carrier statuses (odds ratio = risk ratio for cases/risk ratio for controls)

**Table 4.23 Tests of Homogeneity of the Odds Ratio for crosstabulation of disease status and CTLA4_CT60, conditioning for TNFSF14rs2279627**

| | Chi-Squared | df | Asymp. Sig. (2-sided) (p-value) |
|---|---|---|---|
| Breslow-Day | 1.520 | 1 | .218 |
| Tarone's | 1.520 | 1 | .218 |

A final two-way interaction effect analysis using crosstabulation was carried out using IL18ht4-607 (G carrier status) and CTLA4_CT60 (non-A carrier status). The number of cases, and percentages, in each sub-group defined by carriage or non-carriage of the alleles at the two SNPs of interest are shown in table 4.24. Again, no significant disparity between the odds ratios for the groups was identified using the test for homogeneity of odds ratios ($p = 0.465$) (table 4.26). The IL18ht4-607 G allele carrier status was found to have an odds ratio of 1.476 when controlling for the non-A allele carrier status at CTLA4_CT60 and an odds ratio of 2.056 when controlling for the A allele carrier status at CTLA4_CT60 (table 4.25).

**Table 4.24 Crosstabulation of disease status and IL18ht4_607, conditioning for CTLA4_CT60**

| CTLA4_CT60 | | | | Status | | Total |
|---|---|---|---|---|---|---|
| | | | | cases | controls | |
| [a]non-A carrier status | IL18ht4_607 | G carrier status | Count | 121 | 82 | 203 |
| | | | % within Status | 87.1% | 82.0% | 84.9% |
| | | non-G carrier status | Count | 18 | 18 | 36 |
| | | | % within Status | 12.9% | 18.0% | 15.1% |
| | Total | | Count | 139 | 100 | 239 |
| | | | % within Status | 100% | 100% | 100% |
| [b]A carrier status | IL18ht4_607 | G carrier status | Count | 168 | 160 | 328 |
| | | | % within Status | 87.5% | 77.3% | 82.2% |
| | | non-G carrier status | Count | 24 | 47 | 71 |
| | | | % within Status | 12.5% | 22.7% | 17.8% |
| | Total | | Count | 192 | 207 | 399 |
| | | | % within Status | 100% | 100% | 100% |

Section (a) reports the number and percentage IL18ht4-607 G allele carriers and non-G allele carriers in patients and controls do not carry the A allele at CTLA4_CT60; Section (b) reports the number and percentage IL18ht4-607 G allele carriers and non-G allele carriers in patients and controls who do carry the A allele at CTLA4_CT60

**Table 4.25 Risk estimates for the crosstabulation of disease status and IL18ht4-607, conditioning for CTLA4_CT60**

| CTLA4_CT60 | | Value | 95% Confidence Interval | |
|---|---|---|---|---|
| | | | Lower | Upper |
| non-A carrier status | Odds Ratio for IL18ht4_607 (G carrier status / non-G carrier status) | 1.476[c] | .725 | 3.004 |
| | For cohort Status = cases | 1.192[a] | .844 | 1.684 |
| | For cohort Status = controls | .808[b] | .560 | 1.166 |
| A carrier status | Odds Ratio for IL18ht4_607 (G carrier status / non-G carrier status) | 2.056[c] | 1.202 | 3.519 |
| | For cohort Status = cases | 1.515[a] | 1.076 | 2.134 |
| | For cohort Status = controls | .737[b] | .603 | .900 |

(a) Risk ratio for cases with the associated allele carrier statuses; (b) Risk ratio for controls with associated allele carrier statuses; (c) Odds ratio for the associated allele carrier statuses (odds ratio = risk ratio for cases/risk ratio for controls)

**Table 4.26 Tests of Homogeneity of the Odds Ratio for Crosstabulation of disease status and IL18ht4-607, conditioning for CTLA4_CT60**

| | Chi-Squared | df | Asymp. Sig. (2-sided) (p-value) |
|---|---|---|---|
| Breslow-Day | .534 | 1 | .465 |
| Tarone's | .534 | 1 | .465 |

A three-way interaction of all SNPs analysed above was also carried out. This analysis, too, failed to identify any significant interaction using the test for homogeneity of odds ratios ($p = 0.418$) (table not shown).

The examination of interaction effects between certain SNPs can also be performed using logistic regression analysis (as described in section 2.12.5). Here, the value of the addition to the model of an interaction term, containing the SNPs of interest, to predict disease outcome is assessed using the Wald statistic with associated p-value and odds ratios (as shown in tables below). This method was preformed for each of the SNP combinations assessed above and the results confirm those seen in the crosstabulation analysis. None of the interaction terms was not found to be a statistically significant indicator of disease outcome, indicating that there are no significant interaction effects between the SNPs analysed here which contribute to disease outcome. Tables 4.27, 4.28 and 4.29 show the final logistic regression model for each of the interaction combinations, showing the individual SNPs and the SNP-SNP interaction term as the indicator variables. Significance levels as assessed using the Wald statistic show that the interaction effects are not associated with disease outcome, and the observed p-values (highlighted in bold) are identical to those obtained using the crosstabulation method.

**Table 4.27 Significance of indicator variables in the logistic regression model for disease status using SNPs TNFSF14rs2279627 and IL18ht4-607**

| | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| TNFSF14rs2279627(1) | .467 | .175 | 7.117 | 1 | .008 | 1.595 | 1.132 | 2.247 |
| IL18ht4_607(1) | .756 | .328 | 5.315 | 1 | .021 | 2.129 | 1.120 | 4.048 |
| IL18ht4_607(1) by TNFSF14rs2279627(1) | .154 | .432 | .128 | 1 | **.721** | .857 | .367 | 1.999 |
| Constant | .533 | .132 | 16.262 | 1 | .000 | .587 | | |

B = the coefficient or the estimated log odds ratio for each variable; S.E. = standard error of B; Wald = Wald test for association for each variable; df = degrees of freedom; Sig. = significance of each variable; Exp(B) = odds ratio; 95% C.I. = 95% confidence interval for the odds ratio for EXP(B)

**Table 4.28 Significance of indicator variables in the logistic regression model for disease status using SNPs TNFSF14rs2279627 and CTLA4_CT60**

| | B | S.E. | Wald | df | Sig. | Exp(B) | 95.0% C.I | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| TNFSF14rs2279627(1) | .599 | .203 | 8.686 | 1 | .003 | .549 | .369 | .818 |
| CTLA4_CT60(1) | .568 | .219 | 6.689 | 1 | .010 | .567 | .369 | .872 |
| CTLA4_CT60(1) by TNFSF14rs2279627(1) | .418 | .340 | 1.518 | 1 | **.218** | 1.519 | .781 | 2.956 |
| Constant | .361 | .136 | 7.066 | 1 | .008 | 1.435 | | |

B = the coefficient or the estimated log odds ratio for each variable; S.E. = standard error of B; Wald = Wald test for association for each variable; df = degrees of freedom; Sig. = significance of each variable; Exp(B) = odds ratio; 95% C.I. = 95% confidence interval for the odds ratio for EXP(B)

**Table 4.29 Significance of indicator variables in the logistic regression model for disease status using SNPs IL18ht4-607 and CTLA4_CT60**

| | B | S.E. | Wald | df | Sig. | Exp(B) | 95.0% C.I | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| IL18ht4_607(1) | .721 | .274 | 6.916 | 1 | .009 | 2.056 | 1.202 | 3.519 |
| CTLA4_CT60(1) | .340 | .181 | 3.545 | 1 | .060 | .712 | .499 | 1.014 |
| CTLA4_CT60(1) by IL18ht4_607(1) | .332 | .455 | .533 | 1 | .466 | .718 | .294 | 1.749 |
| Constant | .049 | .110 | .195 | 1 | .659 | .952 | | |

B = the coefficient or the estimated log odds ratio for each variable; S.E. = standard error of B; Wald = Wald test for association for each variable; df = degrees of freedom; Sig. = significance of each variable; Exp(B) = odds ratio; 95% C.I. = 95% confidence interval for the odds ratio for EXP(B)

**4.5 Discussion**

Chromosomal region 11q23 has previously been shown to be linked to coeliac disease in three independent genome scans (Greco *et al.* 1998; Holopainen *et al.* 2001; Naluai *et al.* 2001), and it contains a number of interesting genes with immunological functions which could be considered as good candidates in CD pathogenesis. This study is the first thorough investigation carried out on this region and contains genetic analysis of six candidate genes; *CD3E*, *CD3D*, *CD3G*, *IL10RA*, *THY1* and *IL18*. Also included in this study is a gene on 19q13.3, *TNFSF14* (*LIGHT*), which is of intense recent interest as it has been shown to induce an intestinal inflammatory phenotype in mice and hence could play a role in CD pathology. This gene occurs within a linkage region for inflammatory bowel disease.

A thorough genetic analysis of the chosen candidate genes was carried out, utilising the genetic information available on these genes at the present time. Haplotype tagging strategies were used to define haplotypes around the genes of interest ensuring the maximum information content is gained with less genotyping efforts. Polymorphisms previously shown to be associated with disease or those with a theoretical or proven functional role (e.g. gene promoter regulation or transcription factor binding site alteration) have been included where possible in this study to provide a comprehensive picture of the genetic variation present in the region.

In this study, 21 SNPs were tested for association with disease. Polymorphisms in two genes were shown to have a significant disease association. The promoter polymorphisms in *IL18*, -137 and -607, were both found to be associated with CD, before correction for multiple testing. These polymorphisms have been extensively investigated in other autoimmune diseases with varying results. One previous study (Rueda *et al.* 2005) has examined the relationship between *IL18* and CD by examination of these two promoter polymorphisms in a Spanish population, and found no association with disease. In contrast, here we have found a positive disease association. No other gene in the 11q23 region was found to contain disease associated polymorphisms. Analysis of the *TNFSF14* (*LIGHT*) gene did, however, highlight two polymorphisms which were associated with disease. The TNFSF14_rs2279627 SNP had the highest frequency differential between patient and control groups of all SNPs assessed. A

differential of 11.6% (patients 49.5%, controls 37.9%, p-value = 0.0023) was observed when carrier status for the minor allele (G allele) was examined. Similarly, significant differences were observed at the allele and genotype level. This polymorphism was identified on the HapMap project and found to occur 3' of the *TNFSF14* gene. Association with disease status was also observed for the ht3_rs11878563 and ht2_rs3760746 SNPs. The non-synonymous SNP rs344560 was not associated with disease and, as this is the only non-synonymous SNP with a minor allele frequency greater than 5% identified in the gene to date, this suggests that the association observed here is not due to an amino acid substitution in the gene, but rather is more probably due to DNA sequence changes affecting *TNFSF14* expression in mucosal tissue. However, it should be noted that there is, at present, no information regarding the functional effects of any of these variants and indeed only a limited amount of data is available on genetic variation in *TNFSF14* at this time. Additional data on the genetic variation will be required if a more detailed genetic analysis of this gene is to be carried out with the aim of conclusively proving an association and identifying an actual functional susceptibility locus.

Linkage disequilibrium analysis of the 11q23 region showed that all SNPs within each gene were in strong LD with other SNPs in the same gene, however, there was little to no inter-gene LD present, except in the *CD3* gene cluster. Thus the haplotype structure of each gene was assessed separately for association with disease. *IL18* haplotype analysis revealed a significant association between disease and the promoter haplotype composed of SNPs -607 and -137 in the gene. Hap1; GG was associated with a disease phenotype while hap3; TG was associated with a protective or neutral phenotype (table 4.13). Since the frequency differential observed between groups for the associated haplotypes are similar to those observed for the -607 polymorphism, and as the associated haplotypes differed only at the -607 position and not at the -137 position, it is probable that the disease association observed for *IL18* is most likely a single point association rather than a haplotype effect. Analysis of haplotypes composed of all SNPs in the gene also revealed a haplotype which was significantly associated with disease (Hap4; TGAA, frequency differential = 4.58%, p-value = 0.0379) (table 4.12). The addition of the extra htSNPs does not increase the frequency difference between groups when compared to that observed for the promoter haplotype -607T -137G (5.53%

frequency differential), or the single -607G allele (5.9% differential). Addition of the extra SNPs actually results in a slight dilution in the difference between groups, though the effect is not very pronounced, perhaps owing to the high levels of LD (as shown in figure 4.9) which results in alleles at linked loci being carried largely on the same haplotype. Notably, the -607G allele occurs only on haplotypes which show an increase in frequency in the patient group compared to the control group and the -607T allele occurs only on haplotypes with decreased frequency in the patient group. No other selected SNP in the *IL18* gene followed this trend. Using the method of inference of non-genotyped allele frequencies from haplotype frequencies (section 2.12.4), only one other SNP (IL18rs3882891) from over 20 SNPs tested was found to have the same allelic spread over haplotypes as the -607 SNP. This is an intronic SNP with no obvious function and its exact allelic distribution would need to be confirmed by direct genotyping.

This analysis of *IL18* has been quite extensive with the direct and indirect analysis of 27 SNPs in the gene by haplotype tagging methods. The conclusion of this analysis is that the most likely disease susceptibility locus in this gene is -607, a promoter polymorphism which has also been shown previously to have a functional effect on the levels of IL18 mRNA produced (Gierdraitis *et al.* 2001).

Haplotype analysis of the *LIGHT* gene using the htSNPs revealed a significant association of disease with hap1; CTG associated with disease and hap2; CCA, associated with a protective or neutral status (table 4.14). As was the case for *IL18*, the frequency differential observed between cases and controls for these haplotypes (hap1, 6.04%; hap2, 5.01%) is similar but slightly less than that seen in the allele frequency difference for a disease associated allele ht3_rs11878563 alone (7.10%). It should also be noted that the T allele carrier frequency of ht2_rs3760746 shows a significant frequency difference between groups (7.3%). Haplotype analysis of the htSNPs does not further increase the differential between groups beyond that seen with the individual disease associated SNPs, hence it is possible that the association observed with the haplotypes is likely to be the result of the independently increased ht3_rs11878563 G allele and ht2_rs3760746 T allele in the patient compared to the control population.

Analysis of the *LIGHT* gene using all *LIGHT* SNPs also revealed two significantly associated haplotypes, hap 2; GGCTG is associated with disease status and hap 3; CGCCA is associated with a protective or neutral effect (table 4.15). Again, these haplotypes consist of the disease associated loci, ht2_rs3760746 and ht3_rs11878563 and now also rs2279627, a SNP which is also independently associated with disease. The frequency gap between these associated haplotypes and the individual SNPs are all similar, with the individuals SNPs independently showing a slightly higher difference between groups. As the individual SNPs found to be associated with disease in this study are in LD with each other, this result is not unexpected, as all alleles at the three disease associated loci which are overrepresented in the patients (rs2279627G, ht2_rs3760746T and ht3_rs11878563G) are found on the haplotype also found to be overrepresented in the patient population. This makes it difficult to assess which, if any, allele is the actual disease susceptibility allele by a genetic analysis alone. Both ht2_ rs3760746 and ht3_rs11878563 SNPs are located in the 5' untranslated region and rs2279627 is located in the 3' region of the gene, hence all polymorphisms may have potential functional roles in regulation of gene activity which may implicate them in disease pathogenesis. As previously discussed, it will be difficult to disentangle the effects of these SNPs in the absence of functional studies and additional variation in the *LIGHT* gene may also need to be identified and tested for association with disease status.

In conclusion, no previous genetic association studies have been carried out on the role of the *LIGHT* gene in autoimmune disease but the results shown here indicate that there is strong evidence suggesting this gene is a susceptibility locus for CD in this population. These results do, however, need to be replicated in additional populations or by functional analysis to validate this claim. In addition, this study supports previous findings of a role for *IL18* in autoimmune disease with the association of the -607 promoter polymorphism with coeliac disease in our population.

# CHAPTER V

## GENETIC ASSOCIATION ANALYSIS OF THE MHC CLASS III REGION

## 5.1 Introduction

### 5.1.1 Major Histocompatibility Complex Structure

The human major histocompatibility complex (MHC), which encodes, among other genes, the human leukocyte antigen complex (HLA), is located at chromosomal position 6p21, and is the most important region in the genome involved in protection against infection thus far identified. It is also strongly implicated in autoimmunity. The MHC region (which extends approximately 3.6 Mb) is one of the most gene dense in the genome and is known to contain at least 239 gene loci (TheMHCSequencingConsortium 1999). The MHC can be divided into three areas. The class II region is the most centromeric, encoding among others the HLA-DP, HLA-DQ, and HLA-DR loci which encode proteins expressed on the surface of antigen presenting cells and are essential in the presentation of peptides to T helper cells. In total, the class II region encompasses approximately 0.7 Mb and contains 34 loci, of which 16 are expressed genes, 3 are candidate or potential genes and 15 are pseudogenes (Shiina *et al.* 2004).

The class I region, which is located at the telomeric end of the MHC, contains 6 classical and a number of non-classical HLA genes. Classical class I molecules are found on the surface of most nucleated cells in the body while non-classical HLA molecules have a limited tissue distribution and are known to be less polymorphic that the classical genes. The class I region, covering approximately 1.8 Mb, contains 122 loci, of which 41 are expressed genes, 12 are candidate or potential genes, 3 are non-coding genes and 66 are pseudogenes (Shiina *et al.* 2004).

Finally, the class III region is located between the class I and II regions and is the most gene dense area of the MHC. It contains 62 loci within a region of approximately 0.9 Mb and 58 of these are expressed genes. This region does not encode any known HLA class I or class II-like genes, but it does contain a number of genes which play a role in immune function or inflammation. For example, a group of cytokine genes including *TNF, LTA* and *LTB* are present in this region of the genome.

### 5.1.2 HLA association with disease

As the MHC is central to the regulation of the immune and inflammatory responses, it is not surprising that this region has been associated with many immune and inflammatory disorders (Dawkins *et al.* 1999). Indeed, coeliac disease has been shown to be highly

associated with the high risk DQ2 heterodimer, which is encoded by the HLA alleles, DQA1*0501 and DQB1*0201 in the class II region of the MHC. Many other autoimmune diseases have been linked to HLA class I and HLA class II alleles. Among these are type-1 diabetes (HLA-DRB1), ulcerative colitis (MICA), rheumatoid arthritis (HLA-DRB1) and inflammatory bowel disease (HLA-DPA1) (Shiina *et al.* 2004). In these case studies, association with disease has been identified by a significant increase or decrease in allele frequency at the observed loci when compared between patient and control groups. However, it is possible that the association observed with these alleles is merely as a result of linkage disequilibrium between the tested marker and the actual disease susceptibility locus, given that the MHC is characterised by high levels of linkage disequilibrium. The association between HLA-DQ2 and coeliac disease is an exception since there is strong functional evidence, as previously discussed, to suggest that the HLA-DQ2 locus is the actual susceptibility locus in this case.

However, there has been speculation that there are additional susceptibility loci for CD, and indeed other autoimmune diseases, elsewhere in the MHC region. Most genome scans for autoimmune disease, including CD, have included the MHC within a region of linkage to disease. Furthermore, given that it is difficult to narrowly define the linked region using family based studies, it is also usually difficult or impossible to determine the actual number, or identity, of susceptibility genes in the linkage region using this approach. Since the MHC is such a gene dense region, with many genes encoding products with immunological functions, it is reasonable to assume that there may be other genes present which could be susceptibility factors for CD.

### 5.1.3 Additional MHC autoimmune susceptibility loci

Studies have been carried out to establish if there are other loci within the MHC which act as susceptibility loci to CD. To search for such loci, it is essential for the researcher to control for the primary disease susceptibility locus in the region (the HLA-DQ locus in coeliac disease). This can be achieved by ensuring case and control populations are matched at the disease locus. For coeliac disease studies, this requires, for example, that individuals in the study are all homozygous for the high risk DQ2 allele, or all individuals are heterozygous for the high risk allele. Selection of individuals based on DQ2-positivity alone is suboptimal when controlling for the disease locus, as this will result in differing proportions of DQ2 alleles between groups, as there will be a higher

proportion of DQ2 homozygous individuals in the patient group compared to the control group. This in turn would suggest that there are more DQ2 haplotypes in the patient population, possibly resulting in an artificial association between an allele on this haplotype and disease, which is however, simply the result of linkage disequilibrium to the primary susceptibility locus.

Studies have identified additional MHC loci which display significant association with disease status in CD. *MICA* is a molecule which serves as a ligand for TCR $V_{\delta}1$ $\gamma\delta$ T cells, numbers of which are unusually elevated in the CD lesion. The MICA-A5.1 microsatellite allele is interesting as it generates a frameshift mutation, in turn giving rise to a nonsense ('stop') mutation in exon 5. Exon 5 encodes the protein transmembrane region, making it likely that this allele is associated with modified protein function, or possibly, complete lack of function. Studies have investigated the association of this allele to CD using a number of different DQ2-controlled populations. One study (Gonzalez *et al.* 2004) found an association between MICA-A5.1 and disease in a DQ2 negative case-control study (i.e. in individuals who do not carry the DQA1*0501/DQB1*0201 combination). This study stated that a disease association existed at this locus which was independent of DQ2 and DQ8. However, they also found an association between some alleles of the *MICB* gene and disease, and because of the high levels of LD between the three loci, they were unable to conclusively pinpoint the disease locus. Another study, also found an association between MICA-A5.1 and disease but suggested that this allele was not the actual susceptibility allele, but instead acted as a marker for a susceptibility locus on the same haplotype (Fernandez *et al.* 2002). Another investigation showed that the MICA-A5.1 allele confers an additive effect to the DR3-DQ2 haplotype, that may modulate development of disease (Lopez-Vazquez *et al.* 2002). A family study (Louka *et al.* 2003) compared the frequency of a number of markers in the HLA region on those DR3-DQ2 haplotypes which were transmitted and not transmitted to CD affected offspring. Results exhibited that alleles at three different markers in the MIC-Class I region, including MICA-A5.1, had a significantly increased frequency on DR3-DQ2 transmitted haplotypes, suggesting the presence of an additional CD risk factor within this region. In addition to this, a later study carried out a family based approach to investigate any preferential transmission of MICA alleles to affected offspring (Rueda *et al.* 2003). This found an association between the MICA-A5.1 allele

and CD development which was independent of the DQ2-extended haplotype. However, no association was observed between MICA-A5.1 and disease in a collective European sample (Louka *et al.* 2003).

The *TNF* gene cluster in the MHC class III region has also been investigated for additional susceptibility loci. An association between the TNF2 allele (and the nearby TNFa2 microsatellite allele) and CD was first identified in an Irish population (McManus *et al.* 1996; McManus *et al.* 1996). A weak association with this allele was also seen in a Spanish population (Fernandez *et al.* 2002). Additional support for an association in this region came from a large study where 109 families were used to test for association using the homozygous parent transmission / disequilibrium test (HPTDT) and extended-TDT (ETDT) (Louka *et al.* 2003). Though no associations reaching statistical significance were identified, there was a tendency towards preferential transmission of six adjacent markers, and included in this region were the TNFd and D6S273 markers in the TNF region, supporting the earlier findings.

### 5.1.4 Linkage Disequilibrium patterns and haplotypes in the MHC

The problems associated with the inability to identify the exact location of disease susceptibility loci due to long range linkage disequilibrium between markers and extended haplotypes are intensified in the MHC region. This is due to a number of factors, namely, the high gene density in the region, the high number of polymorphisms in many of these genes, and the high extent of linkage disequilibrium and extended haplotypes in the region. Together, these factors create an environment where many polymorphic loci are in close proximity with high levels of LD, hence making it difficult to identify the actual disease susceptibility variant. Indeed, the MHC region is an extreme example of the presence of numerous conserved extended haplotypes extending over a distance of about 6 Mb from HLA-A to the D6S276 marker. One of the haplotypes spanning this area is termed Ancestral Haplotype 8.1 (AH 8.1), and is carried by about 10% of Northern Europeans (Allcock *et al.* 2004). This haplotype (which carries the variants HLA-A1, Cw7, B8, TNFAB*a2b3, TNFN*S, C2*C, Bf*s, C4A*Q0, C4B*1, DRB1*0301, DRB3*0101, DQA1*0501, DQB1*0201), has been shown to be associated with an unusually large number of autoimmune diseases, including CD, thus making it a high risk factor for disease (Allcock *et al.* 2004). Though this haplotype has

been identified, the location of the actual disease susceptibility locus or loci has not been determined for the majority of diseases linked to it. Neither has it been determined (as discussed above) whether the HLA-DQ locus is the only risk factor carried by this haplotype and suspicions remain that the haplotype may harbour additional risk modifying loci. As discussed, the high level of LD between all loci on this haplotype makes the use of positional markers (i.e. SNP markers chosen for analysis based solely on their location, without consideration of additional factors, such as possible functional effects) to identify significant associations with disease difficult. Thus a more informed approach needs to be undertaken to identify the actual susceptibility locus or loci.

### 5.1.5 Rational of this study

Recent efforts by the MHC Haplotype Project (Allcock *et al.* 2004; Stewart *et al.* 2004) to sequence and characterise extended haplotypes present in the MHC region have greatly improved the information available to dissect and identify actual disease susceptibility loci in the MHC. Such studies provide a fully informative polymorphic map of these haplotypes and hence, allow the selection of polymorphisms which may be the actual susceptibility variant based not only on their location but also on their potential functional effect. Until now, no such map of the variation in the region was available. Stewart *et al.* (2004) reported on the sequencing of two consanguineous, HLA-homozygous cell lines for each of two common haplotypes in the MHC; the AH 8.1 haplotype as previously discussed, and the AH 7.1 haplotype, also called the B7, DR2 haplotype. This also has a frequency of approximately 10% in Northern Europeans and has been shown to provide very significant protection against type 1 diabetes (Stewart *et al.* 2004). Both haplotypes were sequenced in their entirety using DNA cloned into bacterial artificial chromosomes (BACs). The variation content of the two haplotypes was then described, revealing more than 18,000 variations between them. In this study, we are interested in variation in the MHC class III region which may be associated with disease. Of particular interest are polymorphic loci which create amino acid substitutions in genes with an established or potential immune function and which could implicate them as candidate genes for analysis. Also of interest were SNPs which may alter potential transcription factor binding sites. The ability to comprehensively identify such variants from these sequencing studies represents a valuable step forward in disease susceptibility analysis.

The TNF region of the MHC has been characterised in even greater detail, therefore it has been possible to use both haplotype tagging information, as used in previous chapters, in combination with very comprehensive information about the polymorphic nature of the area, to conduct a detailed analysis of the TNF region. This region is of considerable interest in view of previous studies which have shown it may harbour an additional susceptibility locus for CD, as has been discussed above. A recent study analysed 19 SNPs (consisting mainly of coding and promoter SNPs), within the TNF block, which encompasses the *BAT1*, *ATP6V1G3*, *IkBL* (*NFKBIL1*), *LTA* and *TNF* genes (Allcock *et al.* 2004). All of the SNPs may have potential functional effects and hence are good candidates for inclusion in the study of CD susceptibility. In fact, previous studies have provided evidence that a number of SNPs in this region of the genome, namely LTA+252 and IkBL-62 (both of which are included for analysis in this study) can alter gene expression (Ozaki *et al.* 2002). As the haplotype structure of the region is now known from the work of Allcock *et al.* (ten haplotypes with a frequency greater than 4% were identified), htSNPs may be used to infer the allele frequencies at all SNPs within these haplotypes.

In this present study, the information obtained from the SNP data in the Allcock *et al.* (2004) study will be used to examine the TNF block of the MHC class III region, while the information obtained from the SNP data in the Stewart *et al.* (2005) study will use a separate SNP dataset to examine a separate region of the MHC class III region, between the genes from *BAT2* to *DDAH*.

**5.2 Objectives**

The aims of the work outlined in this chapter were to:

1. Identify the HLA-DR and HLA-DQ genotype for a subset of coeliac patient and control individuals.
2. Identify tagging SNPs to tag all variation within the 19-SNP haplotype in the TNF region as identified by Allcock *et al.* (2004).
3. Identify the polymorphic variation between the AH 8.1 and AH 7.1 haplotypes as outlined by Stewart *et al.* (2004), which results in

a. non-synonymous amino acid changes in selected genes in the MHC class III region.
b. possible alteration of transcription factor binding sites in the promoter regions of selected genes.

4. Test for statistically significant allelic, genotypic or carrier status associations between genotyped polymorphisms as identified above and coeliac disease in a DQ2 controlled population.
5. Establish the extended haplotypes present for each HLA-DQB1*0201 heterozygous and HLA-DQB1*0201 homozygous individual in the patient group and for each HLA-DQB1*0201 heterozygous individual in the control group.
6. Identify and compare the allele frequencies on the high risk haplotypes (DQB1*0201 haplotypes) in DQ2 heterozygous patients and controls to establish if the high risk haplotype carries any additional SNPs in the MHC class III region at statistically different frequencies between the groups. Such a finding may be indicative of an additional disease effect on this haplotype in the class III region which is independent of the DQB1*0201 association.
7. Identify and compare allele frequencies on the high risk haplotypes (DQB1*0201 haplotypes) in DQ2 homozygous and DQ2 heterozygous patients. This is to establish if there are any statistical differences in allele frequencies at the additional MHC class III SNPs which may explain why only one copy of the high risk allele is present in some patients who develop CD.
8. Similarly, to compare the haplotype frequencies of each high risk (HLA-DQB1*0201-carrying) haplotype in DQ2 heterozygous versus DQ2 homozygous patients.
9. Compare the frequencies of alleles at the DQB1 locus in DQ2 heterozygous patients and controls.
10. Examine and compare the low risk (non DQB1*0201) haplotypes in DQ2 heterozygous patients and controls.

## 5.3 Results

### 5.3.1 HLA-DRB1, HLA-DQB1 typing

281 coeliac patients and 305 healthy control individuals were genotyped to establish their HLA genotypes. 252 control individuals have been genotyped at the HLA-DRB1, HLA-DQB1 (and HLA-DPB1) loci by the Irish Blood Transfusion Service (IBTS), St. James's Hospital, Dublin. Individuals who have one copy of the DRB1*0301, DQB1*0201 (DR3-DQ2) haplotype were classified as having one copy of the high disease risk allele, DQ2. Individuals with two copies of the DRB1*0301, DQB1*0201 haplotype were classified as "DQ2 homozygous individuals" and individuals with no copies of the DRB1*0301, DQB1*0201 haplotype, were classified as "DQ2 negative individuals". The DRB1*0701 DQB1*0202 allelic combination is known to increase risk of CD when accompanied by the DRB1*0301, DQB1*0201 allelic combination on the opposing chromosome. Hence, to clearly assess for the presence of any additional risk factor or factors on the DRB1*0301, DQB1*0201 haplotype without the added complexity of the risk imposed by the DRB1*0701 DQB1*0202 haplotype, individuals were only included for analysis in the study as "DQ2 heterozygous individuals" if they possessed one copy of the DRB1*0301, DQB1*0201 haplotype and a low risk haplotype (i.e. any DQB1 allele except DQB1*0201 or DQB1*0202).

194 patients were genotyped by the Irish Blood Transfusion Service, St. James's Hospital, Dublin, at the HLA-DRB1 and HLA-DQB1 loci. These individuals were genotyped at a resolution to identify whether samples were, for example, DQB1*02. In this case, a DQB1*02 typing represents the allelic string 0201 or 0202. However, the haplotype structure of the region, as identified in past research and in the previous 252 control individuals (who were typed at high resolution at both sites) shows that the DQB1*02 allele found in combination with a DRB1*03 typing is always a DQB1*0201 allele, giving the high risk DR3-DQ2 combination (R. Hagan (IBTS), pers. comm.). Hence, individuals with the typing DRB1*03, DQB1*02 were identified as carrying the high risk allele. Individuals who had one copy of the DRB1*03, DQB1*02 haplotype were classified as "DQ2 heterozygous individuals" (again this indicates the presence of the high risk DQB1*0201 allele and any other allele on the other chromosome except DQB1*0201 or DQB1*0202). Individuals with two copies of the DRB1*03, DQB1*02 haplotype were classified as "DQ2 homozygous individuals", and individuals with no

copies of the DRB1*03, DQB1*02 haplotype, were classified as "DQ2 negative individuals".

87 patients and 53 controls were genotyped in an earlier HLA genotyping study at the DQB1 locus only at high resolution genotyping which could distinguish whether the allele was, for example, DQB1*0201 or DQB1*0202. Again, the haplotype structure of the region dictates that a DQB1*0201 allele indicates linkage to the DRB1*0301 allele, i.e. the DQ2 high risk allele. For this group, individuals with one copy of the DQB1*0201 allele and any other allele except DQB1*0201 or DQB1*0202 were classified as "DQ2 heterozygous individuals". Individuals with two copies of the DQB1*0201 allele were classified as "DQ2 homozygous individuals", and individuals with no copies of the DQB1*0201 allele, were classified as "DQ2 negative individuals".

It should be noted that for the purpose of this study, only individuals who express the high risk DQ2 molecule *in cis* (DR3-DQ2 haplotypes) are included in the analysis. This approach was taken since our principle objective was to compare haplotypes in order to discern the presence of additional risk modifying loci on the high risk haplotype. Individuals who express the DQ2 high risk molecule *in trans* (DR5-DQ7 and DR7-DQ2 heterozygote individuals) have different haplotype combinations and were not at sufficiently high frequency (10 patients and 12 controls) to be informative in this analysis.

Table 5.1 lists the number of DQ2 homozygous, DQ2 heterozygous and DQ2 negative individuals identified in the patient and control population.

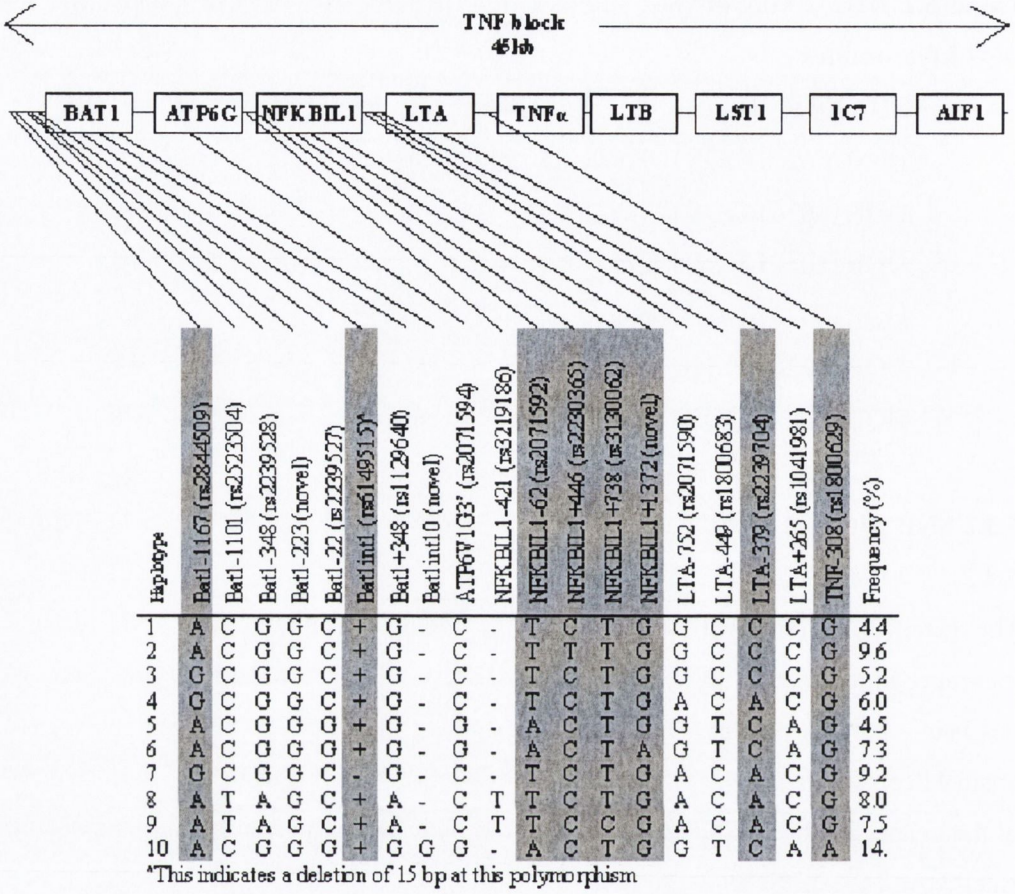
**Table 5.1 DQ2 status of individuals genotyped for the presence of the high risk DQ2 heterodimer**

| HLA-DQB1 genotype | Classification | Patients | Controls |
|---|---|---|---|
| DQB1*0201/DQB1*0201 | DQ2 homozygous | 75 | 6 |
| DQB1*0202/DQB1*0202 | | 3 | 5 |
| DQB1*0201/DQB1*0202 | | 87 | 10 |
| DQB1*0201/DQB1*other | DQ2 heterozygous | 89 | 83 |
| DQB1*0202/DBQ1*other | | 14 | 34 |
| DQB1*other/DQB1*other | DQ2 negative | 13 | 167 |

### 5.3.2 SNP selection results

#### *5.3.2.1 TNF block*

The genetic diversity within the haplotype block surrounding the TNF region, and incorporating the genes, *TNF*, *LTA*, *LTB*, *AIF1*, *NCR3*, *NFKBIL1*, *ATP6V1G* and *BAT1*, has been characterised by a previous study using high density SNP genotyping of the area (Allcock *et al.* 2004). This study identified 19 SNPs in coding and promoter regions of the genes, and these SNPs are listed within their haplotype combinations in figure 5.1. Haplotype-tagging SNPs were identified to tag all SNPs listed in the region using the SNPtagger program, as described previously (section 2.7.2). Identified htSNPs are shaded in grey in figure 5.1. These htSNPs can tag all haplotypes with a predicted frequency greater than 4%.




| Haplotype | Bat1-1167 (rs2844509) | Bat1-1101 (rs2523504) | Bat1-348 (rs2239528) | Bat1-223 (novel) | Bat1-22 (rs2239527) | Bat1int1 (rs61495157)* | Bat1+348 (rs1129640) | Bat1int10 (novel) | ATP6V1G3' (rs2071594) | NFKBIL1-421 (rs3219186) | NFKBIL1-62 (rs2071592) | NFKBIL1+446 (rs2230365) | NFKBIL1+738 (rs3130062) | NFKBIL1+1372 (novel) | LTA-752 (rs2071590) | LTA-449 (rs1800683) | LTA-379 (rs2239704) | LTA+265 (rs1041981) | TNF-308 (rs1800629) | Frequency (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | A | C | G | G | C | + | G | - | C | - | T | C | T | G | G | C | C | C | G | 44 |
| 2 | A | C | G | G | C | + | G | - | C | - | T | T | T | G | G | C | C | C | G | 96 |
| 3 | G | C | G | G | C | + | G | - | C | - | T | C | T | G | G | C | C | C | G | 52 |
| 4 | G | C | G | G | C | + | G | - | C | - | T | C | T | G | A | C | A | C | G | 60 |
| 5 | A | C | G | G | G | + | G | - | G | - | A | C | T | G | G | T | C | A | G | 45 |
| 6 | A | C | G | G | G | + | G | - | G | - | A | C | T | A | G | T | C | A | G | 73 |
| 7 | G | C | G | G | C | - | G | - | C | - | T | C | T | G | A | C | A | C | G | 92 |
| 8 | A | T | A | G | C | + | A | - | C | T | T | C | T | G | A | C | A | C | G | 80 |
| 9 | A | T | A | G | C | + | A | - | C | T | T | C | C | G | A | C | A | C | G | 75 |
| 10 | A | C | G | G | G | + | G | G | G | - | A | C | T | G | G | T | C | A | A | 14. |

*This indicates a deletion of 15 bp at this polymorphism

**Figure 5.1 TNF block haplotype tagging SNP selection. HtSNPs are shaded in grey, and describe all haplotypes with a frequency >4%. Haplotype structure and the relative positions of the htSNPs within the genes of the TNF block are shown in the above diagram and table.**

### *5.3.2.2 SNP variation between the AH 8.1 and AH 7.1 haplotypes*

The polymorphic variation identified by Stewart *et al.* (2004), which exists between the two extended ancestral haplotypes in the MHC region (AH 7.1 & 8.1), was assessed between the class III genes *NFKBIL1* (telomeric) to *EGFL8* (centromeric). Non-synonymous SNPs differing between the haplotypes were identified from the Stewart data set as a part of the present study (described in section 2.8) and these are listed in table 5.2. Figure 2.2 lists the genes present in this region. Seven SNPs located in a subset of this region between (and including) the *BAT2* and *DDAH* genes are included for association testing in this study.

**Table 5.2 Non-synonymous SNPs identified by Stewart *et al.* to differ between the MHC extended haplotypes AH 8.1 and AH 7.1 (Genes and SNPs are listed in order of their location, from telomeric to centromeric position).**

| Gene | SNP (rs number) | Contig position (NT_007592) | Distance between SNPs | Genotyped in this study |
|---|---|---|---|---|
| NFKBIL1 | rs3130062 | 22384163 | 14644 | |
| LTA | rs2857713 | 22398807 | 59366 | |
| BAT2 | rs1046080 | 22458173 | 1422 | Genotyped |
| BAT2 | rs2272593 | 22459595 | 1623 | Genotyped |
| BAT2 | rs1046089 | 22461218 | 1077 | Genotyped |
| BAT2 | rs3132453 | 22462295 | 547 | Genotyped |
| BAT2 | rs10885 | 22462842 | 27543 | Genotyped |
| BAT4 | rs3130618 | 22490385 | 7845 | Genotyped |
| LY6G5B | rs9267532 | 22498230 | 138293 | Genotyped |
| DDAH | | | | |
| HSPA1L | rs2227956 | 22636523 | 5591 | |
| HSPA1A | rs562047 | 22642114 | 134601 | |
| BF | rs4151659 | 22776715 | 10550 | |
| SKIV2L | rs437179 | 22787265 | 19434 | |
| STK19 | rs616634 | 22806699 | 46690 | |
| C4B | rs429329 | 22853389 | 1158 | |
| C4B | rs2258218 | 22854547 | 301 | |
| C4B | rs2734319 | 22854848 | 3 | |
| C4B | rs2734318 | 22854851 | 8 | |
| C4B | rs2229399 | 22854859 | 3 | |
| C4B | rs2229400 | 22854862 | 354 | |
| C4B | rs17874654 | 22855216 | 93 | |
| C4B | rs2734316 | 22855309 | 9 | |
| C4B | rs2746413 | 22855318 | 455 | |
| C4B | rs17874655 | 22855773 | 19640 | |
| TNXB | rs9267795 | 22875413 | 8944 | |
| TNXB | rs1009382 | 22884357 | 20837 | |
| TNXB | rs9267799 | 22905194 | 17782 | |
| TNXB | rs1150752 | 22922976 | 57660 | |
| PPT2 | rs3134604 | 22980636 | 86 | |
| PPT2 | rs3096696 | 22980722 | 12038 | |
| EGFL8 | rs3096697 | 22992760 | | |

We identified SNPs that may alter potential regulatory sites in promoters within this region and which differ between the ancestral haplotypes AH 7.1 and AH 8.1. This strategy involved the identification of all SNPs within 1,000 bp of the transcription start site which occur in, and alter the consensus sequence of transcription factor binding sites. Such SNPs may potentially alter transcription factor binding and consequently, the activity of the gene. Table 5.3 lists all SNPs which were found to occur within 1,000 bp of a transcription start site of a gene in the MHC class III region. 7 SNPs occur within

1,000 bp of the transcription start site of genes located in the interval delimited by *BAT2* and *DDAH*. Two of these (BAT2rs3130068 and CSNK2Brs805257) were identified (using TESS software – see section 2.7.3) as altering consensus transcription factor binding sites within the promoter regions of *BAT2* and *CSNK2B*, and these SNPs were included in this study.

**Table 5.3 SNPs which alter potential transcription factor binding sites, which differ between the MHC extended haplotypes AH 8.1 and AH 7.1, identified in this study from the data of Stewart *et al.***

| Gene | rs number | contig position (NT_007592) | TF binding alteration? | Genotyped in this study |
|---|---|---|---|---|
| ATP6V1G2 | rs2071594 | 22370971 | c-Myc | |
| LTA | rs2239704 | 22398392 | AP-2 gamma, GCF | |
| LTA | rs909253 | 22398564 | No TF sites | |
| LTA | rs746868 | 22398680 | No TF sites | |
| BAT2 | rs3115665 | 22447515 | No TF change | |
| BAT2 | rs3130068 | 22448605 | Arnt, AhR | genotyped |
| BAT4 | rs7029 | 22488204 | No TF sites | |
| CSNK2B | rs805257 | 22492444 | c-Myc | genotyped |
| CSNK2B | rs9267529 | 22492821 | No TF sites | |
| LY6G5C | rs13295 | 22502891 | No TF sites | |
| LY6G6C | rs15574 | 22544748 | No TF sites | |
| HSPA1A | rs1043618 | 22641758 | NF-ATp | |
| NEU1 | rs3130678 | 22685691 | No TF sites | |
| TNXB | rs1058152 | 22867214 | No TF sites | |
| AGPAT1 | rs1061808 | 22994797 | IL6 RE-BP | |
| AGPAT1 | rs1061807 | 22995088 | NF-GMb | |
| PBX2 | rs1004095 | 23011659 | No TF sites | |

As previously discussed, some studies have speculated that the *MIC* genes (*MICA* and *MICB*) may be susceptibility genes in CD pathogenesis, hence, these genes have also been included in our study. The *MICA* and *MICB* genes were investigated for the presence of any SNPs identified by Stewart *et al.* (2004) which differ between haplotypes AH 8.1 and AH 7.1. Both non-synonymous SNPs and SNPs located within 1,000 bp of the transcription site of both genes were investigated. This study did not identify any SNPs in the *MICA* gene which fulfil the criteria but did identify five non-synonymous *MICB* SNPs which meet these conditions (listed in table 5.4).

**Table 5.4 Non-synonymous SNPs in the *MICA* and *MICB* genes identified in this study from the data of Stewart *et al.* to differ between the MHC extended haplotypes AH 8.1 and AH 7.1**

| Gene | rs number | Genotyped in this study |
|---|---|---|
| MICB | **rs3131639** | |
| MICB | **rs1065075** | genotyped |
| MICB | **rs3134900** | genotyped |
| MICB | **rs1051788** | |
| MICB | **rs1065076** | genotyped |

A number of additional SNPs in the MHC class III region had been previously selected and genotyped by the laboratory group. These SNPs were CSNK2B1942, CSNK2B3854, DDAH_rs805305, LTA+252, TNFa -863 and TNFa -857 and these SNPs were also included in this study.

### 5.3.3 Hardy-Weinberg Equilibrium (HWE) Testing

To test for HWE in this dataset, only genotyping results from the control population were assessed, as the coeliac disease population is known to be enriched for particular alleles at the HLA-DQ locus which is in the same chromosomal region as the SNPs under investigation in this study. Therefore it is likely that SNP alleles in LD with the disease associated HLA-DQ locus may be over-selected in patients, hence, this SNP may not be in HWE, even if the genotyping assay is valid.

Results of HWE analysis for the genotyping assays in the control population were carried out using GENEPOP and are listed in table 5.5. The results show that all SNPs are in Hardy-Weinberg equilibrium, which minimises the probability of error in the genotyping assays.

**Table 5.5 Hardy Weinberg Equilibrium analysis results for all candidate SNPs in the MHC class III region and *MICB***

| SNP | Control group p-value |
|---|---|
| BAT1-1167 | 0.7598 |
| IkBL-62 | 0.3620 |
| IkBL+446 | 0.7901 |
| IkBL+738 | 1.0000 |
| LTA-379 | 0.3556 |
| LTA+252 | 0.5656 |
| LTA-863 | 0.3975 |
| TNF-857 | 0.6661 |
| TNF-308 | 0.1376 |
| BAT2rs3130068 | 1.0000 |
| BAT2rs1046080 | 0.6055 |
| BAT2rs2272593 | 0.6930 |
| BAT2rs1046089 | 0.6451 |
| BAT2rs3132453 | 0.7533 |
| BAT2rs10885 | 0.7458 |
| BAT4rs3130618 | 0.2689 |
| CSNK2Brs805257 | 1.0000 |
| CSNK2B1942 | 1.0000 |
| CSNK2B3854 | 0.7405 |
| LY6G5Brs9267532 | 1.0000 |
| DDAH2 | 0.1211 |
| MICBrs3134900 | 0.2005 |
| MICB-7595(rs1065075) | 0.0700 |
| MICB-11715(rs1065076) | 0.1594 |
| DQB1 | 0.4275 |

### 5.3.4 Allele, genotype and carrier status frequencies and disease association testing in genotyped polymorphisms in MHC class III in DQ2 heterozygous individuals

To test for association to disease, allele, genotype and carrier status frequencies were compared in a DQ2-controlled (DQ2 heterozygous) population. Results for the association testing of the MHC class III SNPs in the DQ2-controlled population are listed in tables 5.6 (genotype and allele associations) and 5.7 (carrier status associations). The p-values and odds ratios for the allele frequency differentials between patient and control groups are displayed graphically in figure 5.2. Bat1-1167 was found to be significantly associated with disease at the 5% level of association when allele (p-value = 0.0423), genotype (p-value = 0.0284) and carrier status association (G allele carrier

status: patients 15.5%, controls 31.6%, p-value = 0.0158) were examined. IkBL+446 was also found to be significantly associated with disease at the 5% level for allele frequency (p-value = 0.0260) and carrier status (T allele: patients 20.5%, controls 8.6%, p-value = 0.0320). Also associated with disease was SNP BAT4rs3130618, which was shown to be significantly associated with disease at the genotype (p-value = 0.0033) and carrier status level (A allele carrier status: patients 100.0%, controls 88.9%, p-value = 0.0013). The BAT2rs1046080 and BAT2rs2272593 polymorphisms were also significantly associated with disease when carrier status was examined (C allele carrier status and T allele carrier status, respectively). Both have identical carrier status frequencies (patients 54.1%, controls 37.0%, p-value = 0.0273). LY6G5Brs9267532 also displayed nominally significant association with disease at the allele (p-value = 0.0464) and carrier status level (patients 4.6%, controls 13.6%, p-value = 0.0413).

**Figure 5.2 P-values and odds ratios for allele frequency differences between patient and control groups for SNP loci in the MHC class III region and MIC genes**

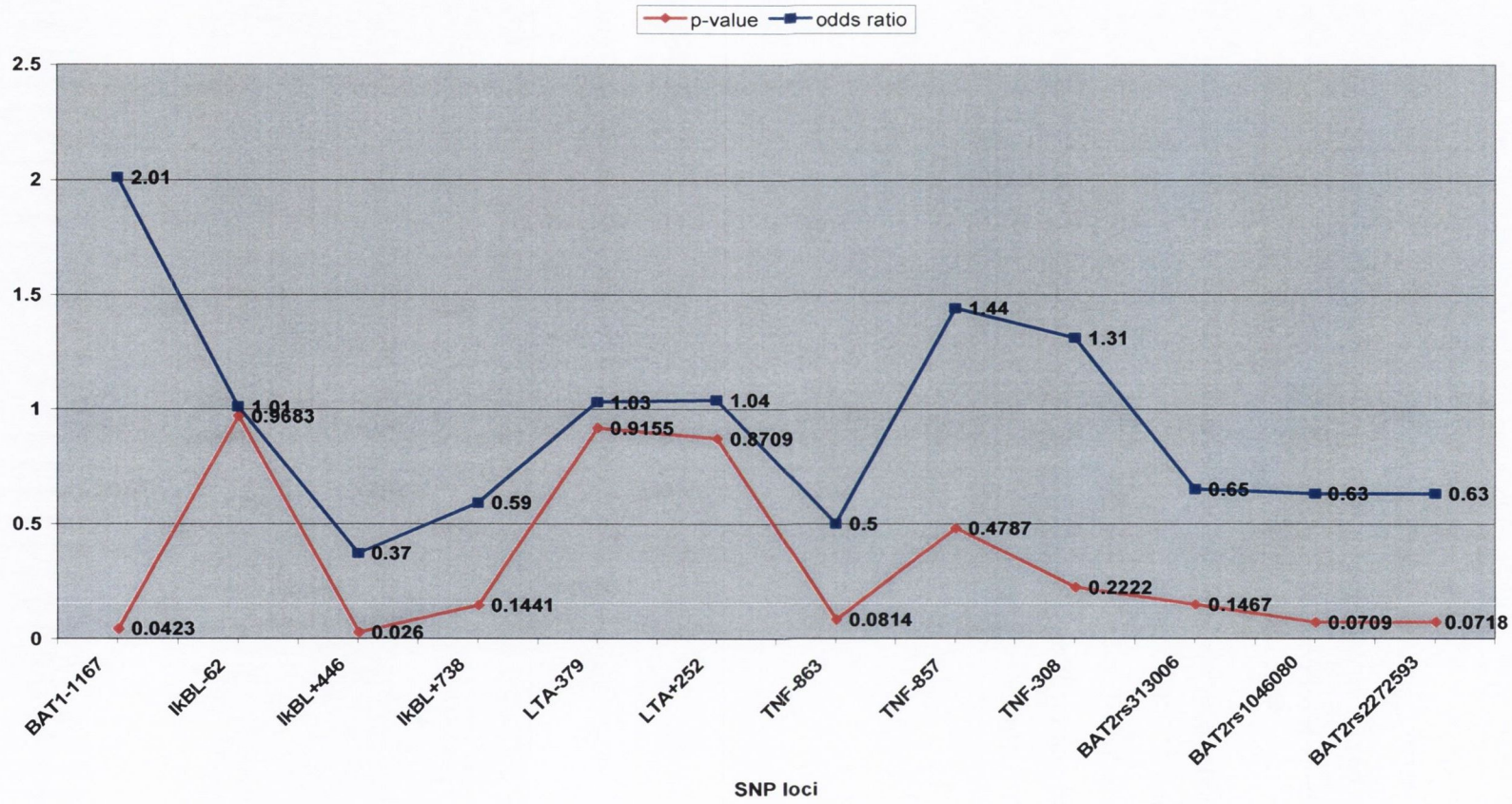
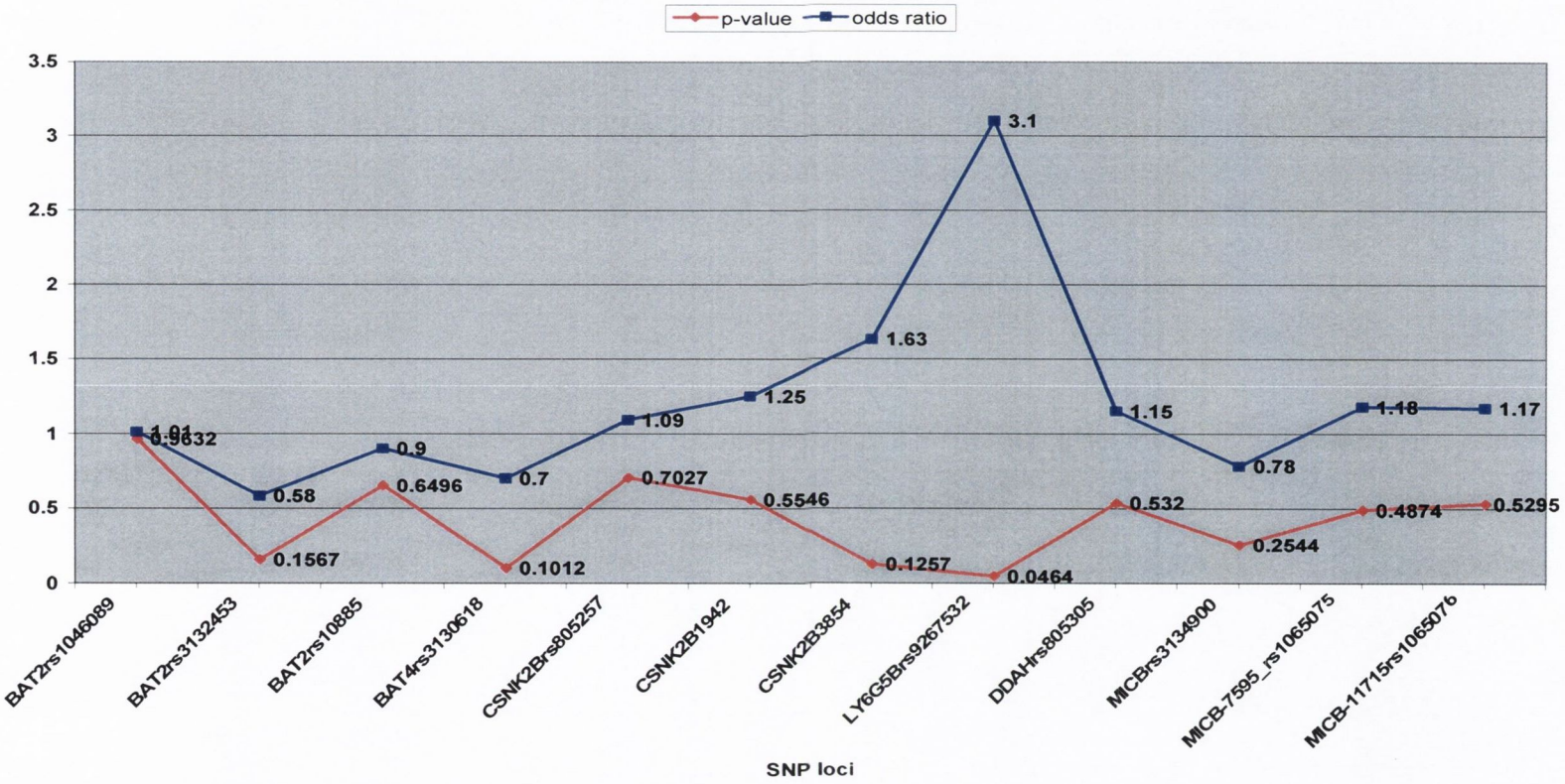


**Figure 5.2 P-values and odds ratios for allele frequency differences between patient and control groups for SNP loci in the MHC class III region and MIC genes (continued)**




Actual allele frequencies listed in table 5.6

**Table 5.6 Genotype and allele frequencies for candidate polymorphisms in the MHC class III region and *MICB* in the DQ2 heterozygous coeliac disease patient and control populations**

| | | Genotype n (%) | | | | Allele n (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SNP | | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | p | $A_1$ | $A_2$ | p | OR | 95%CI |
| BAT1 | Pat | 71 (84.5) | 11 (13.1) | 2 (2.4) | **0.0284** | 153 (91.1) | 15 (8.9) | **0.0423** | 2.01 | (0.97,4.20) |
| -1167A/G | Con | 52 (68.4) | 23 (30.3) | 1 (1.3) | | 127 (83.6) | 25 (16.4) | | | |
| IkBL | Pat | 16 (19.0) | 63 (75.0) | 5 (6.0) | 0.2848 | 95 (56.5) | 73 (43.5) | 0.9683 | 1.01 | (0.64,1.60) |
| -62T/A | Con | 19 (24.1) | 51 (64.6) | 9 (11.4) | | 89 (56.3) | 69 (43.7) | | | |
| IkBL | Pat | 66 (79.6) | 16 (19.3) | 1 (1.2) | 0.0839 | 148 (89.0) | 18 (10.8) | **0.0260** | 0.37 | (0.14,0.98) |
| +446C/T | Con | 74 (91.4) | 7 (8.6) | 0 (0.0) | | 155 (95.7) | 7 (4.3) | | | |
| IkBL | Pat | 64 (73.6) | 23 (26.4) | 0 (0.0) | N/A | 151 (86.8) | 23 (13.2) | 0.1441 | 0.59 | (0.27,1.27) |
| +738T/C | Con | 66 (83.5) | 13 (16.5) | 0 (0.0) | | 145 (91.8) | 13 (8.2) | | | |
| LTA | Pat | 40 (46.5) | 45 (52.3) | 1 (1.2) | 0.7992 | 125 (72.7) | 47 (27.3) | 0.9155 | 1.03 | (0.62,1.71) |
| -379C/A | Con | 37 (46.8) | 40 (50.6) | 2 (2.5) | | 114 (72.2) | 44 (27.8) | | | |
| LTA | Pat | 17 (19.8) | 64 (74.4) | 5 (5.8) | 0.2168 | 98 (57.0) | 74 (43.0) | 0.8709 | 1.04 | (0.66,1.63) |
| +252G/A | Con | 20 (24.4) | 52 (63.4) | 10 (12.2) | | 92 (56.1) | 72 (43.9) | | | |
| TNF | Pat | 66 (76.7) | 20 (23.3) | 0 (0.0) | N/A | 152 (88.4) | 20 (11.6) | 0.0814 | 0.50 | (0.21,1.17) |
| -863C/A | Con | 71 (87.7) | 10 (12.3) | 0 (0.0) | | 152 (93.8) | 10 (6.1) | | | |

| SNP | | Genotype n (%) $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | p | Allele n (%) $A_1$ | $A_2$ | p | OR | 95%CI |
|---|---|---|---|---|---|---|---|---|---|---|
| TNF | Pat | 81 (92.0) | 7 (8.0) | 0 (0.0) | N/A | 169 (96.0) | 7 (4.0) | 0.4787 | 1.44 | (0.48,4.41) |
| -857C/T | Con | 71 (88.8) | 9 (11.3) | 0 (0.0) | | 151 (94.4) | 9 (5.6) | | | |
| TNF | Pat | 11 (12.6) | 67 (77.0) | 9 (10.3) | 0.1437 | 89 (51.1) | 85 (48.9) | 0.2222 | 1.31 | (0.83,2.05) |
| -308A/G | Con | 4 (4.9) | 65 (79.3) | 13 (15.9) | | 73 (44.5) | 91 (55.5) | | | |
| BAT2 | Pat | 51 (60.7) | 32 (38.1) | 1 (1.2) | 0.2452 | 134 (79.8) | 34 (20.2) | 0.1467 | 0.65 | (0.35,1.21) |
| rs3130068A/G | Con | 59 (72.8) | 21 (25.9) | 1 (1.2) | | 139 (85.8) | 23 (14.2) | | | |
| BAT2 | Pat | 39 (45.9) | 44 (51.8) | 2 (2.4) | 0.0796 | 122 (71.8) | 48 (28.2) | 0.0709 | 0.63 | (0.36,1.08) |
| rs1046080A/C | Con | 51 (63.0) | 28 (34.6) | 2 (2.5) | | 130 (80.2) | 32 (19.8) | | | |
| BAT2 | Pat | 40 (46.0) | 45 (51.7) | 2 (2.3) | 0.0788 | 125 (71.8) | 49 (28.2) | 0.0718 | 0.63 | (0.36,1.08) |
| rs2272593C/T | Con | 51 (63.0) | 28 (34.6) | 2 (2.5) | | 130 (80.2) | 32 (19.8) | | | |
| BAT2 | Pat | 5 (6.1) | 61 (74.4) | 16 (19.5) | 0.9902 | 71 (43.3) | 93 (56.7) | 0.9632 | 1.01 | (0.63,1.61) |
| rs1046089G/A | Con | 5 (6.3) | 58 (73.4) | 16 (20.3) | | 68 (43.0) | 90 (57.0) | | | |
| BAT2 | Pat | 63 (75.9) | 20 (24.1) | 0 (0.0) | N/A | 146 (88.0) | 20 (12.0) | 0.1567 | 0.58 | (0.26,1.31) |
| rs3132453G/T | Con | 69 (85.2) | 12 (14.8) | 0 (0.0) | | 150 (92.6) | 12 (7.4) | | | |
| BAT2 | Pat | 5 (6.1) | 72 (87.8) | 5 (6.1) | 0.4855 | 82 (50.0) | 82 (50.0) | 0.6496 | 0.90 | (0.57,1.43) |
| rs10885C/T | Con | 9 (11.4) | 65 (82.3) | 5 (6.3) | | 83 (52.5) | 75 (47.5) | | | |

| SNP | | Genotype n (%) $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | p | Allele n (%) $A_1$ | $A_2$ | p | OR | 95%CI |
|---|---|---|---|---|---|---|---|---|---|---|
| BAT4 | Pat | 0 (0.0) | 73 (86.9) | 11 (13.1) | **0.0033** | 73 (43.5) | 95 (56.5) | 0.1012 | 0.70 | (0.44,1.10) |
| rs3130618C/A | Con | 9 (11.1) | 67 (82.7) | 5 (6.2) | | 85 (52.5) | 77 (47.5) | | | |
| CSNK2B | Pat | 5 (6.3) | 58 (72.5) | 17 (21.3) | 0.8708 | 68 (42.5) | 92 (57.5) | 0.7027 | 1.09 | (0.68,1.75) |
| rs805257A/G | Con | 4 (5.1) | 55 (70.5) | 19 (24.4) | | 63 (40.4) | 93 (59.6) | | | |
| CSNK2B | Pat | 73 (84.9) | 12 (14.0) | 1 (1.2) | 0.8182 | 158 (91.9) | 14 (8.1) | 0.5546 | 1.25 | (0.56,2.83) |
| 1942A/G | Con | 65 (81.3) | 14 (17.5) | 1 (1.3) | | 144 (90.0) | 16 (10.0) | | | |
| CSNK2B | Pat | 70 (80.5) | 15 (17.2) | 2 (2.3) | 0.2199 | 155 (89.1) | 19 (10.9) | 0.1257 | 1.63 | (0.83,3.21) |
| 3854G/A | Con | 56 (68.3) | 23 (28.0) | 2 (2.4) | | 135 (83.3) | 27 (16.7) | | | |
| LY6G5B | Pat | 83 (95.4) | 4 (4.6) | 0 (0.0) | N/A | 170 (97.7) | 4 (2.3) | **0.0464** | 3.10 | (0.89,11.80) |
| rs9267532C/T | Con | 70 (86.4) | 11 (13.6) | 0 (0.0) | | 151 (93.2) | 11 (6.8) | | | |
| DDAH | Pat | 16 (18.8) | 65 (76.5) | 4 (4.7) | 0.6295 | 97 (57.1) | 73 (42.9) | 0.5320 | 1.15 | (0.73,1.81) |
| rs805305C/G | Con | 12 (14.6) | 64 (78.0) | 6 (7.3) | | 88 (53.7) | 76 (46.3) | | | |
| MICB | Pat | 11 (13.3) | 66 (79.5) | 6 (7.2) | 0.1052 | 88 (53.0) | 78 (47.0) | 0.2544 | 0.78 | (0.49,1.23) |
| rs3134900C/G | Con | 16 (19.8) | 64 (79.0) | 1 (1.2) | | 96 (59.3) | 66 (40.7) | | | |
| MICB-7595 | Pat | 12 (16.0) | 55 (73.3) | 8 (10.7) | 0.5752 | 79 (52.7) | 71 (47.3) | 0.4874 | 1.18 | (0.73,1.90) |
| rs1065075G/A | Con | 10 (13.7) | 51 (69.9) | 12 (16.4) | | 71 (48.6) | 75 (51.4) | | | |
| MICB-11715 | Pat | 48 (55.8) | 37 (43.0) | 1 (1.2) | 0.7511 | 133 (77.3) | 39 (22.7) | 0.5295 | 1.17 | (0.69,2.00) |
| rs1065076G/A | Con | 41 (50.0) | 40 (48.8) | 1 (1.2) | | 122 (74.4) | 42 (25.6) | | | |

**Table 5.7 Carrier status frequencies for candidate polymorphisms in the MHC class III region and *MICB* in the DQ2 heterozygous coeliac disease patient and control populations**

| SNP | | Allele$A_1$CarrierStatus $A_1A_1/A_1A$ | $A_2A_2$ | p | OR[(a)] | 95%CI | Allele$A_2$CarrierStatus $A_1A_1$ | $A_1A_2/A_2A_2$ | p | OR[(b)] | 95%CI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BAT2 -1167 | Pat | 82 (97.6) | 2 (2.4) | 1.0000* | 0.55 | (0.01,10.74) | 71 (84.5) | 13 (15.5) | **0.0158** | 0.40 | (0.17,0.91) |
| | Con | 75 (98.7) | 1 (1.3) | | | | 52 (68.4) | 24 (31.6) | | | |
| IkBL -62 | Pat | 79 (94.0) | 5 (6.0) | 0.2154 | 2.03 | (0.59,7.37) | 16 (19.0) | 68 (81.0) | 0.4369 | 1.35 | (0.60,3.04) |
| | Con | 70 (88.6) | 9 (11.4) | | | | 19 (24.1) | 60 (75.9) | | | |
| IkBL +446 | Pat | 82 (98.8) | 1 (1.2) | 1.0000* | 0.00 | (0.00,39.96) | 66 (79.5) | 17 (20.5) | **0.0320** | 2.72 | (0.98,7.78) |
| | Con | 81 (100.0) | 0 (0.0) | | | | 74 (91.4) | 7 (8.6) | | | |
| IkBL +738 | Pat | 87 (100.0) | 0 (0.0) | N/A | N/A | N/A | 64 (73.6) | 23 (26.4) | 0.1191 | 1.82 | (0.80,4.19) |
| | Con | 79 (100.0) | 0 (0.0) | | | | 66 (83.5) | 13 (16.5) | | | |
| LTA -379 | Pat | 85 (98.8) | 1 (1.2) | 0.6072* | 2.21 | (0.11,131.76) | 40 (46.5) | 46 (53.5) | 0.9668 | 1.01 | (0.52,1.96) |
| | Con | 77 (97.5) | 2 (2.5) | | | | 37 (46.8) | 42 (53.2) | | | |
| LTA +252 | Pat | 81 (94.2) | 5 (5.8) | 0.1471 | 2.25 | (0.67,7.99) | 17 (19.8) | 69 (80.2) | 0.4698 | 1.31 | (0.59,2.90) |
| | Con | 72 (87.8) | 10(12.2) | | | | 20 (24.4) | 62 (75.6) | | | |
| TNF -863 | Pat | 86 (100.0) | 0 (0.0) | N/A | N/A | N/A | 66 (76.7) | 20 (23.3) | 0.0664 | 2.15 | (0.88,5.36) |
| | Con | 81 (100.0) | 0 (0.0) | | | | 71 (87.7) | 10 (12.3) | | | |

| SNP | | Allele $A_1$ Carrier Status | | | | | Allele $A_2$ Carrier Status | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1/A_1A$ | $A_2A_2$ | p | OR[(a)] | 95%CI | $A_1A_1$ | $A_1A_2/A_2A_2$ | p | OR[(b)] | 95%CI |
| TNF | Pat | 88 (100.0) | 0 (0.0) | N/A | N/A | N/A | 81 (92.0) | 7 (8.0) | 0.4674 | 0.68 | (0.22,2.13) |
| -857 | Con | 80 (100.0) | 0 (0.0) | | | | 71 (88.8) | 9 (11.3) | | | |
| TNF | Pat | 78 (89.7) | 9 (10.3) | 0.2875 | 1.63 | (0.61,4.45) | 11 (12.6) | 76 (87.4) | 0.0760 | 0.35 | (0.09,1.28) |
| -308 | Con | 69 (84.1) | 13 (15.9) | | | | 4 (4.9) | 78 (95.1) | | | |
| BAT2 | Pat | 83 (98.8) | 1 (1.2) | 1.000* | 1.04 | (0.01,82.41) | 51 (60.7) | 33 (39.3) | 0.0986 | 1.74 | (0.86,3.53) |
| rs3130068 | Con | 80 (98.8) | 1 (1.2) | | | | 59 (72.8) | 22 (27.2) | | | |
| BAT2 | Pat | 83 (97.6) | 2 (2.4) | 1.000* | 1.05 | (0.07,14.81) | 39 (45.9) | 46 (54.1) | **0.0273** | 2.01 | (1.03,3.92) |
| rs1046080 | Con | 79 (97.5) | 2 (2.5) | | | | 51 (63.0) | 30 (37.0) | | | |
| BAT2 | Pat | 85 (97.7) | 2 (2.3) | 1.000* | 1.08 | (1.08,15.16) | 40 (46.0) | 47 (54.1) | **0.0273** | 2.00 | (1.03,3.89) |
| rs2272593 | Con | 79 (97.5) | 2 (2.5) | | | | 51 (63.0) | 30 (37.0) | | | |
| BAT2 | Pat | 66 (80.5) | 16 (19.5) | 0.9062 | 1.05 | (0.45,2.43) | 5 (6.1) | 77 (93.9) | 1.0000* | 1.04 | (0.23,4.72) |
| rs1046089 | Con | 63 (79.7) | 16 (20.3) | | | | 5 (6.3) | 74 (93.7) | | | |
| BAT2 | Pat | 83 (100.0) | 0 (0.0) | N/A | N/A | N/A | 63 (75.9) | 20 (24.1) | 0.1337 | 1.83 | (0.77,4.35) |
| rs3132453 | Con | 81 (100.0) | 0 (0.0) | | | | 69 (85.2) | 12 (14.8) | | | |
| BAT2 | Pat | 77 (93.9) | 5 (6.1) | 1.000* | 1.04 | (0.23,4.72) | 5 (6.1) | 77 (93.9) | 0.2333 | 1.98 | (0.57,7.19) |
| rs10885 | Con | 74 (93.7 | 5 (6.3) | | | | 9 (11.4) | 70 (88.6) | | | |

| SNP | | Allele $A_1$ Carrier Status | | | | | Allele $A_2$ Carrier Status | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $A_1A_1/A_1A$ | $A_2A_2$ | p | OR[(a)] | 95%CI | $A_1A_1$ | $A_1A_2/A_2A_2$ | p | OR[(b)] | 95%CI |
| BAT4 | Pat | 73 (86.9) | 11 (13.1) | 0.1330 | 0.44 | (0.12,1.45) | 0 (0.0) | 84 (100.0) | **0.0013*** | N/A | N/A |
| rs3130618 | Con | 76 (93.8) | 5 (6.2) | | | | 9 (11.1) | 72 (88.9) | | | |
| CSNK2B | Pat | 63 (78.8) | 17 (21.3) | 0.6414 | 1.19 | (0.53,2.68) | 5 (6.3) | 75 (93.8) | 1.0000* | 0.81 | (0.15,3.94) |
| rs805257 | Con | 59 (75.6) | 19 (24.4) | | | | 4 (5.1) | 74 (94.9) | | | |
| CSNK2B | Pat | 85 (98.8) | 1 (1.2) | 1.0000* | 1.08 | (0.01,85.44) | 73 (84.9) | 13 (15.1) | 0.5322 | 0.77 | (0.32,1.87) |
| 1942 | Con | 79 (98.8) | 1 (1.3) | | | | 65 (81.3) | 15 (18.8) | | | |
| CSNK2B | Pat | 85 (97.7) | 2 (2.3) | 1.0000* | 1.08 | (1.08,15.16) | 70 (80.5) | 17 (19.5) | 0.0903 | 0.54 | (0.25,1.17) |
| 3854 | Con | 79 (97.5) | 2 (2.5) | | | | 56 (69.1) | 25 (30.9) | | | |
| LY6G5B | Pat | 87 (100.0) | 0 (0.0) | N/A | N/A | N/A | 83 (95.4) | 4 (4.6) | **0.0413** | 0.31 | (0.08,1.11) |
| rs9267532 | Con | 81 (100.0) | 0 (0.0) | | | | 70 (86.4) | 11 (13.6) | | | |
| DDAH | Pat | 81 (95.3) | 4 (4.7) | 0.5301* | 1.60 | (0.36,7.99) | 16 (18.8) | 69 (81.2) | 0.4688 | 0.74 | (0.30,1.80) |
| rs805305 | Con | 76 (92.7) | 6 (7.3) | | | | 12 (14.6) | 70 (85.4) | | | |
| MICB | Pat | 77 (92.8) | 6 (7.2) | 0.1171* | 0.16 | (0.00,1.38) | 11 (13.3) | 72 (86.7) | 0.2618 | 1.61 | (0.65,4.04) |
| rs3134900 | Con | 80 (98.8) | 1 (1.2) | | | | 16 (19.8) | 65 (80.2) | | | |
| MICB | Pat | 67 (89.3) | 8 (10.7) | 0.3045 | 1.65 | (0.58,4.78) | 12 (16.0) | 63 (84.0) | 0.6940 | 0.83 | (0.31,2.25) |
| -7595 | Con | 61 (83.6) | 12(16.4) | | | | 10 (13.7) | 63 (86.3) | | | |

| SNP | | Allele$A_1$CarrierStatus $A_1A_1/A_1A$ | $A_2A_2$ | p | OR[(a)] | 95%CI | Allele$A_2$CarrierStatus $A_1A_1$ | $A_1A_2/A_2A_2$ | p | OR [(b)] | 95%CI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MICB-11715 | Pat | 85 (98.8) | 1 (1.2) | 1.0000* | 1.05 | (0.01,83.33) | 48 (55.8) | 38 (44.2) | 0.4504 | 0.79 | (0.41,1.52) |
| | Con | 81 (97.6) | 1 (2.4) | | | | 41 (50.0) | 41 (50.0) | | | |

a; Odds ratio for comparision of $A_1A_1/ A_1A_2$ to $A_2A_2$ , b; Odds ratio for $A_1A_1/ A_1A_2$ compared to $A_1A_1$; * = Fisher's exact test and exact confidence limits used

### 5.3.5 Haplotype reconstruction accuracy and testing

The frequencies obtained in our study were compared against those observed in the previous study by Allcock *et al.* (2004) to establish if similar haplotype frequency patterns were observed in both populations. For comparison purposes, we use the haplotype frequencies obtained from control individuals only. The frequencies in the Allcock *et al.* study are based on haplotypes reconstructed from 300 Australian Caucasian control individuals. Results of the comparison between populations are presented in table 5.8 and these show similar frequency trends in both groups, although some haplotype frequencies are statistically different between groups. This result is not unexpected when comparing inter-population haplotype frequencies.

Two htSNPs were not genotyped for this haplotype tagging project. HtSNP BAT1int1(rs6149515) was a 15 bp deletion and was not included for analysis, and the assay for htSNP IkBL+1372 (novel to the Allcock *et al* (2004) study) failed. Failure to include these htSNPs, resulted in the collapse of two pairs of haplotypes into single undifferentiated haplotypes, (see figure 5.1). Thus, haplotypes 5 (frequency = 4.5%) and 6 (frequency = 7.3%), converge into one undifferentiated haplotype upon failure to include htSNP IkBL+1372; while haplotypes 4 (frequency = 6.0%) and 7 (frequency = 9.2%), converge into one undifferentiated haplotype upon failure to include htSNP BAT1int1(rs6149515)).

It should be noted that based on published allele frequencies, it appears very likely that Allcock *et al.* (2004) have mislabelled the alleles for the IkBL-62 polymorphism. This is also apparent from the complete mismatch in haplotype frequencies observed between the Irish and Australian populations when the original allele labelling as presented by Allcock *et al.* (2004) is used. Table 5.8 presents the data with what we assume is the correct allele labelling for the IkBL-62 polymorphism in the Allcock *et al.* (2004) data.

**Table 5.8 Comparison of haplotype frequencies between Allcock *et al* study and present control population**

| BAT1-1167 | IkBL-62 | IkBL+446 | IkBL+738 | LTA-379 | TNF-308 | Present study (% haps) | Allcock et al (% haps) | p-value | OR(95%CI) |
|---|---|---|---|---|---|---|---|---|---|
| G | A | C | T | A | A | 11.9 | 15.24 | 0.0804 | 0.75 (0.54,1.05) |
| A | T | C | T | C | G | 20.5 | 14.43 | 0.0045 | 1.52 (1.23,2.05) |
| A | T | C | T | C | A | 14.6 | 11.82 | 0.1497 | 1.27 (0.91,1.78) |
| A | A | T | T | C | A | 8.9 | 9.55 | 0.7072 | 0.93 (0.63,1.38) |
| A | A | C | T | A | A | 13.6 | 7.98 | 0.0014 | 1.80 (1.23,2.64) |
| A | A | C | C | A | A | 8.8 | 7.45 | 0.4085 | 1.18 (0.78,1.80) |
| G | A | C | T | C | A | 11.7 | 5.18 | <0.0005 | 2.44 (1.56,3.83) |
| A | A | C | T | C | A | 3.8 | 4.38 | 0.6347 | 0.88 (0.49,1.57) |

The MHC region is known to be highly polymorphic and the TNF block region investigated here is of great interest in the study of inflammatory diseases, hence it is interesting to investigate if the haplotype tagging approach considered in this study captures additional variation in the region which has not been previously characterised by Allcock *el al.* To investigate this, we have identified and assayed three additional SNPs in the TNF block region, TNF-863, TNF-857 and LTA+252. These SNPs all occur in promoter regions of their respective genes and may individually have a functional role in gene expression. The allele, genotype and carrier status of these SNPs have been assessed individually in DQ2-controlled patient and control populations in tables 5.6 and 5.7 above. No significant association was observed between any of these polymorphisms and disease.

To establish if the information content of these SNPs was captured using the haplotype tagging method employed in this study, the Entropy Maximisation Method (EMM) was carried out by using a web-based program (http://www.well.ox.ac.uk/ rmott/SNPS/webserver_frame.shtml). EMM identifies the percentage haplotype diversity represented by each SNP in the haplotype. Haplotypes constructed from all previous htSNPs in the TNF block, as identified earlier, and the three additional SNPs, TNF-863, TNF-857 and LTA+252 were analysed using EMM, to establish which group of SNP loci were best able to define the total haplotype diversity, or structure, observed. This empirically addresses the question of whether or not the htSNPs, as identified in

this study, capture additional uncharacterised variation in the region, and whether or not there is any information to be gained regarding chromosomal variation by adding extra SNPs. On analysis of these haplotypes, figure 5.3 shows the percentage haplotype diversity explained for these haplotypes reaches 100% using only the htSNPs identified previously. Addition of extra polymorphisms to the haplotype does not create extra diversity or information (i.e. subdivision of haplotypes).

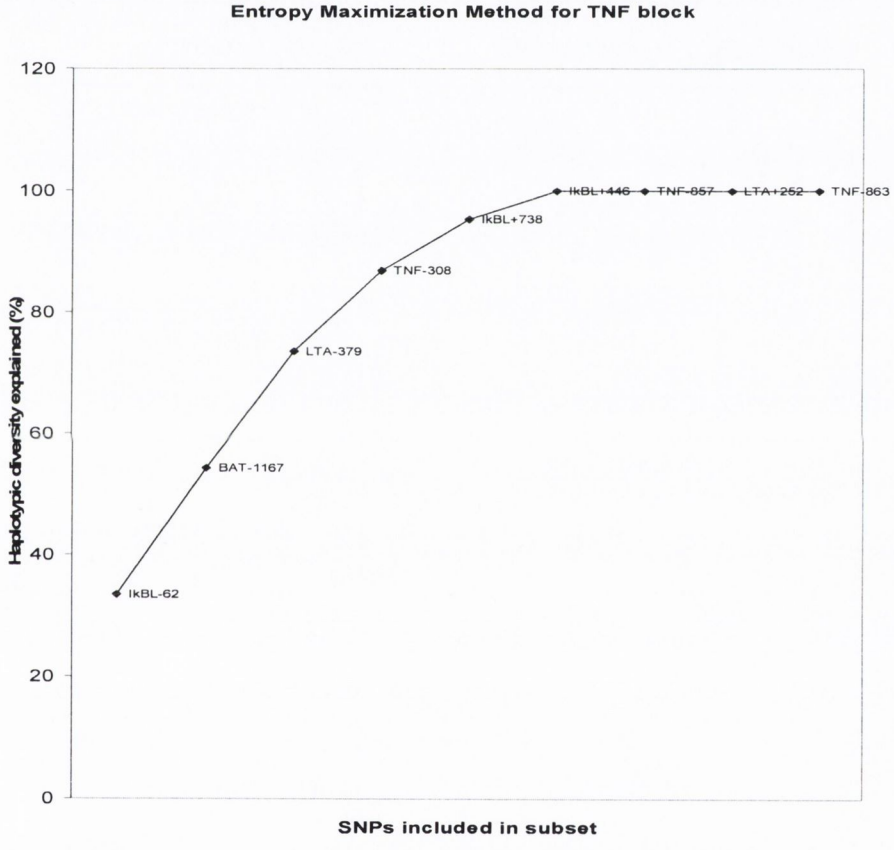



**Figure 5.3 Entropy Maximisation Method allowing identification of the SNP loci which contribute most information on haplotype diversity.**

Due to the high levels of LD present in the region, it is not surprising that only two TNF block haplotypes were found with a frequency greater that 5% in both patient and control groups when the HLA-DQB1 locus was included in the haplotype structure. These haplotypes are listed in table 5.9 below. Haplotypes were generated using PHASEv2.0.2 and compared using the chi-square test. Haplotype 1 contains the high

risk HLA-DQB1*0201 allele since this population was selected such that each individual has only one copy of this allele, and consequently it has a similar frequency in both groups. Haplotype 2 is over-represented in the patient population (patients 10.4%, controls 3.2%, p = 0.0151). This haplotype carries HLA-DQB1*0602 alleles and the reason for its overrepresentation in the patient population is unknown, however this may indicate a contribution from the "low risk" haplotype to disease.

**Table 5.9 TNF block haplotype frequencies (>5%) in combination with the HLA-DQB1 locus in DQ2-controlled patient and control populations**

| DQB1 | BAT1-1167 | IkBL-62 | IkBL+446 | IkBL+738 | LTA-379 | LTA+252 | TNF-863 | TNF-857 | TNFa-308 | p-value | Frequencies | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | Patient | Control |
| 0201 | A | T | C | T | C | G | C | C | A | 0.4733 | 0.44290099 | 0.40769899 |
| 0602 | A | A | C | C | A | A | C | C | G | 0.0151 | 0.11728700 | 0.03750300 |

### 5.3.6 Linkage disequilibrium analysis

To establish the patterns of linkage disequilibrium in the MHC region, both the HAPLOVIEW and GENEPOP programs were utilised. The results from the HAPLOVIEW program are shown in figure 5.4. Shown here are the LD patterns for the SNPs present in the region as measured using D'. This analysis package does not allow for the measurement of LD patterns for microsatellites or HLA alleles such as the HLA-DQ2 locus under investigation here, hence GENEPOP was used to investigate LD between DQB1 and the SNPs in the region. Table 5.10 lists the p-values associated with linkage disequilibrium levels as measured by D' values between DQB1 and all MHC class III SNPs studied in the patient and control groups. Significant LD was found between DQB1 and all SNPs in the patient group, except between DQB1 and LY6G5Brs9267532 (p-value = 0.0980). Only two SNPs failed to show significant linkage with DQB1 in the control group, IkBL+738 (p-value = 0.4192) and BAT2rs3132453 (p-value = 0.1684). Figure 5.4a shows the LD pattern present in the total patient population, and figure 5.4b shows results for the control population. Both result sets show a block of high LD incorporating all BAT2 and BAT4 SNPs and the

neighbouring CSNK2B SNPs. However, the SNPs in these three genes do show slightly higher levels of LD in the control population when compared to the patient population and the block containing the BAT2 / BAT4 SNPs extends slightly further into the CSNK2B SNP cluster in the control population.

Of interest also is the TNF gene cluster. In the patient population, the TNF SNPs are found in high LD and form part of a haplotype block which also incorporates SNPs in the *LTA*, *IkBL* and *BAT1* genes. This extended haplotype block is not seen in the control population, with the TNF SNPs alone forming a haplotype block. It should be noted however, that there are still quite high levels of LD found between the majority of SNPs extending from MICB-7595 to TNF-308 in both populations.

The levels and patterns of LD were also compared and contrasted between DQ2 heterozygous patients, DQ2 homozygous patients and DQ2 heterozygous controls to establish if there were any differences present in the LD structure between these groups. It may be postulated that the DQ2 heterozygous patient group may show higher levels of LD, if the presence of a more highly conserved haplotype structure in the region contributes to disease in heterozygotes. Thus high LD may be a possible signature for the conservation of a secondary disease locus on the haplotype, and may explain why only one high risk haplotype is present in some patients. Results are shown in figures 5.4c-e. Though slight differences are observed between groups, no substantial differences can be detected at a significant level.

**Table 5.10 P-values associated with linkage disequilibrium levels as measured by D' values between DQB1 and all MHC class III SNPs studied in the total patient and control groups**

| SNP | Patient group p-value | Control group p-value |
|---|---|---|
| BAT1-1167 | 0.0000 | 0.0000 |
| IkBL-62 | 0.0000 | 0.0000 |
| IkBL+446 | 0.0000 | 0.0034 |
| IkBL+738 | 0.0000 | 0.4192 |
| LTA-379 | 0.0000 | 0.0000 |
| LTA+252 | 0.0000 | 0.0000 |
| TNF-863 | 0.0000 | 0.0000 |
| TNF-857 | 0.0030 | 0.0372 |
| TNF-308 | 0.0000 | 0.0000 |
| BAT2rs3130068 | 0.0000 | 0.0000 |
| BAT2rs1046080 | 0.0000 | 0.0000 |
| BAT2rs2272593 | 0.0000 | 0.0000 |
| BAT2rs1046089 | 0.0000 | 0.0000 |
| BAT2rs3132453 | 0.0000 | 0.1684 |
| BAT2rs10885 | 0.0000 | 0.0000 |
| BAT4rs3130618 | 0.0000 | 0.0000 |
| CSNK2Brs805257 | 0.0000 | 0.0000 |
| CSNK2B1942 | 0.0000 | 0.0006 |
| CSNK2B3854 | 0.0000 | 0.0001 |
| LY6G5Brs9267532 | 0.0980 | 0.0376 |
| DDAHrs805305 | 0.0000 | 0.0000 |
| MICBrs3134900 | 0.0000 | 0.0000 |
| MICB-7595(rs1065075) | 0.0000 | 0.0000 |
| MICB-11715(rs1065076) | 0.0000 | 0.0000 |

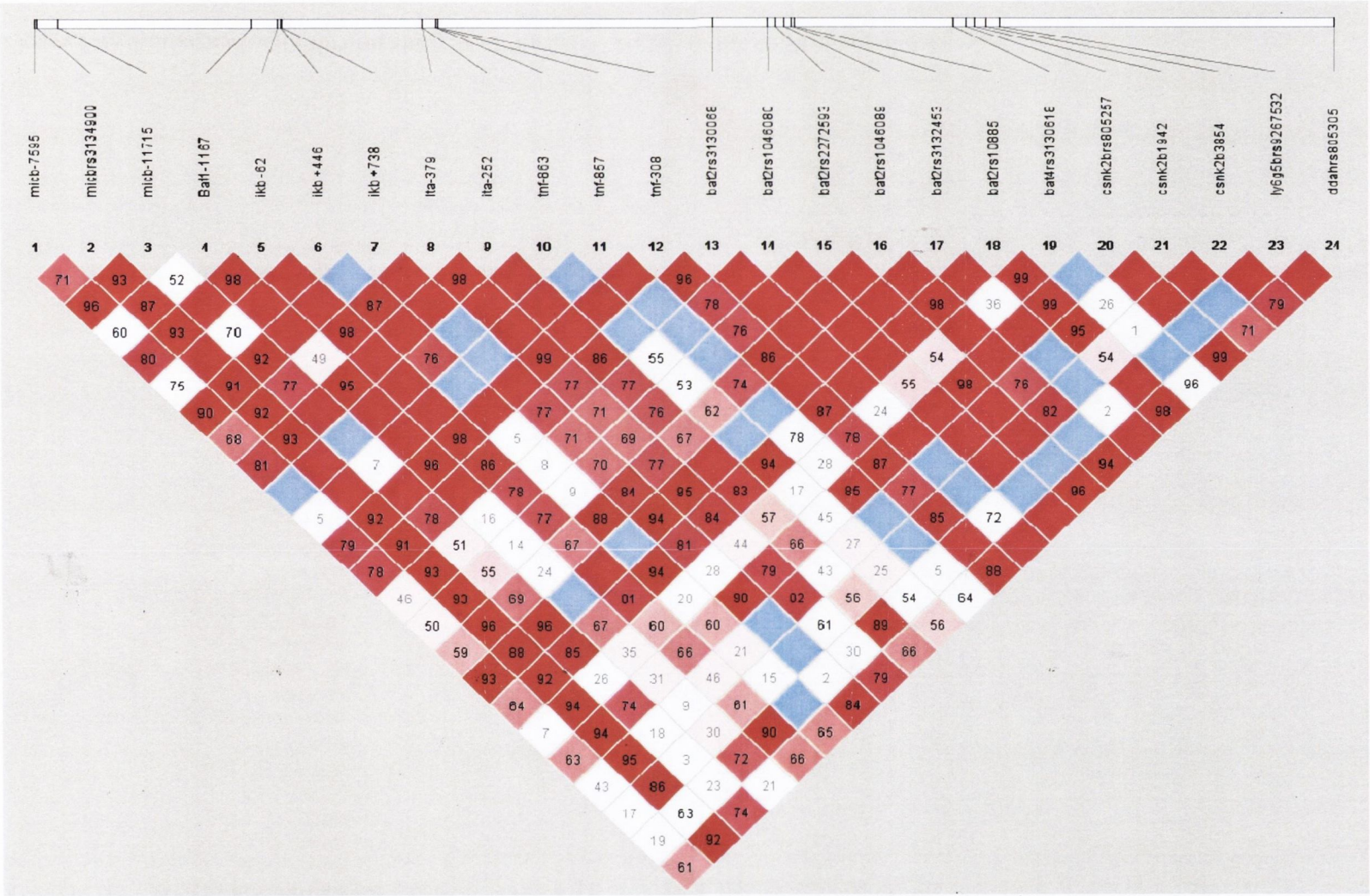



**Figure 5.4a Linkage disequilibrium analysis of the MHC region in coeliac disease cases**

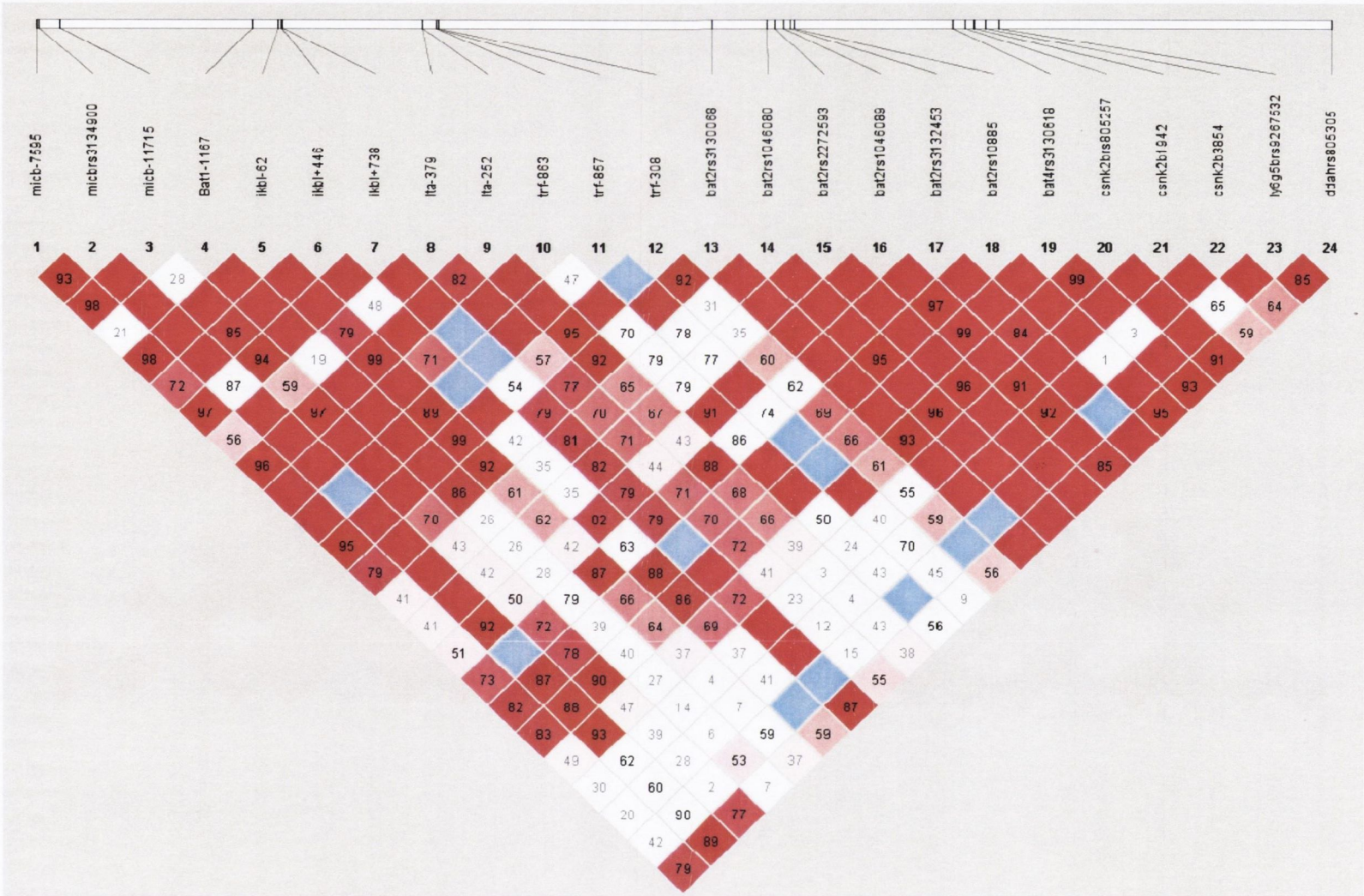



**Figure 5.4b Linkage disequilibrium analysis of the MHC region in the control population**

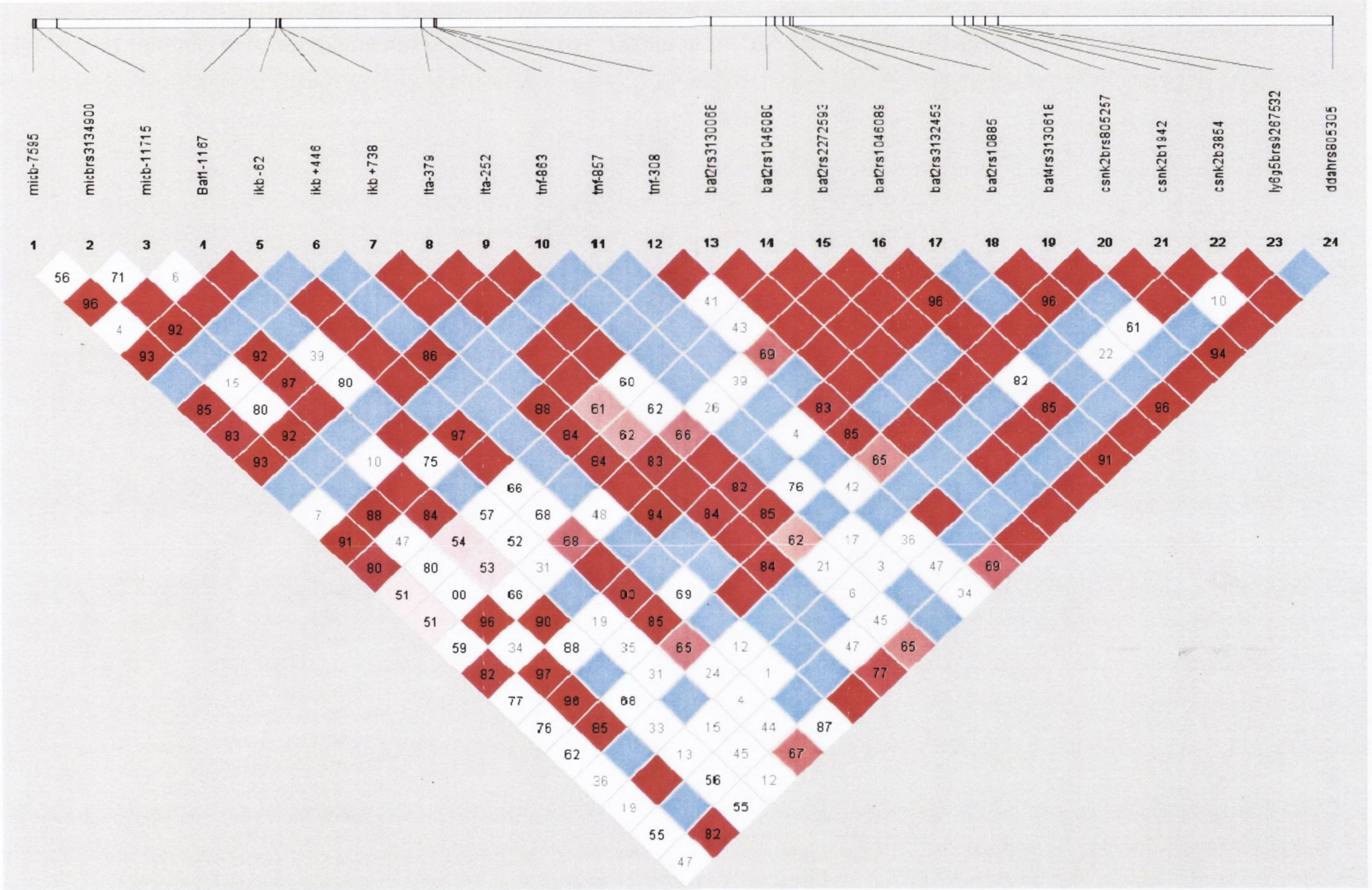



**Figure 5.4c Linkage disequilibrium analysis of the MHC region in the DQ2 heterozygous patient population**

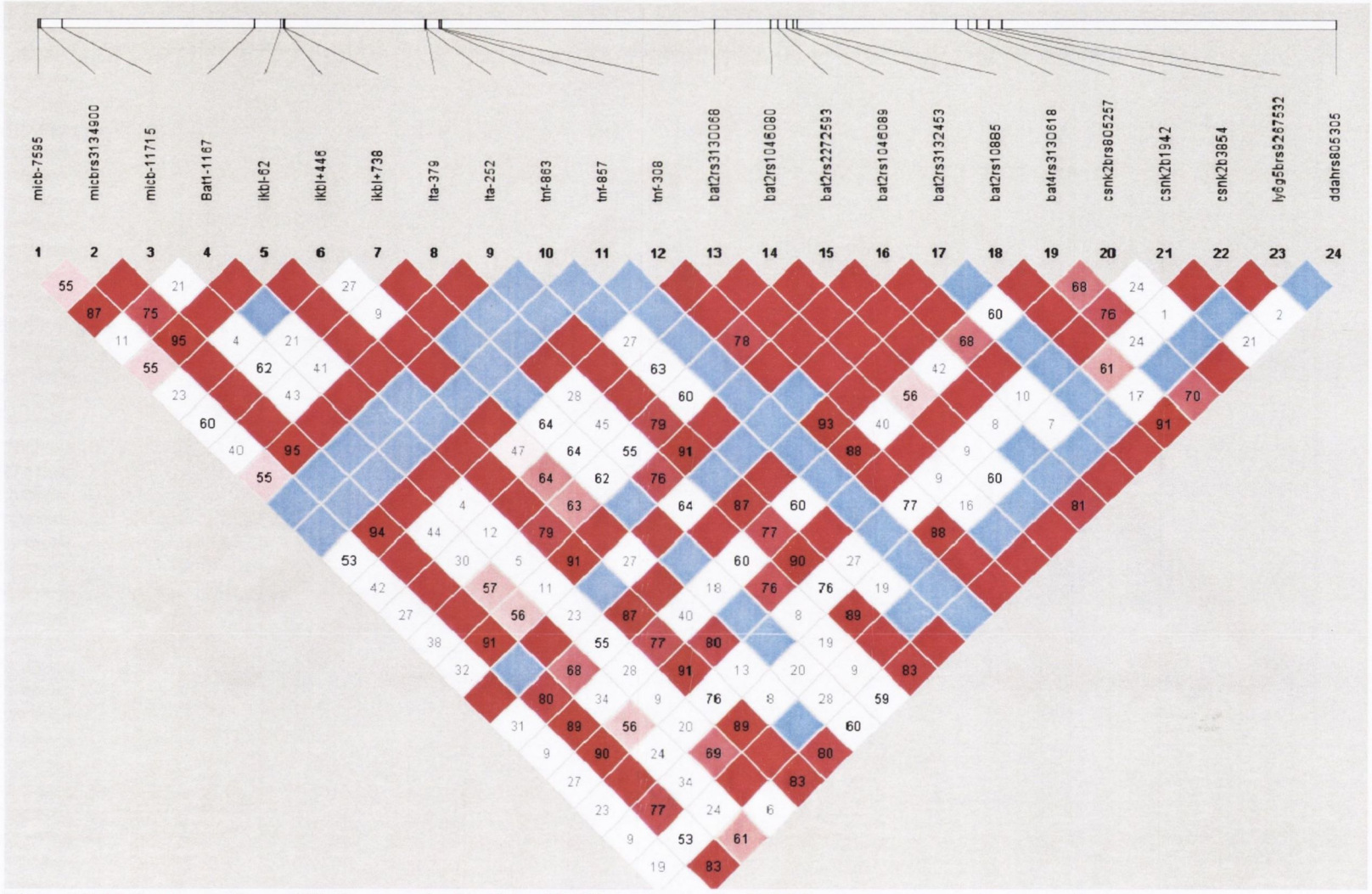



**Figure 5.4d Linkage disequilibrium analysis of the MHC region in the DQ2 homozygous patient population**

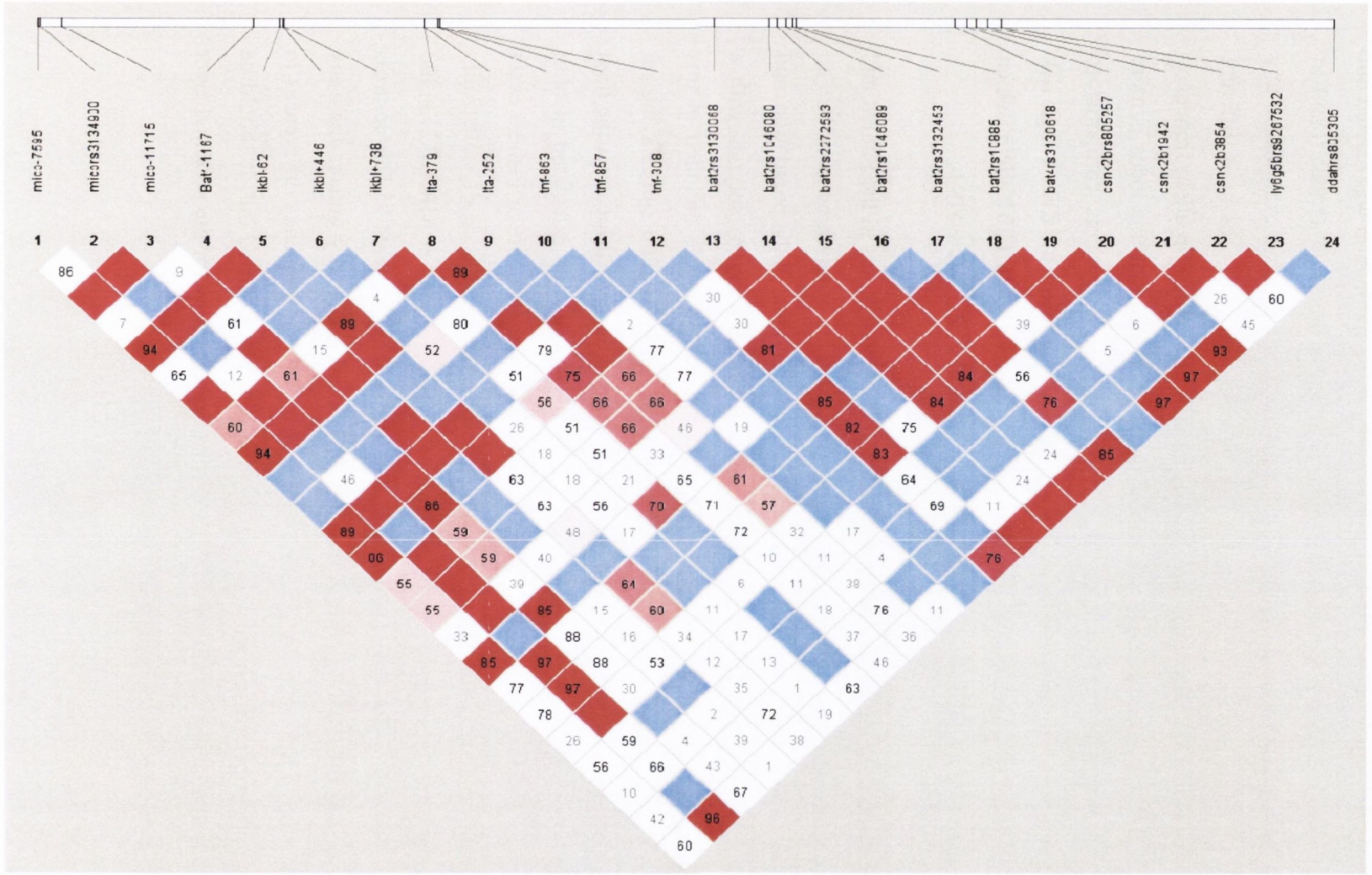



**Figure 5.4e Linkage disequilibrium analysis of the MHC region in the DQ2 heterozygous control population**

### 5.3.7 Analysis of SNP frequency differentials in high risk haplotypes

The aim of this section was to identify any additional CD risk factors which occur on the disease associated extended AH 8.1 haplotype. This examination was carried out to establish if the high risk haplotype found in cases is the same as that found in controls, or whether subtle differences exist between them. Thus we examined if any SNPs in the class III region (and in the *MIC* gene region) were present at different frequencies between cases and controls. Any significant increase in allele frequency at a particular locus on the high risk haplotype may indicate that this locus is more highly conserved on the extended high risk haplotype in patients compared to controls and therefore, that this locus is an additional risk factor for CD. A high risk haplotype is defined here as a haplotype which carries the DQB1*0201 allele (which is indicative of the DRB1*0301 DQB1*0201 extended AH 8.1 haplotype).

In order to do this, the haplotype structure for DQ2 heterozygotes in the patient and control groups was reconstructed using PHASEv2.0.2. This programme generates an output file showing the predicted haplotypes of each individual in the population, and gives the probability that the reconstruction is accurate. Only haplotypes with a probability of reconstruction accuracy of greater than 75% were included in this analysis, with approximately 90% of the haplotype reconstruction calls having on accuracy of greater than 90%. The allele frequencies at each SNP locus on the high risk haplotype of the DQ2 heterozygous individuals were then calculated in both groups and statistical comparisons were carried out using chi-square or Fishers exact test. Results are shown in table 5.11. These results show evidence of statistically significant differences between the patient and control groups at MICBrs3134900 and MICB-7595, which both had similar allele frequencies (patients 95.5%, controls 82.6%. p-value = 0.0162). This frequency differential is actually greater than any difference observed when the general examination of allele frequency differences in the DQ2 heterozygous patient and control populations was carried out. It should be noted that this is an exploratory analysis and we cannot allow for the potential error caused by using computationally generated diplotypes.

**Table 5.11. SNP allele frequencies at each locus on the high risk haplotype of the DQ2 heterozygous patient and control groups**

| SNP | | Allele 1 counts | Allele 2 counts | p-value | OR (95% CI) |
|---|---|---|---|---|---|
| BAT1-1167 | patient | 60 (93.8) | 4 (6.3) | 1.0000 | 1.09(0.22,5.73) |
| | control | 69 (93.2) | 5 (6.8) | | |
| IKBL-62 | patient | 57 (89.1) | 7 (10.9) | 0.5704 | 0.72(0.20,2.56) |
| | control | 68 (91.9) | 6 (8.1) | | |
| IKBL+446 | patient | 62 (96.9) | 2 (3.1) | 0.2133* | 0.00(0.00,4.59) |
| | control | 74 (100) | 0 (0.0) | | |
| IKBL+738 | patient | 64 (100) | 0 (0.0) | N/A | N/A |
| | control | 74 (100) | 0 (0.0) | | |
| LTA-379 | patient | 59 (92.2) | 5 (7.8) | 0.7190* | 0.73(0.10,4.49) |
| | control | 72 (97.3) | 2 (2.7) | | |
| LTA+252 | patient | 57 (89.1) | 7 (10.9) | 0.2495* | 0.33(0.32,4.43) |
| | control | 68 (91.9) | 6 (8.1) | | |
| TNF-863 | patient | 62 (96.9) | 2 (3.1) | 0.5966* | 0.42(0.01,8.39) |
| | control | 73 (98.6) | 1 (1.4) | | |
| TNF-857 | patient | 63 (98.4) | 1 (1.6) | 0.4638* | 0.00(0.00,33.73) |
| | control | 74 (100) | 0 (0.0) | | |
| TNF-308 | patient | 55 (85.9) | 9 (14.1) | 0.5622 | 0.74(0.24,2.27) |
| | control | 66 (89.2) | 8 (10.8) | | |
| BAT2 rs3130068 | patient | 63 (98.4) | 1 (1.6) | 1.0000* | 0.86(0.01,86.81) |
| | control | 73 (98.6) | 1 (1.4) | | |
| BAT2 rs1046080 | patient | 62 (96.9) | 2 (3.1) | 1.0000* | 1.31(0.14,16.12) |
| | control | 71 (95.9) | 3 (4.1) | | |
| BAT2 rs2272593 | patient | 62 (96.9) | 2 (3.1) | 1.0000* | 1.31(0.14,16.12) |
| | control | 71 (95.9) | 3 (4.1) | | |
| BAT2 rs1046089 | patient | 4 (6.3) | 60 (93.8) | 1.0000* | 0.86(0.15,4.82) |
| | control | 4 (5.4) | 70 (94.6) | | |
| BAT2 rs3132453 | patient | 64 (100) | 0 (0.0) | 1.0000* | undefined |
| | control | 73 (98.6) | 1 (1.4) | | |
| BAT2 rs10885 | patient | 5 (7.8) | 59 (92.2) | 1.0000* | 0.86(0.19,3.92) |
| | control | 5 (6.8) | 69 (93.2) | | |

| SNP | | Allele 1 counts | Allele 2 counts | p-value | OR (95% CI) |
|---|---|---|---|---|---|
| BAT4 rs3130618 | patient | 2 (3.1) | 62 (96.9) | 0.4499* | 2.25(0.35,24.26) |
| | control | 5 (6.8) | 69 (93.2) | | |
| CSNK2B rs805257 | patient | 3 (4.7) | 61 (95.3) | 1.0000* | 0.86(0.11,6.66) |
| | control | 3 (4.1) | 71 (95.9) | | |
| CSNK2B 1942 | patient | 62 (96.9) | 2 (3.1) | 0.5966* | 0.42(0.01,8.39) |
| | control | 73 (98.6) | 1 (1.4) | | |
| CSNK2B 3854 | patient | 61 (95.3) | 3 (4.7) | 0.5040* | 1.79(0.36,11.51) |
| | control | 68 (91.9) | 6 (8.1) | | |
| LY6G5B rs9267532 | patient | 63 (98.4) | 1 (1.6) | 0.2159* | 4.57(0.49,219.28) |
| | control | 69 (93.2) | 5 (6.8) | | |
| DDAH rs805305 | patient | 61 (95.3) | 3 (4.7) | 1.0000* | 1.16(0.19,8.24) |
| | control | 70 (94.6) | 4 (5.4) | | |
| MICB rs3134900 | patient | 3 (4.5) | 64 (95.5) | **0.0162** | 4.49(1.10,21.21) |
| | control | 12 (17.4) | 57 (82.6) | | |
| MICB -7595 | patient | 64 (95.5) | 3 (4.5) | **0.0162** | 4.49(1.10,21.21) |
| | control | 57 (82.6) | 12 (17.4) | | |
| MICB -11715 | patient | 66 (98.5) | 1 (1.5) | 0.6195* | 3.00(0.23,159.83) |
| | control | 66 (95.7) | 3 (4.3) | | |

* = Fishers exact test and exact confidence limits used

Patients who were heterozygous for the high risk DQB1*0201 allele were then compared to patients who were homozygous for the high risk DQB1*0201 allele and its extended haplotype. This investigation was carried out using the same statistical analysis as that described above for analysis of DQ2 heterozygous patients versus DQ2 heterozygous controls. The aim here was to investigate if there are additional disease susceptibility factors present at a significantly higher frequency on the high risk haplotype of DQ2 heterozygous patients compared to DQ2 homozygous patients. Again this may go some way to explain why only one copy of the high risk allele is required for some patients to develop CD. Results are shown in table 5.12. A significantly higher frequency of the MICBrs3134900 C allele was observed on the high risk haplotype of DQ2 heterozygous individuals compared to DQ2 homozygotes (DQ2 heterozygotes 95.5%, DQ2 homozygotes 84.8%, p = 0.0249). A similar finding was observed for the MICB-7595 G allele (DQ2 heterozygotes 95.5 %, DQ2 homozygotes 85.5%, p = 0.0202). It is of interest to note that the allele

frequencies of MICBrs3134900 and MICB-7595 in the DQ2 homozygous patients are similar to (i.e. do not differ significantly from) the allele frequencies of these SNPs in the control group (For MICBrs3134900 C allele; DQ2 homozygous patients 84.8%, DQ2 heterozygous controls 82.6%, $p = 0.6871$. For MICB-7595 G allele; DQ2 homozygous patients 85.5%, DQ2 heterozygous controls 82.6%, $p = 0.7574$ (data not shown)).

**Table 5.12 SNP allele frequencies at each locus on the high risk haplotype of the DQ2 heterozygous and DQ2 homozygous patients**

| SNP | | Allele 1 counts | Allele 2 counts | p-value | OR(95%CI) |
|---|---|---|---|---|---|
| BAT1-1167 | Heter | 60 (93.8) | 4 (6.3) | 1.0000* | 1.00(0.26,4.72) |
| | Homo | 120 (93.8) | 8 (6.3) | | |
| IKBL-62 | Heter | 58 (90.6) | 6 (9.4) | 0.6888 | 1.38(0.47,4.20) |
| | Homo | 112 (87.5) | 16 (12.5) | | |
| IKBL+446 | Heter | 62 (96.9) | 2 (3.1) | 1.0000* | 0.74(0.08,9.14) |
| | Homo | 125 (97.7) | 3 (2.3) | | |
| IKBL+738 | Heter | 64 (100) | 0 (0.0) | 0.5532* | undefined |
| | Homo | 126 (98.4) | 2 (1.6) | | |
| LTA-379 | Heter | 59 (92.2) | 5 (7.8) | 0.1196* | 0.28(0.04,1.52) |
| | Homo | 125 (97.7) | 3 (2.3) | | |
| LTA+252 | Heter | 57 (89.1) | 7 (10.9) | 0.7533 | 1.16(0.42,3.33) |
| | Homo | 112 (87.5) | 16 (12.5) | | |
| TNF-863 | Heter | 62 (96.9) | 2 (3.1) | 0.6018* | 0.49(0.04,6.96) |
| | Homo | 126 (98.4) | 2 (1.6) | | |
| TNF-857 | Heter | 63 (98.4) | 1 (1.6) | 0.3333* | 0.00(0.00,19.50) |
| | Homo | 128 (100.0) | 0 (0.0) | | |
| TNF-308 | Heter | 55 (85.9) | 9 (14.1) | 1.0000 | 1.00(0.39,2.59) |
| | Homo | 110 (85.9) | 18 (14.1) | | |
| BAT2 rs3130068 | Heter | 63 (98.4) | 1 (1.6) | 1.0000* | 1.51(0.12,80.62) |
| | Homo | 125 (97.7) | 3 (2.3) | | |
| BAT2 rs1046080 | Heter | 62 (96.9) | 2 (3.1) | 1.0000* | 1.00(0.14,11.34) |
| | Homo | 124 (96.9) | 4 (3.1) | | |
| BAT2 rs2272593 | Heter | 62 (96.9) | 2 (3.1) | 1.0000* | 0.77(0.09,9.52) |
| | Homo | 120 (93.8) | 8 (6.3) | | |
| BAT2 rs1046089 | Heter | 4 (6.3) | 60 (93.8) | 1.0000* | 1.12(0.30,5.16) |
| | Homo | 9 (7.0) | 121 (94.5) | | |
| BAT2 rs3132453 | Heter | 64 (100) | 0 (0.0) | 0.5532* | undefined |
| | Homo | 126 (98.4) | 2 (1.6) | | |
| BAT2 rs10885 | Heter | 5 (7.8) | 59 (92.2) | 0.5588 | 1.57(0.50,5.21) |
| | Homo | 15 (11.7) | 113 (88.3) | | |

| SNP | | Allele 1 counts | Allele 2 counts | p-value | OR(95%CI) |
|---|---|---|---|---|---|
| BAT4 rs3130618 | Heter<br>Homo | 2 (3.1)<br>10 (7.8) | 62 (96.9)<br>118 (92.2) | 0.3431* | 2.63(0.53,25.29) |
| CSNK2B rs805257 | Heter<br>Homo | 3 (4.7)<br>9 (7.0) | 61 (95.3)<br>119 (93.0) | 0.7538* | 1.54(0.37,9.13) |
| CSNK2B 1942 | Heter<br>Homo | 62 (96.9)<br>124 (96.9) | 2 (3.1)<br>4 (3.1) | 1.0000* | 1.00(0.14,11.34) |
| CSNK2B 3854 | Heter<br>Homo | 61 (95.3)<br>119 (93.0) | 3 (4.7)<br>9 (7.0) | 0.7538* | 1.54(0.37,9.13) |
| LY6G5B rs9267532 | Heter<br>Homo | 63 (98.4)<br>123 (96.1) | 1 (1.6)<br>5 (3.9) | 0.6655* | 2.56(0.28,123.02) |
| DDAH rs805305 | Heter<br>Homo | 61 (95.3)<br>118 (92.2) | 3 (4.7)<br>10 (7.8) | 0.5495* | 1.72(0.42,10.07) |
| MICB rs3134900 | Heter<br>Homo | 3 (4.5)<br>21 (15.2) | 64 (95.5)<br>117 (84.8) | **0.0249** | 3.83(1.03,16.81) |
| MICB -7595 | Hetero<br>Homo | 64 (95.5)<br>118 (85.5) | 3 (4.5)<br>22 (15.9) | **0.0202** | 3.98(1.07, 17.39) |
| MICB -11715 | Heter<br>Homo | 66 (98.5)<br>127 (92.0) | 1 (1.5)<br>11 (8.0) | 0.1085* | 5.72(0.79,249.72) |

* = Fishers exact test and exact confidence limits used

Based on the results presented in table 5.12, there are observed differences in allele frequency for a number of polymorphisms when DQ2 homozygous and DQ2 heterozygous patients are compared. We wished to assess whether this is the result of underlying differences in the frequencies of the high risk extended haplotype between these two patient groups. To do this, the frequencies of the high risk extended haplotype (as defined by the presence of the DQB1*0201 allele and including all other linked SNP loci genotyped in this study) were estimated and compared between groups. Results are listed in table 5.13. Only one haplotype had a frequency greater than 5% in both the patient groups. DQ2 heterozygous individuals carry a significantly higher proportion of the conserved high risk haplotype (suggesting less conservation and more breakdown of the disease haplotype in homozygous patients). These results show that in DQ2 heterozygous patients, 89.2% of all haplotypes present were an identical conserved haplotype. While in DQ2

homozygous patients, only 74.4% of all haplotypes present were the same identical conserved haplotype.

However, 86.6% of the DQ2 homozygous patients carried one or more copies of this conserved haplotype (table 5.14). Therefore, in both DQ2 heterozygous and DQ2 homozygous patient groups, a comparable number of individuals had at least one copy of the conserved haplotype. Furthermore, these results suggest that only one copy of the high risk extended haplotype is present in a significant proportion of DQ2 homozygous individuals. The other haplotype in these individuals tends to be less conserved in the class III region and hence there is a lower proportion of high risk conserved haplotypes in this group. These results suggest that only one copy of the extended haplotype is necessary for CD development.

**Table 5.13 DQB1*0201 allele-carrying MHC haplotype frequencies (>5%) in DQ2 heterozygous and DQ2 homozygous patient populations**

| Batl-1167 | ikbl-62 | ikbl+446 | ikbl+738 | lta-379 | ltat252 | tnf-863 | tnf-857 | tnf-308 | bat2rs3130068 | bat2rsl046080 | bat2rs2272593 | bat2rsl046089 | bat2rs3132453 | bat2rsl0885 | bat4rs3130618 | csnk2brs8052257 | csnk2b1942 | csnk2b3854 | ly6gSbrs9267532 | ddah | micbrs3134900 | micb-7595 | micb-11715 | DQB1 | Frequencies | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | | | | | | | | | | p-value | DQ2 hetero | DQ2 homo |
| A | T | C | T | C | G | C | C | A | A | A | C | A | G | T | A | G | A | G | C | C | G | G | G | 0201 | 0.0102 | 0.8916 | 0.7440 |

hetero = heterozygous; homo = homozygous

Based on individuals with haplotypes with a probability of reconstruction accuracy >75%

**Table 5.14 Number of HLA-DQB1*0201 allele-carrying conserved haplotypes in DQ2 homozygous patients**

| | DQ2 homozygous Patients | % |
|---|---|---|
| 2 copies | 44 | 65.7 |
| 1 copy | 14 | 20.9 |
| no copies | 9 | 13.4 |
| Total | 67 | 100 |

### 5.3.8 Analysis of low risk (non-DQB1*0201) haplotypes

The results in this analysis have, to date, found that when allele, genotype and carrier status frequencies are compared between DQ2 heterozygous patients and controls, a number of MHC class III SNPs show significant differences between the two groups (section 5.3.4). The high risk (DQB1*0201-carrying) haplotype was then analysed to see if there were any differences in this extended haplotype between DQ2 heterozygous patients and controls which may explain the variation in allele frequency observed in the total group (section 5.3.7). In this section, the low risk (non DQB1*0201-carrying) haplotype has been analysed for differences between the DQ2 heterozygous patient and control groups.

Table 5.15 lists the DQB1 allele frequencies observed in both DQ2 heterozygous patients and controls. The high risk allele (DQB1*0201) is present in equal proportions in both groups (50%) by definition, as these groups were selected based on the fact that they carry one copy of this allele. There are no DQB1*0202 alleles present in these groups, for reasons outlined previously (section 5.3.1). Table 5.15 shows that there is indeed an overall significant difference in the frequency distribution of these alleles between DQ2 heterozygous patient and control groups ($p = 0.0014$). The largest allele frequency differential is observed at the DQB1*0602 allele (patients 24.7%, controls 12.7%). These results suggest that the DQB1*0602 allele may be an additional, previously unidentified CD susceptibility or modifier allele at the DQB1 locus, or that it is linked to other susceptibility markers.

**Table 5.15 DQB1 allele frequencies in the DQ2 heterozygous patient and control groups**

| DQB1 Allele | Patient group n (%) | Control group n (%) |
|---|---|---|
| 0201 | 89 (50.0) | 83 (50.0) |
| 0301 | 9 (5.1) | 12 (7.2) |
| 0302 | 11 (6.2) | 14 (8.4) |
| 0303 | 4 (2.2) | 1 (0.6) |
| 03032 | 3 (1.7) | 0 (0.0) |
| 0304 | 9 (5.1) | 13 (7.8) |
| 0402 | 0 (0.0) | 1 (0.6) |
| 0501 | 3 (1.7) | 6 (3.6) |
| 0502 | 1 (0.6) | 1 (0.6) |
| 0503 | 1 (0.6) | 0 (0.0) |
| 0504 | 0 (0.0) | 7 (4.2) |
| 0602 | **44 (24.7)** | **21 (12.7)** |
| 0603 | 3 (1.7) | 0 (0.0) |
| 0604 | 4 (2.2) | 2 (1.2) |
| 0609 | 0 (0.0) | 1 (0.6) |
| Total | 178 | 166 |

overall p-value = 0.0014

When the frequency of the low risk haplotypes (non DQB1*0201-carrying) are compared between DQ2 heterozygous patients and controls, a similar increase in the patient frequency of the DQB1*0602-carrying haplotypes which mirrors the increase in the patient frequency of the DQB1*0602 allele is observed. Table 5.16 lists the haplotype frequencies of the DQB1*0602-carrying haplotypes based as a percentage of the overall frequencies of the low risk haplotypes with a probability of reconstruction accuracy of greater than 75%. (i.e. all low risk haplotypes with 75% or greater reconstruction accuracy total 100%).

**Table 5.16 DQB1*0602 allele-carrying MHC haplotype frequencies as a proportion of the low risk (non DQB1*0201-carrying) in DQ2 heterozygous patient and control populations**

| Bat1-1167 | ikbl-62 | ikbl+446 | ikbl+738 | lta-379 | lta+252 | tnf-863 | tnf-857 | tnf-308 | bat2rs3130068 | bat2rs1046080 | bat2rs2272593 | bat2rs1046089 | bat2rs3132453 | bat2rs10885 | bat4rs3130618 | csnk2brs8052257 | csnk2b1942 | csnk2b3854 | ly6g5brs9267532 | ddah | micbrs3134900 | micb-7595 | micb-11715 | DQB1 | n (frequencies) p-value | DQ2 het Patients | DQ2 het controls |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | A | C | C | A | A | C | C | G | G | C | T | G | T | C | C | A | A | G | C | G | C | A | G | 0602 | 0.0009 | 17(29.8) | 6(7.9) |
| A | A | C | T | A | A | C | C | G | G | C | T | G | G | C | C | A | A | G | C | G | C | G | G | 0602 | 0.1519 | 8(14.0) | 5(6.6) |
| A | T | C | T | C | G | C | C | A | A | A | C | G | G | C | C | A | A | G | C | G | C | G | G | 0602 | n/s | 1(1.8) | 0(0.0) |
| A | T | C | T | C | G | C | C | A | A | C | T | G | G | C | C | A | A | G | C | G | C | A | G | 0602 | n/s | 1(1.8) | 0(0.0) |
| A | T | C | T | C | G | C | C | G | A | A | C | G | G | C | C | A | A | G | C | G | C | A | G | 0602 | n/s | 1(1.8) | 3(3.9) |
| A | T | C | T | C | G | C | C | G | A | A | C | A | G | C | C | G | G | A | C | C | C | A | G | 0602 | n/s | 1(1.8) | 1(1.3) |
| G | A | C | T | C | A | C | C | G | A | A | C | G | G | C | C | A | A | G | C | G | C | A | A | 0602 | n/s | 1(1.8) | 0(0.0) |
| G | A | C | T | C | A | C | T | G | A | A | C | G | G | C | C | A | A | G | C | G | C | G | G | 0602 | n/s | 0(0.0) | 2(2.6) |

Total number of low risk haplotypes with a probability of reconstruction accuracy greater than 75% in patients and controls is 57 and 76, respectively.

het = heterozygous; n/s = not significant

Thus this data shows that there are significant differences between the groups in the frequency of the DQB1*0602-carrying haplotype which extends into the MHC class III region. This is therefore likely to perturb the observed SNP allele frequencies in the MHC class III region. It is possible that there may be additional (MHC class III) risk factors on the DQB1*0602-carrying haplotype which predispose to CD. To test this hypothesis, allele frequency differences were compared in a number of different case scenarios;

1. Comparison of allele frequencies on all low risk haplotypes (non DQB1*0201-carrying) in DQ2 heterozygous patients and DQ2 heterozygous controls (table 5.17).
2. Comparison of allele frequencies on the DQB1*0602-carrying haplotypes in DQ2 heterozygous patients and DQ2 heterozygous controls (table 5.18).
3. Comparison of allele frequencies on all low risk haplotypes except DQB1*0602-carrying haplotypes in DQ2 heterozygous patients and DQ2 heterozygous controls (table 5.19).

To carry out these analyses we used the same method as used in section 5.3.7, whereby we computationally reconstructed haplotypes in each individual of the population using Phasev2.0.2. software. All haplotypes included for analysis here had a probability of haplotype reconstruction accuracy greater than 75%. Results show a number of loci are associated with disease at the 5% level when allele frequencies on all low risk haplotypes are compared between the DQ2 heterozygous patients and controls (table 5.17). All loci which we originally found to be associated with disease when total allele, genotype or carrier status frequencies in the DQ2 heterozygous individuals were compared (tables 5.6 and 5.7) were also found to be associated when the allele frequencies on the low risk haplotypes alone in this group was studied. This would suggest that the low risk haplotype is a contributory factor in the frequency differences observed at the disease associated loci shown in tables 5.6 and 5.7. An exception, however, was SNP BAT4rs3130618 which was not found to have a significant disease association when low risk haplotype alleles were examined. Also SNP LY6G5Brs9267532 which displayed a borderline significant association with disease in the original total DQ2 heterozygous groups analysis (p-value = 0.0464), had a p-value just above the 5% threshold for association (p-value =

0.0677) upon examination of low risk haplotype allele frequencies. This analysis also revealed some additional disease associations which were not seen when all alleles in the DQ2 heterozygous groups were compared. IkBL+738 was found to be associated with disease in the present analysis (patients 67.8%, controls 86.9%, p-value = 0.0075), as was SNP TNF-863 (patients 78.0%, controls 90.8%, p-value = 0.0375), and BAT2rs3130068 (patients 57.8%, controls 76.3%, p-value = 0.0208).

Previous analysis carried out in this study has shown that there is a significant difference in the allele frequency of the DQB1*0602 allele between DQ2 heterozygous patients and controls. All other DQB1 alleles have comparable frequencies between groups. Here we have now tested whether the difference in DQB1*0602-carrying haplotype frequencies is responsible for the variation in class III allele frequencies, seen on the low risk haplotypes between groups.

Table 5.18 lists the allele frequencies and differences observed on DQB1*0602-carrying haplotypes only, in the DQ2 heterozygous patient and control populations. All loci fail to reach a statistically significant level of disease association, although there are large differences in allele frequencies observed between groups at some loci. Unfortunately, the small sample sizes tested here preclude a meaningful statistical analysis of the results. The next investigation assessed the remaining (non DQB1*0602-carrying) low risk haplotypes. This assessment examined whether the differences in allele frequencies, seen when the low risk haplotypes are compared in heterozygotes, are caused only by the enrichment for haplotypes carrying DQB1*0602 in patients, with the remaining low risk haplotypes playing no role in allele differentiation. To do this, the allele frequencies on all non DQB1*0602, low risk haplotypes were compared between groups.

Results are shown in table 5.19. Examining alleles found only on non DQB1*0602 low risk haplotypes, IkBL+738, BAT2rs3130068, BAT2rs1046080, BAT2rs2272593 and BAT2rs3132453 were no longer found to be associated with disease. These SNPs now display similar allele frequencies on the remaining low risk haplotypes in the DQ2 heterozygous patient and control groups. This suggests that the disease association observed for particular alleles at these MHC

class III SNPs is caused by linkage disequilibrium with DQB1*0602 on extended haplotypes which are at significantly different frequencies between the DQ2 heterozygous patient and control groups. After removal of the DQB1*0602-carrying haplotypes, only SNPs BAT1-1167, IkBL+446 and TNF-863 are found to be associated with disease.

**Table 5.17 SNP allele frequencies at each locus on the low risk haplotype of the DQ2 heterozygous patient and control groups**

| SNP | | Allele 1 counts | Allele 2 counts | p-value | OR (95% CI) |
|---|---|---|---|---|---|
| BAT1-1167 | patient | 54 (91.5) | 5 (8.5) | **0.0007** | 5.29(1.74,17.17) |
| | control | 51 (67.1) | 25 (32.9) | | |
| IKBL-62 | patient | 14 (23.7) | 45 (76.3) | 0.3985 | 0.72(0.31,1.66) |
| | control | 23 (30.3) | 53 (69.7) | | |
| IKBL+446 | patient | 47 (76.7) | 12 (20.3) | **0.0072** | 0.22(0.06,0.79) |
| | control | 72 (94.7) | 4 (5.3) | | |
| IKBL+738 | patient | 40 (67.8) | 19 (32.2) | **0.0075** | 0.32(0.12,0.81) |
| | control | 66 (86.9) | 10 (13.2) | | |
| LTA-379 | patient | 30 (50.8) | 29 (49.2) | 0.8370 | 0.93(0.45,1.95) |
| | control | 40 (52.6) | 36 (47.4) | | |
| LTA+252 | patient | 14 (23.7) | 45 (76.3) | 0.4964 | 0.76(0.33,1.78) |
| | control | 22 (28.9) | 54 (71.1) | | |
| TNF-863 | patient | 46 (78.0) | 13 (22.0) | **0.0375** | 0.36(0.12,1.06) |
| | control | 69 (90.8) | 7 (9.2) | | |
| TNF-857 | patient | 57 (96.6) | 2 (3.4) | 0.1852* | 3.35(0.63,33.39) |
| | control | 68 (89.5) | 8 (10.5) | | |
| TNF-308 | patient | 8 (13.6) | 51 (86.4) | 0.4252 | 0.65(0.20,2.13) |
| | control | 7 (9.2) | 69 (90.8) | | |
| BAT2 rs3130068 | patient | 34 (57.6) | 25 (42.4) | **0.0208** | 0.42(0.19,0.94) |
| | control | 58 (76.3) | 18 (23.7) | | |
| BAT2 rs1046080 | patient | 24 (40.7) | 35 (59.3) | **0.0146** | 0.42(0.20,0.90) |
| | control | 47 (61.8) | 29 (38.2) | | |
| BAT2 rs2272593 | patient | 24 (40.7) | 35 (59.3) | **0.0146** | 0.42(0.20,0.90) |
| | control | 47 (61.8) | 29 (38.2) | | |
| BAT2 rs1046089 | patient | 53 (89.8) | 6 (10.2) | 0.0615 | 2.55(0.86,7.86) |
| | control | 59 (77.6) | 17 (22.4) | | |
| BAT2 rs3132453 | patient | 42 (71.2) | 17 (28.8) | **0.0415** | 0.42(0.16,1.06) |
| | control | 65 (85.5) | 11 (14.5) | | |
| BAT2 rs10885 | patient | 56 (94.9) | 3 (5.1) | 0.7310* | 1.60(0.32,10.29) |
| | control | 70 (92.1) | 6 (7.9) | | |

| SNP | | Allele 1 counts | Allele 2 counts | p-value | OR (95% CI) |
|---|---|---|---|---|---|
| BAT4 rs3130618 | patient | 57 (96.6) | 2 (3.4) | 0.4652* | 2.44(0.41,25.49) |
| | control | 70 (92.1) | 6 (7.9) | | |
| CSNK2B rs805257 | patient | 51 (86.4) | 8 (13.6) | 0.1389 | 1.98(0.73,5.46) |
| | control | 58 (76.3) | 18 (23.7) | | |
| CSNK2B 1942 | patient | 53 (89.8) | 6 (10.2) | 0.3407 | 1.66(0.53,5.36) |
| | control | 64 (84.2) | 12 (15.8) | | |
| CSNK2B 3854 | patient | 53 (89.8) | 6 (10.2) | 0.0615 | 2.55(0.86,7.86) |
| | control | 59 (77.6) | 17 (22.4) | | |
| LY6G5B rs9267532 | patient | 59 (100.0) | 0 (0.0) | 0.0677* | undefined |
| | control | 71 (93.4) | 5 (6.6) | | |
| DDAH rs805305 | patient | 10 (16.9) | 49 (83.1) | 0.8564 | 0.92(0.34,2.53) |
| | control | 12 (15.8) | 64 (84.2) | | |
| MICB rs3134900 | patient | 56 (94.9) | 3 (5.1) | 0.3181* | 0.25(0.00,3.22) |
| | control | 75 (98.7) | 1 (1.3) | | |
| MICB -7595 | patient | 12 (20.3) | 47 (79.7) | 0.3686 | 0.66(0.25,1.78) |
| | control | 11 (14.5) | 65 (85.5) | | |
| MICB -11715 | patient | 32 (54.2) | 27 (45.8) | 0.2954 | 0.69(0.33,1.46) |
| | control | 48 (63.2) | 28 (36.8) | | |

* = Fisher's exact test and exact confidence intervals used

**Table 5.18 SNP allele frequencies at each locus on the DQB1*0602-carrying haplotypes only for the DQ2 heterozygous patient and control groups**

| SNP | | Allele 1 counts | Allele 2 counts | p-value | OR (95% CI) |
|---|---|---|---|---|---|
| BAT1-1167 | patient | 29 (96.7) | 1 (3.3) | 0.5439* | 3.87(0.18,235.80) |
| | control | 15 (88.2) | 2 (11.8) | | |
| IKBL-62 | patient | 4 (13.3) | 26 (86.7) | 0.4350* | 2.00(0.31,12.37) |
| | control | 4 (23.5) | 13 (76.5) | | |
| IKBL+446 | patient | 30 (100.0) | 0 (0.0) | 0.3617* | undefined |
| | control | 16 (94.1) | 1 (5.9) | | |
| IKBL+738 | patient | 12 (40.0) | 18 (60.0) | **0.0438** | 0.28(0.06,1.16) |
| | control | 12 (70.6) | 5 (29.4) | | |
| LTA-379 | patient | 5(16.7) | 25 (83.3) | 0.7027* | 1.54(0.26,8.52) |
| | control | 4 (23.5) | 13 (76.5) | | |
| LTA+252 | patient | 4 (13.3) | 26 (86.7) | 0.4350* | 2.00(0.31,12.47) |
| | control | 4 (23.5) | 13 (76.5) | | |
| TNF-863 | patient | 30 (100.0) | 0 (0.0) | N/A | N/A |
| | control | 17 (100.0) | 0 (0.0) | | |
| TNF-857 | patient | 30 (100.0) | 0 (0.0) | 0.1258* | undefined |
| | control | 15 (88.2) | 2 (11.8) | | |
| TNF-308 | patient | 2 (6.7) | 28 (93.3) | 0.5282* | 0.00(0.00,9.45) |
| | control | 0 (0.0) | 17 (100.0) | | |
| BAT2 rs3130068 | patient | 5 (16.7) | 25 (83.3) | 0.1705* | 2.73(0.55,13.74) |
| | control | 6 (35.3) | 11 (64.7) | | |
| BAT2 rs1046080 | patient | 4 (13.3) | 26 (86.7) | 0.1357* | 3.55(0.67,20.16) |
| | control | 6 (35.3) | 11 (64.7) | | |
| BAT2 rs2272593 | patient | 4 (13.3) | 26 (86.7) | 0.1357* | 3.55(0.67,20.16) |
| | control | 6 (35.3) | 11 (64.7) | | |
| BAT2 rs1046089 | patient | 29 (96.7) | 1 (3.3) | 1.0000* | 1.81(0.02,147.10) |
| | control | 16 (94.1) | 1 (5.9) | | |
| BAT2 rs3132453 | patient | 13 (43.3) | 17 (56.7) | 0.1590 | 0.42(0.10,1.67) |
| | control | 11 (64.7) | 6 (35.3) | | |
| BAT2 rs10885 | patient | 30 (100.0) | 0 (0.0) | N/A | N/A |
| | control | 17 (100.0) | 0 (0.0) | | |

| SNP | | Allele 1 counts | Allele 2 counts | p-value | OR (95% CI) |
|---|---|---|---|---|---|
| BAT4 | patient | 30 (100.0) | 0 (0.0) | N/A | N/A |
| rs3130618 | control | 17 (100.0) | 0 (0.0) | | |
| CSNK2B | patient | 29 (96.7) | 1 (3.3) | 1.0000* | 1.81(0.02,147.10) |
| rs805257 | control | 16 (94.1) | 1 (5.9) | | |
| CSNK2B | patient | 29 (96.7) | 1 (3.3) | 1.0000* | 1.81(0.02,147.10) |
| 1942 | control | 16 (94.1) | 1 (5.9) | | |
| CSNK2B | patient | 29 (96.7) | 1 (3.3) | 1.0000* | 1.81(0.02,147.10) |
| 3854 | control | 16 (94.1) | 1 (5.9) | | |
| LY6G5B | patient | 30 (100.0) | 0 (0.0) | N/A | N/A |
| rs9267532 | control | 17 (100.0) | 0 (0.0) | | |
| DDAH | patient | 1 (3.3) | 29 (96.7) | 1.0000* | 1.81(0.02,147.10) |
| rs805305 | control | 1 (5.9) | 16 (94.1) | | |
| MICB | patient | 29 (96.7) | 1 (3.3) | 1.0000* | 0.00(0.00,68.82) |
| rs3134900 | control | 17 (100.0) | 0 (0.0) | | |
| MICB | patient | 2 (6.7) | 28 (93.3) | 0.6130* | 1.87(0.12,27.79) |
| -7595 | control | 2 (11.8) | 15 (88.2) | | |
| MICB | patient | 7 (23.3) | 23 (76.7) | 0.5005* | 1.79(0.39,7.95) |
| -11715 | control | 6 (35.3) | 11 (64.7) | | |

* = Fisher's exact test and exact confidence intervals used

**Table 5.19 SNP allele frequencies on non DQB1*0602–carrying, low risk haplotypes, in the DQ2 heterozygous patient and control groups**

| SNP | | Allele 1 counts | Allele 2 counts | p-value | OR (95% CI) |
|---|---|---|---|---|---|
| BAT1-1167 | patient | 25 (86.2) | 4 (13.8) | **0.0160** | 3.99(1.11,15.60) |
| | control | 36 (61.0) | 23 (39.0) | | |
| IKBL+446 | patient | 17 (58.6) | 12 (41.4) | **0.00002** | 0.08(0.01,0.34) |
| | control | 56 (94.9) | 3 (5.1) | | |
| IKBL+738 | patient | 28 (96.6) | 1 (3.4) | 0.6590* | 2.59(0.27,126.93) |
| | control | 54 (91.5) | 5 (8.5) | | |
| TNF-863 | patient | 16 (55.2) | 13 (44.8) | **0.0005** | 0.17(0.05,0.54) |
| | control | 52 (88.1) | 7 (11.9) | | |
| BAT2 rs3130068 | patient | 29 (100.0) | 0 (0.0) | 0.0904* | undefined |
| | control | 52 (88.1) | 7 (11.9) | | |
| BAT2 rs1046080 | patient | 20 (69.0) | 9 (31.0) | 0.9599 | 0.98(0.34,2.85) |
| | control | 41 (69.5) | 18 (30.5) | | |
| BAT2 rs2272593 | patient | 20 (69.0) | 9 (31.0) | 0.9599 | 0.98(0.34,2.85) |
| | control | 41 (69.5) | 18 (30.5) | | |
| BAT2 rs3132453 | patient | 29 (100.0) | 0 (0.0) | 0.1666* | undefined |
| | control | 54 (91.5) | 5 (8.5) | | |

### 5.4 Discussion

Genetic association studies of the MHC class III region for polymorphisms which contribute to coeliac disease require a unique approach in the methodologies used. Simple association testing of polymorphisms of interest will not suffice in this genomic region, because linkage disequilibrium between candidate SNPs and the known disease locus at HLA-DQB1 can lead to false positive results. In this study, we have controlled for the LD to HLA-DQB1 by investigating a DQ2 stratified population (heterozygous for DQ2). This population included individuals who had one copy of the HLA-DQB1*0201 allele, known to be a high risk allele for coeliac disease, and one copy of any other HLA-DQB1 allele except the high risk HLA-DQB1*0201 allele or the HLA-DQB1*0202 allele. Removal of the HLA-DQB1*0202 allele from the population of interest was necessitated for a number of reasons including the gene dosage effect seen at the HLA-DQB1 locus in CD (Vader *et al.* 2003). Individuals who have a genotype of HLA-DQB1*0201/0202 have a risk of disease lower but similar to the HLA-DQB1*0201 homozygous individuals, even though the HLA-DQB1*0202 allele alone does not confer any disease risk. Another factor which excludes the use of the HLA-DQB1*0202 allele in the population of interest is that fact that the HLA-DRB1*0301 DQB1*0201 allelic combination is indicative of the extended ancestral haplotype, AH 8.1. This is the extended haplotype of interest in this study as it has been associated with a number of autoimmune diseases. Previous investigations on the genetics of CD point to additional risk factors carried on this haplotype independent of the HLA-DQ2 association (Lopez-Vazquez *et al.* 2002; Louka *et al.* 2003). Hence the use of the HLA-DQB1*0201 allele alone, and the exclusion of the HLA-DQB1*0202 allele, provides a clearly defined population. This is an excellent population, and arguably, the only possible population, in which to assess for the presence of additional disease risk factors in the MHC class III region which are independent of DQ2. It was not possible to compare HLA-DQB1*0201 homozygous individuals between groups due to the small numbers of HLA-DQB1*0201 homozygous individuals in the healthy control population (n = 6). Also, arguably, the comparison of the HLA-DQB1*0201 homozygous versus the HLA-DQB1*0201 heterozygous patients may be useful in this study. This comparison can investigate if individuals who have only one copy of the high risk HLA-DQB1*0201 allele possess any additional disease risk alleles in the MHC class III region, which are not seen in DQB1*0201 homozygous

individuals. The occurrence of such alleles in heterozygotes may be to, perhaps, compensate for the presence of only one DQB1*0201 allele.

Results from this study show, on initial examination of the allele, genotype and carrier status frequencies of the DQ2 heterozygous populations (as defined earlier), that a number of polymorphisms in the MHC class III region are associated with disease at the 5% level of association. These SNPs include BAT1-1167, IkBL+446, BAT4rs3130618, BAT2rs1046080 and BAT2rs2272593 (table 5.6 and 5.7). Though these results have a certain level of interest, this method of examination is not fully controlled for the LD present in the region. The low risk HLA-DQB1 (i.e. non DQB1*0201) allele in the patient and control groups is not controlled. As there is no previous evidence suggesting a role for any other HLA-DQB1 allele in coeliac disease, it is plausible to presume, that there would be similar allele frequencies of all other low risk (non DQB1*0201) DQB1 alleles in DQ2 heterozygous patients and control groups. However, it is possible that there are differing proportions of the different low risk HLA-DQB1 alleles between the groups (A hypothesis which is also examined in this study). As this region is known to have extended levels of high LD, the differences observed in the total examination of allele, genotype and carrier status frequency differences, though interesting, may be due simply to differing low risk haplotypes proportions. Due to the extended levels of LD, these haplotypes, incorporating the MHC class III loci, may be characterised by the specific low risk allele present at the HLA-DQB1 locus. Hence, analysis of the allele, genotype and carrier status frequencies in the DQ2 heterozygous patient and control populations provides a semi-controlled environment in which to assess for the presence of additional coeliac disease risk factors in the MHC class III region. The results presented here, list a number of possible candidates for further study.

Analysis of the allele frequencies at the DQB1 locus in the DQ2 heterozygous patient and control populations revealed that there is an significant frequency difference of the DQB1*0602 allele between groups. Analysis of the low risk (non DQB1*0201-carrying) haplotypes in these individuals revealed that there is also a significant frequency difference in the DQB1*0602-carrying haplotypes which extended into the MHC class III region between groups. Our analysis has shown that linkage disequilibrium to the DQB1*0602 locus can account for some of the original disease associations observed with SNPs in the total DQ2 heterozygous population. Here, we

have divided the low risk haplotypes into two group; DQB1*0602-carrying haplotypes and non-DQB1*0602 carrying haplotypes. The majority of SNP allele frequency differences originally observed in the total DQ2 heterozygous populations were also identified between the DQB1*0602-carrying haplotypes in these groups. Non-DQB1*0602 carrying haplotypes displayed similar allele frequency differences between patient and control groups. Therefore, our results suggest that the DQB1*0602-carrying haplotype, or one or more of the variants present on this haplotype, may be associated with CD. Dissection of the exact causal variant is complicated by the high levels of LD in the region.

In fact, a number of disease associations which were observed with SNP loci, were attributed to variations in allele frequencies on the low risk haplotype. Some of these associations were attributed to linkage to the DQB1*0602 allele at the DQB1 locus which was found to differ significantly between patient and control groups. Disease associations at some other SNP loci could not be attributed to the 0602 allele. However, all disease associations at these loci were as a result of the low risk haplotypes. These disease associations with potentially functional SNPs were found in a number of genes including *IkBL*, *BAT1*, *TNF*, *BAT2* and *BAT4*. All these genes have immune-related functions which make them plausible candidates in disease pathogenesis. This study highlights these SNPs as potential targets for further research. Further genetic studies on larger independent populations will need to be carried out to confirm these findings. If these findings are replicated, functional studies will need to be carried out to identify the exact causal variant or variants present.

The next analysis carried out in this study, aimed to provide a situation where the presence of the high risk DQB1*0201 allele was completely controlled for. Analysis was carried out to examine if there were additional risk factors on the high risk haplotype (DQB1*0201-carrying haplotype) which cause disease susceptibility independently of the DQ2 association. To do this, allele frequencies present on the high risk haplotype were compared between patient and control groups. Using this method, significant frequency differences were observed at the MICB-7595 and MICBrs31314900 loci, with frequency differentials of 12.9% ($p = 0.0162$). The *MIC* gene region has been previously implicated as an additional independent susceptibility locus, as has been discussed earlier. This lends support to the validity

of this method for discovering additional susceptibility loci. It should be noted, however, that the population samples used in this study are small. This is due to the high cost of HLA typing and the inevitable discrepancy between the numbers of patients and controls which fall into each DQ2 subclass (DQ2 homozygous, DQ2 heterozygous and DQ2 negative). Therefore, in order to achieve adequate power, much greater numbers of individuals would need to be typed for the HLA-DQ loci. Alternatively, these results should be replicated in an independent population to verify our findings.

Further studies wished to examine if the polymorphisms found previously to be at elevated frequencies on the high risk haplotypes of DQ2 heterozygous patients were similarly increased on the high risk haplotypes of DQ2 homozygous patients. This implicitly addresses the issue of whether there were any differences between the high risk haplotype in DQ2 heterozygous, compared with DQ2 homozygous, individuals, or whether these are in fact the same. This was also to establish if there were any additional factors differing between the high risk haplotypes of both groups which account for the ability of just one copy of the high risk haplotype to cause disease in some individuals. This was an exploratory investigation, as it is clear that the presence of two high risk haplotypes is superfluous in some individuals, who develop disease despite having only one DQB1*0201 carrying haplotype. However, the findings from these studies were interesting. Again, both MICBrs3134900 and MICB-7595 showed significant differences between the DQ2 heterozygous and DQ2 homozygous patient groups. It is noteworthy that DQB1*0201 homozygous individuals had allele frequencies at these loci similar to those seen in DQB1*0201 heterozygous control individuals. Though the reasons for this may be speculative, it is possible that the presence of two high risk alleles at HLA-DQB1, overcomes the need for an additional disease susceptibility locus in the nearby region. Hence the additional risk factor is only selected for in patients who have one copy of the HLA-DQB1*0201 allele. This leaves DQB1*0201 homozygous patients with allele frequencies at these loci similar to those seen in the general population (represented here by the DQB1*0201 heterozygous healthy controls).

To further investigate the possible reasons for the allele frequency differences observed between the DQ2 heterozygous and DQ2 homozygous patients, the extended haplotype structure and associated frequencies were examined in both

groups. This investigation found only one high risk (i.e. DQB1*0201 carrying) haplotype to be present at a frequency greater than 5% across the region of interest. This result is not unexpected, as it has previously been shown that there is a high level of LD, and hence haplotype conservation, in this region. The DQ2 heterozygous patient group experienced a significantly higher frequency of this conserved haplotype when compared to the DQ2 homozygous group (89.2% versus 74.4%, respectively). This haplotype contained the allelic variants of the SNPs which were shown previously in this study to be elevated in the DQ2 heterozygous group (SNPs MICB-7595 and MICBrs3134900). However, it was then noted that the number of DQ2 homozygous individuals who possessed one or more copies of the conserved high risk allele was comparable to the number of DQ2 heterozygous individuals carrying the haplotype (86.6% versus 89.2%, respectively). These results imply that, even in homozygotes, one copy of the high risk extended haplotype is useful for CD development. This is suggested by the fact, that in DQ2 homozygous patients, one copy of the conserved haplotype is present at the rate seen in heterozygous patients, but the "other" DQB1*0201-carrying haplotype in these individuals is not conserved at such a high frequency and experiences lower levels of linkage disequilibrium. This suggests the need for conservation of only one of the DQB1*0201 allele-carrying extended haplotypes in CD development.

An interesting point to be noted from this analysis, is that the ability to detect a disease association at the MICB SNPs would have been diluted and the association would most probably have been missed, if all DQB1*0201-carrying haplotypes from both DQ2 heterozygous and DQ2 homozygous patients had been grouped together and the frequency of the MHC class III SNPs on these haplotypes had been compared with those on similar haplotypes in the control population. This is because the frequency of the associated MICB SNPs are similar between DQ2 homozygous patients and DQ2 heterozygous controls.

The observation of the differences in the frequency of the conserved high risk extended haplotype between homozygotes and heterozygotes is interesting. So too is the finding of significant differences at the MICB SNPs on the high risk haplotype between DQ2 heterozygous patients and controls, and between DQ2 heterozygous and homozygous patients. However, the practical implications or outcome of these allele frequency differences remains unclear. This is in light of the fact that these

SNPs do not show significant differences when the total DQ2 heterozygous patient population was compared against the DQ2 heterozygous control population. In this situation, the allele frequencies on the low risk haplotype seem to counteract the significant allele frequency differences found on the high risk haplotype, leading to an overall situation of similar allele frequencies between patients and control groups.

The method used above to identify any allele frequency differences between varying groups on the high risk or low risk haplotype should be viewed as an entirely exploratory method. As this method is based initially on the computational reconstruction of haplotypes from unphased genotype data, there will always be an intrinsic level of error associated with the haplotype frequencies generated. Because haplotype frequencies are used to calculate all allele frequencies in the populations of interest, the error rates associated with the haplotype frequency estimation will be passed onto the allele frequency estimation. In the present study, we have not accounted for this error. Here we are simply looking at crude allele frequency differences to identify any loci which display large and significant frequency differences between groups, thus identifying such loci as potential disease susceptibility loci, and targets for further research. The statistics and p-values used in these particular analyses should be viewed only as a guide to direct further research. To further address the weakness apparent in the present methodology used, the development of appropriate software needs to be introduced to account for the error rates associated with looking at allele frequencies on haplotypes generated from computational data. The generation of such software has been discussed with Hitachi Dublin Laboratories, and may lead to new software development to allow for accurate assessments of the currently unquantifiable levels of error associated with this approach. Another means of minimising error associated with this approach is to again conduct these analyses in larger study populations.

In conclusion, in this study of the MHC class III region and *MIC* genes, we have found evidence for DQ2-independent additional risk factors at *MICB* which is specific to the DR3-DQ2 haplotype. It was noted that only one copy of the extended high risk (DQB1*0201-carrying) haplotype is necessary for disease development. Due to practical reasons such as cost restraints, this was the only CD associated haplotype to be examined in this study, and this is by far the most common haplotype in the Irish population. This study also identified an increase in the frequency of

DQB1*0602 in the DQ2 heterozygous patient population when compared to the DQ2 heterozygous control population. This may suggest that the DQB1*0602 allele or, perhaps more likely, an allele in LD with DQB1*0602, may be be an additional risk factor. A number of MHC class III loci were also found to be associated with disease in the DQ2 heterozygous population, and unlike the MICB associations, these findings are most likely attributed to linkage on a low risk haplotype. Functional studies characterising the effects of these genetic variants on gene regulation, expression and function would need to be carried out to identify the exact causal variant, or indeed if combinations of SNPs in this region are required. All SNPs examined in this study have potential functional effects on the immune-related genes in which they occur, which experimental evidence already available outlining the effects of some SNPs such as LTA+252 and IkBL-62 on gene regulation (Ozaki *et al.* 2002). Replication of these findings in larger and or independent population samples is required to confirm these results.

# CHAPTER VI

## GENERAL DISCUSSION

### 6.1 Coeliac disease and chromosomal region 2q33

The aim of the research presented in this study is to add to the current knowledge and understanding regarding the underlying genetic causes of coeliac disease. Though it has long been established through familial and twin studies that coeliac disease is indeed a genetic disorder, with the exception of the HLA-DQ locus, the identification of the actual genes responsible have, so far, eluded researchers. Confounding factors such as inadequate study design and lack of knowledge regarding the polymorphic content and the behaviour of the genome have all resulted in the inability of studies to conclusively identify susceptibility genes. This study has attempted to implement the recent advances, both in knowledge and technology, to investigate for the presence of coeliac disease susceptibility loci.

To date, the only non-HLA gene to be shown in a number of independent populations to be associated with coeliac disease is the *CTLA4* gene on chromosome 2q33 (Djilali-Saiah *et al.* 1998; Holopainen *et al.* 1999; Naluai *et al.* 2000; King *et al.* 2002; Popat *et al.* 2002; Popat *et al.* 2002; Mora *et al.* 2003; Amundsen *et al.* 2004; Holopainen *et al.* 2004; van Belzen *et al.* 2004; Hunt *et al.* 2005). However, the association of this gene with CD is not above controversy, with some studies failing to provide any evidence of association with disease (Clot *et al.* 1999; Wijmenga *et al.* 2000; King *et al.* 2003; Martin-Pagola *et al.* 2003; Haimila *et al.* 2004). Not only have results from studies on the *CTLA4* gene been inconsistent, so too have study designs. The majority of studies have focused on only one or a small number of individual SNPs, paying no attention to other variation present, the possibility of linkage between markers and the haplotype structure present in a given region. These are all factors which we now know from recent developments in genetic mapping are highly important to consider when constructing and carrying out genetic association testing. Indeed, many of the above studies of the 2q33 region failed to include analysis of the neighbouring genes, *CD28* and *ICOS*, both of which play important roles in the same immune regulatory pathway as the *CTLA4* gene. This present study is among the first to collectively investigate the contribution of these three genes to the pathogenesis of coeliac disease in a single population. It is also among the first to utilise the information available at the present time on the genetic variation and haplotype structure of the region, making it easier to cross-compare these results with those from other populations. We have shown that there is an association between a number of individual SNPs in the region with disease

status, but perhaps, more interestingly, the highest frequency differential observed between patient and control groups was identified with an extended haplotype which stretched from the *CD28* gene, through the *CTLA4* gene and to the 5' end of the *ICOS* gene. This haplotype was constructed to mimic the linkage disequilibrium block structure of the region, and this finding highlights the major importance of considering and utilising the patterns of linkage disequilibrium present in the genomic regions of interest. Although, in past studies, the phenomenon of haplotype block patterns in the human genome was not well understood and hence, not utilised, recent genetic association studies have begun to incorporate this information in the methodologies used to examine their data. Our investigation of the 2q33 region concluded that there was a significant association between this region and coeliac disease status. It is most likely however, that this association is not due solely to any of the genotyped markers used in this study, which include the functional variants CTLA4+49 and CTLA4_CT60, both of which have been shown in previous studies to be associated with disease. It appears more likely to be due either to an un-genotyped marker which is on or in LD with the extended haplotype of the region, which is shown to have a higher frequency differential between groups than any individual polymorphism analysed, or alternatively, a combination of SNP alleles on this haplotype.

A major objective of this project was to incorporate and analyse as much genetic variation as possible in the region of interest in a cost effective manner. The main methodological approach undertaken to obtain this goal was the use of haplotype tagging SNPs (htSNPs). This method has been used extensively throughout this study, both in the 2q33 region and also in additional genomic regions (11q23 and the *TNFSF14* gene). Using this method, we can obtain allelic information from non-genotyped SNPs, based on the ability of the htSNPs to tag the haplotypes on which these non-genotyped SNPs occur, hence extracting the information from these SNPs.

At the present time, with the recent availability of large amounts of sequencing and genomic variation data, the use of cost-reduction techniques such as the use of htSNPs is essential as it is not feasible for association studies to incorporate and genotype all genetic variation which has been identified in the genome. It should be noted, of course, that a method such as htSNP usage does fundamentally contain some margin for error. In the present study, only haplotypes with a frequency greater

than 4 - 5% were incorporated in the method used to infer non-genotyped SNP frequencies. Haplotypes below this level were ignored, therefore introducing a certain level of error in allele frequency estimation. It must also be noted that haplotypes reconstructed computationally from unphased genotype data, as is the case in this study, are estimated frequencies and have associated error rates. These error rates will, in turn, also be present in the allele frequency estimations of non-genotyped SNPs. Hence, while the use of htSNPs is an extremely useful method of genetic variation analysis, for identifying non-genotyped SNPs which may have significant allele frequency differentials between patient and control groups, these results should be viewed bearing in mind the estimates of error involved in the calculations. This method provides a guide to identifying disease associated SNPs, which can be further selected for genotyping to confirm or reject any disease association.

### 6.2 Coeliac disease and chromosomal region 11q23

The general approach taken to identify candidate genes for analysis, was to combine information from a number of sources. These included information from past genomic scans which used strictly genetic information to identify regions linked to disease. Information on the function of the genes located in these areas was evaluated, in an effort to select genes which might have a function in an immune-related inflammatory disease, such as coeliac disease. While both genome scans and candidate gene identification approaches are individually both good strategic methods for association studies, with approximately 18,000 – 20,000 genes in the human genome, the utilisation of information from both methods is clearly a better strategic approach for choosing candidate genes to analyse, improving the chances of detecting an actual susceptibility gene and reducing the chances of false positive results. Using this combined approach in this study, we have investigated possible candidate genes with immune-related functions in chromosomal region 11q23, which has been identified in a number of independent studies as linked to disease (Greco *et al.* 1998; Holopainen *et al.* 2001; Naluai *et al.* 2001). This approach identified six candidate genes for analysis; *CD3E*, *CD3D*, *CD3G*, *IL10RA*, *THY1* and *IL18*. Markers were again selected in this region to maximise the information return while minimising the numbers of markers used, an important consideration for cost-effectiveness in large scale study designs as discussed above. Markers were again selected based on their ability to tag haplotypes (and hence, non-genotyped variation)

and also for their possible functional effects on protein structure or gene regulation. From this research, we have concluded that there was a tentative association between the *IL18* gene and coeliac disease. Interestingly, two polymorphisms found to be associated with disease were the promoter polymorphisms -137 and -607, both of which have been implicated in the regulation of gene expression of *IL18* (Gierdraitis *et al.* 2001). The *IL18* gene, with a role as a proinflammatory cytokine, is a good candidate gene for involvement in the pathogenesis of coeliac disease, and this gene has already shown association with other autoimmune diseases such as type 1 diabetes (Kretowski *et al.* 2002), hence indicating that it can act as a potential pathogenic agent in such conditions. This study, however, provides the first evidence to suggest that there is a genetic association between *IL18* and CD.

### 6.3 Coeliac disease and the *TNFSF14* gene

Another gene to be included as a candidate gene in this present study was the *TNFSF14* (*LIGHT*) gene. This gene, located at position 19p13.3, is located outside a region previously shown to be linked to disease (19p13.1) (Van Belzen *et al.* 2003). Though this gene does not fall directly within the linkage region, recent studies have provided evidence to show that the *TNFSF14* gene plays an important role in inflammation of the gut. Mouse model experiments have shown that constitutive T cell expression of *TNFSF14* results in lymphocytic activation, mucosal inflammation and tissue damage (Shaikh *et al.* 2001; Wang *et al.* 2001). Also, elevated constitutive *TNFSF14* expression on human intestinal T cells and the requirement of *LIGHT* for interferon-gamma production have been shown (Cohavy *et al.* 2005), suggesting an intestinal mucosa-specific *TNFSF14* mediated inflammation mechanism. This would suggest that *TNFSF14* is an excellent candidate gene for study in a disease characterised by such intestinal mucosa-specific inflammation. Indeed, here we present evidence to suggest that *TNFSF14* is associated with CD. Three polymorphisms from the five selected for analysis in this gene have been shown to be significantly associated with disease before correction. After correction for multiple testing for all SNPs (n = 20) analysed on chromosomes 11 and 19 using even the most stringent method available, the Bonferroni method (corrected p-value = 0.05/20 = 0.0025), one of these SNPs remains significantly associated with disease (TNFSF14_rs2279627 G allele carrier status; p-value = 0.0023), and another displays borderline significant association (TNFSF14_ht3_rs11878563 G allele carrier status; p-value = 0.0027). Hence these results increase our confidence that this represents a

real disease association with this gene. Of course, replication of these findings in an independent population is desirable to confirm these results. These findings provide a useful addition to the present body of knowledge available on the genetics of coeliac disease, as the specific immune regulatory function of this gene in humans is a relatively new finding. In light of its functional effects, and the identified genetic association with CD presented in this study, *TNFSF14* appears a most promising coeliac disease susceptibility candidate gene at the present time.

**6.4 Gene-gene interaction effects in coeliac disease susceptibility genes**

Another area of study within this project investigated whether there was a gene-gene interaction effect between those genes found to be individually associated with disease. Disease associations were found at individual SNPs within *CTLA4* (at 2q33), *IL18* (at 11q23) and the *TNFSF14* gene (at 19p13.3). It was postulated that, as each of these genes provides evidence of individually attributing to disease, interactions between them may be important in promoting disease. Crosstabulation and logistic regression analysis methods were employed to investigate this hypothesis.

In some of the regions of interest (2q33, 11q23 and the *TNFSF14* gene), more than one SNP was found to be associated with disease. However, using the methods employed in this study, it is not possible to investigate gene-gene interaction effects between variables which are not independent of each other. As all variants in each region experience some level, and in the majority of cases, high levels of linkage disequilibrium with each other, they are clearly not independent from each other and hence, such an investigation cannot be carried out between genes or SNPs within the same genomic region. Thus gene-gene interaction analysis may only be carried out between variants in unlinked genomic regions, for example on different chromosomes.

Gene-gene interaction analysis did not reveal any significant disease effect when analysing all combinations of interactions between the three SNPs selected in the regions of interest. Therefore these three variants of the genes of interest do not show evidence of interacting to increase disease susceptibility. Of course, the possible number of combinations of SNPs selected from the three regions of interest is much larger than that selected here. However, analysing all combinations of SNPs would be extremely labour intensive, and not at all feasible under the present conditions and

using the software available. A program, with the ability to select all combinations of SNPs (one from each selected area) needs to be made available for this type of analysis to be undertaken, especially for large datasets. In this study, we have chosen to assess those SNPs which individually display the highest level of association with disease in each region for gene-gene interaction analysis to increase the probability of detecting interaction effects. However, as stated, no such effects were detected in this case.

### 6.5 Coeliac disease and the MHC class III region/*MIC* genes

Previous studies on the genetics of coeliac disease have also focused on the possibility of an additional disease susceptibility locus or loci in the MHC region, which is independent of the already identified DQ2 association. Studies on this region have to ensure that they are controlled for the known and relatively strong association between the DQ2 locus and disease when testing for association at other neighbouring loci. Extended haplotypes and high levels of linkage disequilibrium (LD) in the region would most probably lead to spurious positive associations with disease at tested markers, due entirely to linkage disequilibrium with the DQ2 locus, if populations tested are not controlled for this locus. Studies have implemented various methods to control for the DQ2 locus, and these involve matching cases and controls for alleles at the DQB1 locus. This is the method which has been employed in this study. A unique and interesting facet of the present study is that all markers selected for analysis have the potential to be the actual disease susceptibility locus. The high levels of linkage disequilibrium and the presence of the disease risk allele at DQB1 make the use of positional markers less desirable than the use of SNPs which may in fact contribute to disease because of a potential functional role. The use of only markers with a potential functional effect is arguably a more cost effective and potentially rewarding means of proceeding in the analysis of the MHC region for additional risk loci for diseases with known associations to HLA class I or II loci such as coeliac disease. Unless we have this evidence of a potential functional effect for a marker, it would not be unreasonable to assume that any identified disease association is simply the result of LD to another marker. This assumption is particularly valid in the MHC region, as this is one of the most gene dense and polymorphic regions of the genome, which also experiences high levels of extended LD between markers.

However, even with the careful selection and analysis of only potentially functional SNPs, there are a number of cautionary factors which need to be considered. One must recognise that there may still be linkage disequilibrium between functional markers, thus identification of the actual disease susceptibility factor may require functional testing. Furthermore, it is possible that markers designated as non-functional may in fact have unknown functions. Such SNPs would be excluded from analysis using the present approach. Similarly, SNPs which are designated as potentially functional (i.e. non-synonymous SNPs or SNPs in transcription factor binding sites) may actually have no functional effect. However, with the knowledge available to date on the function and architecture of genetic variation in the human genome, the approach undertaken for SNP selection in this region is a good approach to increase the probability of finding the actual disease susceptibility locus early in the detection process and to do so in a cost effective manner.

An interesting finding we observed in the genetic association analysis of the MHC region in this study is the presence of additional risk factors in the *MICB* gene (SNPs MICBrs3134900 and MICB-7595). These additional MHC risk factors were most evident on the DR3-DQ2 haplotype in DQ2 heterozygous patients. These same risk factors in the *MICB* gene were not identified in DQ2 homozygous individuals, who had allele frequencies similar to those seen in DQ2 heterozygous controls.

These findings are similar to previous studies which also identified a DQ2 independent risk factor on the DR3-DQ2 haplotype in the *MIC* genes (mainly the MICA-A5.1 microsatellite) as previously discussed. Unique to this study however is the finding that the increased frequency of the additional risk factors at the *MIC* genes is observed on the DR3-DQ2 haplotypes of DQ2 heterozygous patients. Previous studies identified to date have failed to compare the allele frequencies of such SNPs between DQ2 heterozygous patients and DQ2 homozygous patients, hence potentially overlooking an important facet of the data.

This finding initiated further interesting investigations to assess what was the underlying cause of the difference in allele frequencies between DQ2 heterozygous and DQ2 homozygous patient groups. Analysis of the haplotype frequencies of the extended high risk (DQB1*0201-carrying) haplotype revealed the presence of one major extended haplotype with a frequency greater that 5%, as was expected in a

region with such high levels of extended LD. This haplotype was found to occur at a significantly higher frequency in the DQ2 heterozygous patient population when compared to the DQ2 homozygous patient population. However, it was of interest to note that the proportion of DQ2 homozygous patients with one or more copies of the conserved high risk extended haplotype was similar to that seen in the DQ2 heterozygous patient group. These results suggest that in DQ2 homozygous patients who carry one copy of the conserved high risk extended haplotype, the "other" DQB1*0201-carrying haplotype tends to be less conserved This suggests that one copy of the conserved high risk extended haplotype may predispose to CD development. This may imply or suggest that any additional MHC class III or MIC region disease susceptibility variants may need to be present in only one copy (i.e. act in a dominant manner) for disease to occur. All other high risk haplotypes in the region had a frequency of approximately 1% or slightly higher, and were very numerous, hence no reasonable examination of the contribution of these haplotypes to disease pathogenesis could be carried out.

The *MIC* genes are expressed on enterocytes under stressful conditions and are ligands for intraepithelial γδ T cells which are preferentially located in intestinal lymphatic tissue and are known to be greatly expanded in coeliac disease (Groh *et al.* 1998). These combined results are interesting as there is now both genetic and functional evidence to suggest that the *MIC* genes play a role in disease susceptibility. However, attempts to locate the actual causal variant within the *MIC* genes may be difficult due to the high level of LD which exists between the polymorphisms within the gene. The three polymorphisms examined in the present study were all found to be in relatively high LD (D'>0.71). To identify the actual disease risk variant or variants, potentially functional disease associated alleles identified through genetic analysis, will need to be tested experimentally to identify if these variants do actually cause a change in the function of the gene which may possibly assist in the pathogenesis of disease.

Another interesting finding in the study of this region is the significantly increased frequency of the DQB1*0602 allele in DQ2 heterozygous patients compared to DQ2 heterozygous controls. This finding has not to our knowledge been reported previously for coeliac disease, and requires replication in additional independent populations. In addition to this finding, an extended conserved haplotype carrying the

DQB1*0602 allele was found to be significantly more frequent in DQ2 heterozygous patients compared to controls. We have also shown that allele frequencies at a number of additional loci in the MHC class III region were significantly different as a result of the presence of the differing DQB1*0602-carrying extended haplotype frequencies. If these observed disease associations do indeed represent a real disease susceptibility locus or loci, the determination remains to be made whether the actual association lies at DQB1*0602 or with a nearby locus.

### 6.6 Future Prospects

The results presented in this study highlight many future lines of investigation which may be followed. We are at present in an era where genetic advances are progressing at astonishing rates both in the knowledge obtained regarding the structure of the human genome, including the variation present and the interaction and linkage between these variants, and also in the technological advances available at present to study this variation and behaviour. In addition to this, coeliac disease is, today, probably, one of the most well-understood autoimmune diseases, with the major environmental trigger (gluten), and the major genetic contribution (the HLA-DQ2 locus) well defined. The combination of these factors would suggest that the ability to locate the remaining genetic components involved in CD pathogenesis is within reach.

Continued recruitment of patients at St. James' hospital and at other centres across the country will be carried out to increase the sample size of the population under investigation, thus increasing the power of future studies. Though recruitment of a large number of patients is a laborious task, it is imperative to the success of future genetic association studies. To date, the most successful identification of risk factors for various diseases have almost all identified those risk factors with high odds ratios. As the task to identify the remaining disease risk factors which will most likely have small effect sizes on disease intensifies, patient and control sample sizes need to be increased substantially from those used in current studies to provide the power necessary to detect these small changes. This will only be achieved after years of patient recruitment and possible collaborations between recruitment centres. The establishment of national biobank facilities, such as that proposed and currently being implemented in Ireland, is also an important and perhaps, necessary advancement for genetic association studies. Such projects will provide a pool of

healthy control individuals sampled randomly throughout the country, from which to draw a control population for analysis. This method is most likely to be better than the most popular currently used method which implements the construction of a healthy control sample population for analysis from colleagues and health workers in hospitals, which usually tend to be biased towards selection of females and there may also be other unseen biases.

This study has provided some tentative positive associations between certain genetic variants and CD. Among these positive disease associations are some genetic associations which have been identified for the first time in this study. *IL18* has been shown to be significantly associated with disease. There is good evidence, based on the function of this gene as a proinflammatory cytokine, and previous evidence showing that the polymorphisms studied and found to be associated with disease in this study are actual functional variants which influence gene expression, that these positive disease associations are actually not the result of false positive findings. To confirm such results, these polymorphisms will need to be replicated in an independent population sample through collaborative works. It will also, in the future, be possible to re-assess this positive finding in a larger Irish patient sample after the continued recruitment of patients has been completed.

The observation of a positive association between the *TNFSF14* gene and CD is also of major interest, as this is the first genetic association study to be carried out between any autoimmune disease and this gene. This may be the result of the relative recent discovery of the importance of this gene in human intestinal mucosa-specific inflammation. The positive findings in this study will most likely direct further genetic association studies to confirm these results in other populations. As, to date, this gene has not been the subject of much interest in genetic association studies, little effort has been expended on sequencing and identification of all genetic variation present in the region, and at the present time, relatively little information is available from projects such as the HapMap project on *TNFSF14*, when compared to other well established inflammatory genes such as *IL18*. To our knowledge, all informative and potentially interesting genotype and haplotype information available at the present time has been included in this study. However, with additional releases of genetic variation information from the HapMap consortium and other sources, it is quite conceivable that information on extra variation in the gene will become

available for researchers to analyse. From the data presented in this study, it is envisaged that *TNFSF14* is a candidate gene worthy of further investigation. For example, since little is known about variation within this gene resequencing of the gene in patients and controls is an obvious starting point to help define its relevance to CD genetics.

The positive finding of disease association with a risk factor in *MICB*, independent of the established DQ2 association, confirms previous studies showing association in these *MIC* genes. Though this study has attempted to assess as much of the potentially functional variation present in the MHC class III and *MIC* genes, all the variation assessed here is only that which is known to differ between the AH 8.1 and AH 7.1 extended haplotypes. This is due to the fact that at the present time, sequence information is only available for these two haplotypes from The MHC Haplotype Project (ttp://www.sanger.ac.uk/HGP/Chr6/MHC/). However, this project is currently in the process of sequencing and releasing information on the other haplotypes present in the region. Upon release of such information, it will be possible to analyse the additional variation in the region. The imminent release of new technologies such as the MHC genotyping micro-array panel (made by Illumina, CA, U.S.A.), which allows genotyping of many variants in one experiment by the use of a microarray-based technology, will increase the ability of researchers, in the near future, to analyse very large numbers of genetic variants in a region of interest.

The future of genetic association studies to identify coeliac disease susceptibility genes looks promising. This study has identified a number of interesting candidates which may provide the direction or foundation for future studies. Improvements in the fields of genetics and coeliac disease research will inevitably lead to the identification of, as yet unknown, coeliac disease genes.

## Appendix I

## List of SNP or haplotype related databases

A comprehensive list of SNP related database (http://hgvbase.cgb.ki.se/databases.htm)

Asthma pharmacogenetics website (http://www.pharmgat.org/index_html)

Genetic association database (http://geneticassociationdb.nih.gov/)

Perlegen database (http://genome.perlegen.com/)

Seattle SNPs variation resource database (http://pga.gs.washington.edu/)

Southwestern program for genomic applications (http://pga.swmed.edu/)

Innate Immunity (http://innateimmunity.net/)

The HapMap project website (http://www.hapmap.org/)

NCBI gene genotype report (http://www.ncbi.nlm.nih.gov/)

## Appendix II

**Table A2.2 Primer sequences for genotype assays**

| SNP | Forward primer | Reverse primer |
|---|---|---|
| CTLA4-658(rs11571317) | gcttttctttggaccttcttca | tcacaagaaataaactgaaaatagcc |
| CTLA4_CT60(rs3087243) | tttcttcaccactatttgggatatacc | tagttgagctgagaaagcaggc |
| CTLA4+49(rs231775) | gattctgtgtgggttcaaac | ttgcagaagacagggatg |
| IL10RA(rs2229113) | caggctgacagaacgctgggaacc | tgggccctgtgctggggc |
| ICOS+1323(rs4270326) | caacagggagacaattccttccctc | ctcttgacctcaatttcctcatctagtttg |
| IkBL-62(rs2071592) | agagccgtaggcccaaggcc | cacagttcacttccgtcctccagc |
| CTLA4-1661(rs4553808) | cttgagggcaggaacatttgtttt | gccagactgggcaacaga |
| CTLA4-1722(rs733618) | caattccagcattgatctcactctatca | gccctcaagaggaacacaca |

Nucleotides which are underlined represented mismatched nucleotides as created by the PIRA program (Ke *et al.* 2001).

**Table A2.2 Primer sequences used for SNP genotyping using Amplifluor**

| SNP | Allele1 | Allele2 | Common |
|---|---|---|---|
| CD28 (rs3181113) | ggagtaggaatcttgagattccag | gaggagtaggaatcttgagattccat | ggaagcccacttaaaaatcaaccaaagta |
| CD28 (rs3769683) | ataagtctcttctcttcctcccc | ctataagtctcttctcttcctccct | gagaggaggacaaatctctaagaagat |
| CTLA4+49 (rs231775) | aaggctcagctgaacctggcta | ggctcagctgaacctggctg | Agagtgcagggccaggtcct |
| ICOS-2394 (rs11571306) | tttcaatgatccttcccatgcctca | tttcaatgatccttcccatgcctcg | Cacaagtacttgtattcttgaccaagattt |
| ICOS-2119 (rs11889031) | ggaaggatacttaggtcctagag | gtggaaggatacttaggtcctagaa | Accctccaattgaccaggcatcttt |
| ICOS+1323 (rs4270326) | gtgattagtttaattagtaggtatgtcttg | gtgattagtttaattagtaggtatgtcttc | gaccaacagggagacaattcctt |
| ICOS+2003(novel (Haaning *et al*) | atagcgcccatggtctaagtctgc | atagcgcccatggtctaagtctgt | gcctttattgtttctttccca |
| ICOS+3990 (rs10172036) | ctggcacatggagagcatttagag | ctggcacatggagagcatttagat | gggtcacatctgtaacaaaaattcagcata |
| ICOS+4031 (rs10183087) | atgcctgggtgccagagttccat | atgcctgggtgccagagttccag | tacagatgtgaccctataa |
| ICOS+5291 (rs10932038) | ggaatcacagtggtctacctgca | aatcacagtggtctacctgcg | gctcttcacagatcctggaattatgaa |
| MICB11715 (rs1065076) | cctggctgtagagtctaggt | gcctggctgtagagtctaggc | gtcagctactgggtccactggtt |
| CSNK2B1942 (rs805256) | caagatcatgtctagagcttgtcga | aagatcatgtctagagcttgtcgg | actcaatgagcaggtccctcacta |
| CSNK2B3854 (rs4569) | tctactttatttcatattcccaccaca | ctactttatttcatattcccaccacg | ccctttcaggaaccctgtatggttt |
| DDAH2 (rs805305) | gactggaagtccagcccgg | gactggaagtccagcccgc | ccagctttctccttctgtcccataa |
| LTA+252 (rs22398322) | cacacattctctgtttctgccatga | acacattctctgtttctgccatgg | agacaggaagggaacagagaggaa |
| TNFα-308 (rs1800629) | ggaggctgaaccccgtcct | gaggctgaaccccgtccc | gaggcaataggttttgaggggcat |
| IkBL-62 (rs2071592) | agttcacttccgtcctccacct | agttcacttccgtcctccacca | caggctggaggaaatggcgcaa |
| TNFα-863 (rs4645836) | cgagtatggggacccccc | gtcgagtatggggaccccca | gccctctacatggccctgtctt |

| SNP | Allele1 | Allele2 | Common |
|---|---|---|---|
| TNFα-857 (rs1799724) | tacatggccctgtcttcg | cctctacatggccctgtcttca | gtccagggctatgaaagtcgagtat |
| IL10RA snp23 (rs2256111) | cccaagatggcccccgca | cccaagatggcccccgcg | ctcatactctcggaagtgactgaagat |
| IL10RA snp24 (rs4252272) | ctcggaagtgactgaagatgct | ctcggaagtgactgaagatgcc | gattcagctacccaggcccaagat |
| IL10RA snp42 (rs2229113) | gacagaacgctgggaaaca | ctgacagaacgctgggaaacg | tctgtgctattgctgctgccacta |
| IL10RA snp46 (rs9610) | cagcagaacctcacatccctat | cagcagaacctcacatccctac | catctgcagaataatgactgacttgtcta |
| MICB-7595 (rs1065075) | ggtctctgtgtcccaggtctt | ggtctctgtgtcccaggtctc | gacaggcagaaacgcagggcaa |
| BAT1-1167 (rs2844509) | cgtttggccctctccctgttaa | gtttggccctctccctgttag | gagaataacggtacctactccttagtatt |
| IkBL738 (rs3130062) | agcacgtgggggtcgacg | cagcacgtgggggtcgaca | cagcagcgagaagcagagggat |
| LTA-379 (rs2239704) | cagagggcaggacactgcg | gcagagggcaggacactgct | cccgtgcttcgtgctttggacta |
| BAT2 (rs3130068) | ggagctaagtagatcatggttgctta | gagctaagtagatcatggttgcttg | ccagtctatgtaaactgcccactctt |
| CSNK2B (rs805257) | ccaagtttatcattctcaagcaagaact | caagtttatcattctcaagcaagaacc | tcgcccccaacccgtgagaaaa |
| BAT2 (rs2272593) | caaggcacaagcctgggcc | ccaaggcacaagcctgggct | gaggggccctgaggggctt |
| BAT2 (rs1046089) | attggcacagaacgatcacagca | ggcacagaacgatcacagcg | caggtcgatgggatggccgaat |
| BAT2 (rs3132453) | gaaactggcccttgagagctac | agaaactggcccttgagagctaa | cctccctcagcactgctctct |
| BAT2 (rs10885) | cacctcctgccccacctc | ccacctcctgccccacctt | aggtgaaggcagcacccgagta |
| BAT4 (rs3130618) | caatcagggcctcatagaaagcta | aatcagggcctcatagaaagctc | gaagccagagtccaccctggat |
| MICB (rs3134900) | ggtgctgctgtcttcatgg | ctggtgctgctgtcttcatgc | gcttgcattccctccaggagatta |
| CD3E_ht1 (rs3782042) | agccaaatactcaacaccagccaaa | ccaaatactcaacaccagccaag | cagggcttctgtcagggctaaataa |
| CD3E_ht2 (rs1945764) | ggcagtgtctaattccactagaaaag | ggcagtgtctaattccactagaaaaa | cacgtgtctgtgtaataccacctgaa |

| SNP | Allele1 | Allele2 | Common |
|---|---|---|---|
| CD3E_ht2 (rs1945764) | ggcagtgtctaattccactagaaaag | ggcagtgtctaattccactagaaaaa | cctgaaaatgaaaaaaggaaacagcagtaa |
| CD3DG_ht1 (rs2276423) | gtaagtggggatagagaggctc | gtaagtggggatagagaggctg | cccttcccacgtgcaaacagcat |
| CD3DG_ht2 (rs3181261) | ggagtagtgcaggaagctgtgt | gagtagtgcaggaagctgtgc | gtacctgaatctgggactggccat |
| CD3DG_ht3 (rs3181259) | gtgacagtctaactgcagcagc | aagtgacagtctaactgcagcagt | tggcctacgaggatctgaaactctt |
| CD3DG_ht4 (rs1561966) | cctcctaccattatgtacccaatg | tcctcctaccattatgtacccaata | gcaagaagcaccaactcagtctctt |
| CD3DG_ht5 (rs7949185) | gagtccctcctaaattcattacctac | agagtccctcctaaattcattacctat | ttaggatcgtagccagttgttacttgtt |
| THY1_ht3 (rs1894006) | gatggctggttctgtgagacca | atggctggttctgtgagaccg | caatgtcttagctgggaccaggaa |
| THY1_ht7 (rs1001205) | cactgctctgcagataggctaga | actgctctgcagataggctagg | cagccttggctgaaccaaccctt |
| TNFSF14_ht1 (rs2291668) | gttgctgatgggggcc | gctgttgctgatgggggct | agctgcaggaggaaccagcctt |
| TNFSF14_ht2 (rs3760746) | gccagagagacaagaaacctcg | agccagagagacaagaaacctca | caacagagcgagactccatctcaaa |
| TNFSF14_ht3 (rs11878563) | cacatccactcctttattccactctt | acatccactcctttattccactctc | gtgttttgatgtgacaagtgcagtgaaaa |
| TNFSF14 (rs344560) | cagcacacggacgaccacctt | gcacacggacgaccacctc | cagcttcctgggtggtgtggta |
| TNFSF14 (rs2279627) | atgagtattaattaccaagtgtgacgg | atgagtattaattaccaagtgtgacgc | acagccactcccagaggtgtca |

Allele1 common 5' tail: gaaggtgaccaagttcatgct; Allele 2 common 5' tail: gaaggtcggagtcaacggatt

## Appendix III

## Genes identified on 11q23.3 between and 100 kb either side of markers D11S4111 and D11S4464 as identified by Human ContigView in Ensembl

| Chromosome Name | Start Position (bp) | Band | External Gene ID | Biotype |
|---|---|---|---|---|
| 11 | 116088932 | q23.3 | XP_499561.1 | protein_coding |
| 11 | 116092106 | q23.3 | | protein_coding |
| 11 | 116124103 | q23.3 | NP_116114.1 | protein_coding |
| 11 | 116154487 | q23.3 | ZNF259 | protein_coding |
| 11 | 116165297 | q23.3 | APOA5 | protein_coding |
| 11 | 116196628 | q23.3 | | protein_coding |
| 11 | 116205834 | q23.3 | APOC3 | protein_coding |
| 11 | 116211679 | q23.3 | APOA1 | protein_coding |
| 11 | 116219330 | q23.3 | NP_079440.2 | protein_coding |
| 11 | 116528374 | q23.3 | PAFAH1B2 | protein_coding |
| 11 | 116554858 | q23.3 | SIDT2 | protein_coding |
| 11 | 116554858 | q23.3 | SIDT2 | protein_coding |
| 11 | 116575250 | q23.3 | TAGLN | protein_coding |
| 11 | 116581004 | q23.3 | PCSK7 | protein_coding |
| 11 | 116581004 | q23.3 | PCSK7 | protein_coding |
| 11 | 116610190 | q23.3 | NP_997226.1 | protein_coding |
| 11 | 116661627 | q23.3 | BACE1 | protein_coding |
| 11 | 116661627 | q23.3 | BACE1 | protein_coding |
| 11 | 116661627 | q23.3 | BACE1 | protein_coding |
| 11 | 116714482 | q23.3 | NP_055771.3 | protein_coding |
| 11 | 116803699 | q23.3 | DSCAML1 | protein_coding |
| 11 | 117196000 | q23.3 | FXYD2 | protein_coding |
| 11 | 117196000 | q23.3 | FXYD2 | protein_coding |
| 11 | 117196000 | q23.3 | FXYD2 | protein_coding |
| 11 | 117196136 | q23.3 | | protein_coding |
| 11 | 117212903 | q23.3 | FXYD6 | protein_coding |
| 11 | 117212903 | q23.3 | FXYD6 | protein_coding |
| 11 | 117276570 | q23.3 | TMPSD_HUMAN | protein_coding |
| 11 | 117276570 | q23.3 | TMPSD_HUMAN | protein_coding |
| 11 | 117362319 | q23.3 | IL10RA | protein_coding |
| 11 | 117452976 | q23.3 | TMPRSS4 | protein_coding |
| 11 | 117452976 | q23.3 | TMPRSS4 | protein_coding |
| 11 | 117509302 | q23.3 | SCN4B | protein_coding |
| 11 | 117541414 | q23.3 | SCN2B | protein_coding |
| 11 | 117569888 | q23.3 | AMICA1 | protein_coding |
| 11 | 117569888 | q23.3 | AMICA1 | protein_coding |
| 11 | 117569888 | q23.3 | AMICA1 | protein_coding |
| 11 | 117602619 | q23.3 | NP_938016.1 | protein_coding |
| 11 | 117629347 | q23.3 | EVA1 | protein_coding |
| 11 | 117680662 | q23.3 | CD3E | protein_coding |
| 11 | 117714999 | q23.3 | CD3D | protein_coding |

| 11 | 117720311 | q23.3 | CD3G | protein_coding |
|---|---|---|---|---|
| 11 | 117735512 | q23.3 | UBE4A | protein_coding |
| 11 | 117777527 | q23.3 | ATP5L | protein_coding |
| 11 | 117810622 | q23.3 | NP_116099.1 | protein_coding |
| 11 | 117812438 | q23.3 | MLL | protein_coding |
| 11 | 117812438 | q23.3 | MLL | protein_coding |
| 11 | 117812438 | q23.3 | MLL | protein_coding |
| 11 | 117812438 | q23.3 | MLL | protein_coding |
| 11 | 117813393 | q23.3 | NP_079167.1 | protein_coding |
| 11 | 117903420 | q23.3 | XP_084672.3 | protein_coding |
| 11 | 117907127 | q23.3 | TMEM25 | protein_coding |
| 11 | 117907127 | q23.3 | TMEM25 | protein_coding |
| 11 | 117907127 | q23.3 | TMEM25 | protein_coding |
| 11 | 117907127 | q23.3 | TMEM25 | protein_coding |
| 11 | 117920471 | q23.3 | NP_064538.2 | protein_coding |
| 11 | 117920471 | q23.3 | NP_064538.2 | protein_coding |
| 11 | 117948321 | q23.3 | ARCN1 | protein_coding |
| 11 | 117948321 | q23.3 | ARCN1 | protein_coding |
| 11 | 117982423 | q23.3 | PHLDB1 | protein_coding |
| 11 | 117982423 | q23.3 | PHLDB1 | protein_coding |
| 11 | 117982423 | q23.3 | PHLDB1 | protein_coding |
| 11 | 117982423 | q23.3 | PHLDB1 | protein_coding |
| 11 | 117982423 | q23.3 | PHLDB1 | protein_coding |
| 11 | 118034209 | q23.3 | TREH | protein_coding |
| 11 | 118099161 | q23.3 | O95283_HUMAN | protein_coding |
| 11 | 118125623 | q23.3 | DDX6 | protein_coding |
| 11 | 118259777 | q23.3 | BLR1 | protein_coding |
| 11 | 118272263 | q23.3 | BCL9L | protein_coding |
| 11 | 118332236 | q23.3 | UPK2 | protein_coding |
| 11 | 118347627 | q23.3 | FOXR1 | protein_coding |
| 11 | 118374062 | q23.3 | NP_940891.1 | protein_coding |
| 11 | 118391635 | q23.3 | RPS25 | protein_coding |
| 11 | 118394451 | q23.3 | TRAPPC4 | protein_coding |
| 11 | 118400276 | q23.3 | SLC37A4 | protein_coding |
| 11 | 118400276 | q23.3 | SLC37A4 | protein_coding |
| 11 | 118400276 | q23.3 | SLC37A4 | protein_coding |
| 11 | 118420110 | q23.3 | HYOU1 | protein_coding |
| 11 | 118443745 | q23.3 | VPS11 | protein_coding |
| 11 | 118460803 | q23.3 | HMBS | protein_coding |
| 11 | 118470883 | q23.3 | H2AFX | protein_coding |
| 11 | 118472424 | q23.3 | DPAGT1 | protein_coding |
| 11 | 118472424 | q23.3 | DPAGT1 | protein_coding |
| 11 | 118472424 | q23.3 | DPAGT1 | protein_coding |
| 11 | 118483303 | q23.3 | TMEM24 | protein_coding |
| 11 | 118497498 | q23.3 | MIZF_HUMAN | protein_coding |
| 11 | 118524960 | q23.3 | ABCG4 | protein_coding |
| 11 | 118544648 | q23.3 | NP_078894.2 | protein_coding |
| 11 | 118556437 | q23.3 | PDZK2 | protein_coding |
| 11 | 118556437 | q23.3 | PDZK2 | protein_coding |
| 11 | 118582199 | q23.3 | CBL | protein_coding |
| 11 | 118678084 | q23.3 | Q8NAQ6_HUMAN | protein_coding |
| 11 | 118684196 | q23.3 | MCAM | protein_coding |
| 11 | 118710447 | q23.3 | RNF26 | protein_coding |

| | | | | |
|---|---|---|---|---|
| 11 | 118714862 | q23.3 | MFRP | protein_coding |
| 11 | 118714862 | q23.3 | MFRP | protein_coding |
| 11 | 118714862 | q23.3 | MFRP | protein_coding |
| 11 | 118714862 | q23.3 | MFRP | protein_coding |
| 11 | 118731136 | q23.3 | USP2 | protein_coding |
| 11 | 118731136 | q23.3 | USP2 | protein_coding |
| 11 | 118731136 | q23.3 | USP2 | protein_coding |
| 11 | 118794097 | q23.3 | THY1 | protein_coding |
| 11 | 118978782 | q23.3 | | protein_coding |
| 11 | 119014018 | q23.3 | PVRL1 | protein_coding |
| 11 | 119014018 | q23.3 | PVRL1 | protein_coding |
| 11 | 119014018 | q23.3 | PVRL1 | protein_coding |
| 11 | 119487205 | q23.3 | TRIM29 | protein_coding |
| 11 | 119487205 | q23.3 | TRIM29 | protein_coding |
| 11 | 119544895 | q23.3 | Q8N8C5_HUMAN | protein_coding |
| 11 | 119586957 | q23.3 | NP_848602.1 | protein_coding |
| 11 | 119616256 | q23.3 | POU2F3 | protein_coding |
| 11 | 119701226 | q23.3 | NP_777586.1 | protein_coding |
| 11 | 119712825 | q23.3 | ARHGEF12 | protein_coding |
| 11 | 119712825 | q23.3 | ARHGEF12 | protein_coding |
| 11 | 119712825 | q23.3 | ARHGEF12 | protein_coding |
| 11 | 119866486 | q23.3 | Q8NI68_HUMAN | protein_coding |
| 11 | 120036238 | q23.3 | GRIK4 | protein_coding |
| 11 | 120400035 | q23.3 | NP_689928.2 | protein_coding |
| 11 | 120478585 | q23.3 | TECTA | protein_coding |
| 11 | 120668708 | q23.3 | SC5DL | protein_coding |
| 11 | 120737769 | q23.3 | XP_208545.3 | protein_coding |
| 11 | 120828169 | q23.3 | SORL1 | protein_coding |
| 11 | 122031654 | q24.1 | NP_116262.2 | protein_coding |
| 11 | 122214465 | q24.1 | NP_062550.2 | protein_coding |
| 11 | 122258601 | q24.1 | NP_079082.2 | protein_coding |
| 11 | 122258601 | q24.1 | NP_079082.2 | protein_coding |
| 11 | 122353567 | q24.1 | XP_372433.2 | protein_coding |
| 11 | 122433411 | q24.1 | HSPA8 | protein_coding |
| 11 | 122433411 | q24.1 | HSPA8 | protein_coding |
| 11 | 122448245 | q24.1 | NP_079045.1 | protein_coding |
| 11 | 122448245 | q24.1 | NP_079045.1 | protein_coding |
| 11 | 122936247 | q24.1 | Q9ULL9_HUMAN | protein_coding |
| 11 | 122973978 | q24.1 | | protein_coding |
| 11 | 123005107 | q24.1 | SCN3B | protein_coding |
| 11 | 123100699 | q24.1 | ZNF202 | protein_coding |
| 11 | 123100699 | q24.1 | ZNF202 | protein_coding |
| 11 | 123129498 | q24.1 | OR6X1 | protein_coding |
| 11 | 123181326 | q24.1 | OR6M1 | protein_coding |

## Appendix IV

## Inference of allele frequencies at non-genotyped SNPs

**Table A3.16 Inferred allele frequencies and association testing for CTLA4-1765**

| Allele | Hap | Patient % | Pat Total n (%) | Control % | Con Total n (%) | p-value (OR(95%CI)) |
|---|---|---|---|---|---|---|
| T | 2 | 35.10 | | 35.52 | | |
| | 5 | 7.58 | | 8.52 | | |
| | | | 336(42.68) | | 371(44.04) | |
| C | 4 | 5.22 | | 5.49 | | |
| | 1 | 30.90 | | 34.88 | | |
| | 3 | 6.57 | | 6.08 | | |
| | 6 | 8.98 | | 7.35 | | 0.9374 |
| | | | 407(51.67) | | 453(53.8) | (1.01(0.82,1.24)) |

**hap = haplotype; Pat = patient; Con = control; n = number of haplotypes (chromosomes); OR(95%CI) = Odds ratio (95% confidence intervals);**

**Total number of patient and control chromosomes used to infer allele numbers are n = 790 and n = 842, respectively; Original haplotype frequencies were obtained from table 3.8. Full haplotype structure containing htSNPs and non-genotyped SNPs obtained from figure 3.2**

**Table A3.17 Inferred allele frequencies and association testing for CTLA4-1577**

| Allele | Hap | Patient % | Pat Total n (%) | Controls % | Con Total n (%) | p-value (OR(95%CI)) |
|---|---|---|---|---|---|---|
| | | | (n=788) | | (n=842) | |
| G | 2 | 35.10 | | 35.52 | | |
| | 5 | 7.58 | | 8.52 | | |
| | 3 | 6.57 | | 6.08 | | |
| | 6 | 8.98 | | 7.35 | | |
| | | | 459(58.23) | | 484(57.47) | |
| A | 4 | 5.22 | | 5.49 | | |
| | 1 | 30.90 | | 34.88 | | 0.2326 |
| | | | 285(36.12) | | 340(40.37) | 0.88(0.72,1.09) |

**hap = haplotype; Pat = patient; Con = control; n = number of haplotypes (chromosomes); OR(95%CI) = Odds ratio (95% confidence intervals);**

**Total number of patient and control chromosomes used to infer allele numbers are n = 790 and n = 842, respectively; Original haplotype frequencies were obtained from table 3.8. Full haplotype structure containing htSNPs and non-genotyped SNPs obtained from figure 3.2**

**Table A3.18 Inferred allele frequencies and association testing for CTLA4-1478**

| Allele | Hap | Patient % | Pat Total n (%) | Controls % | Con Total n (%) | p-value (OR(95%CI)) |
|---|---|---|---|---|---|---|
| | | | (n=788) | | (n=842) | |
| G | 2 | 35.10 | | 35.52 | | |
| | 5 | 7.58 | | 8.52 | | |
| | 4 | 5.22 | | 5.49 | | |
| | 1 | 30.90 | | 34.88 | | |
| | | | 621(78.80) | | 711(84.41) | |
| A | 3 | 6.57 | | 6.08 | | |
| | 6 | 8.98 | | 7.35 | | 0.1191 |
| | | | 123(15.55) | | 113(13.43) | (1.25(0.94,1.66)) |

**hap = haplotype; Pat = patient; Con = control; n = number of haplotypes (chromosomes); OR(95%CI) = Odds ratio (95% confidence intervals);**

**Total number of patient and control chromosomes used to infer allele numbers are n = 790 and n = 842, respectively; Original haplotype frequencies were obtained from table 3.8. Full haplotype structure containing htSNPs and non-genotyped SNPs obtained from figure 3.2**

**Table A3.19 Inferred allele frequencies and association testing for CTLA4+923**

| Allele | Hap | Patient % | Pat Total n (%) | Controls % | Con Total n (%) | p-value (OR(95%CI)) |
|---|---|---|---|---|---|---|
| | | | (n=788) | | (n=842) | |
| C | 2 | 35.10 | | 35.52 | | |
| | 5 | 7.58 | | 8.52 | | |
| | 4 | 5.22 | | 5.49 | | |
| | 1 | 30.90 | | 34.88 | | |
| | | | 621(78.80) | | 711(84.41) | |
| T | 3 | 6.57 | | 6.08 | | |
| | 6 | 8.98 | | 7.35 | | 0.1191 |
| | | | 123(15.55) | | 113(13.43) | 1.25(0.94,1.66) |

**hap = haplotype; Pat = patient; Con = control; n = number of haplotypes (chromosomes); OR(95%CI) = Odds ratio (95% confidence intervals);**

**Total number of patient and control chromosomes used to infer allele numbers are n = 790 and n = 842, respectively; Original haplotype frequencies were obtained from table 3.8. Full haplotype structure containing htSNPs and non-genotyped SNPs obtained from figure 3.2**

**Table A3.20 Inferred allele frequencies and association testing for CTLA4+1822**

| Allele | Hap | Patient % | Pat Total n (%) | Control % | Con Total n (%) | p-value OR (95%CI) |
|---|---|---|---|---|---|---|
| | | | (n=788) | | (n=842) | |
| T | 2 | 35.10 | | 35.52 | | |
| | 5 | 7.58 | | 8.52 | | |
| | | | 336(42.68) | | 371(44.04) | |
| C | 4 | 5.22 | | 5.49 | | |
| | 1 | 30.90 | | 34.88 | | |
| | 3 | 6.57 | | 6.08 | | |
| | 6 | 8.98 | | 7.35 | | 0.9374 |
| | | | | | | 1.01 |
| | | | 407(51.67) | | 453(53.8) | (0.82,1.24) |

**hap = haplotype; Pat = patient; Con = control; n = number of haplotypes (chromosomes); OR(95%CI) = Odds ratio (95% confidence intervals);**

**Total number of patient and control chromosomes used to infer allele numbers are n = 790 and n = 842, respectively; Original haplotype frequencies were obtained from table 3.8. Full haplotype structure containing htSNPs and non-genotyped SNPs obtained from figure 3.2**

**Table A3.21 Inferred allele frequencies and association testing for CTLA4+6249**

| Allele | Hap | Patient % | Pat Total n (%) | Contros % | Con Total n (%) | p-value OR(95%CI) |
|---|---|---|---|---|---|---|
| G | 2 | 35.10 | | 35.52 | | |
| | 5 | 7.58 | | 8.52 | | |
| | 4 | 5.22 | | 5.49 | | |
| | 1 | 30.90 | | 34.88 | | |
| | | | 621(78.80) | | 711(84.41) | |
| A | 3 | 6.57 | | 6.08 | | |
| | 6 | 8.98 | | 7.35 | | 0.1191 |
| | | | | | | 1.25 |
| | | | 123(15.55) | | 113(13.43) | (0.94,1.66) |

**hap = haplotype; Pat = patient; Con = control; n = number of haplotypes (chromosomes); OR(95%CI) = Odds ratio (95% confidence intervals);**

**Total number of patient and control chromosomes used to infer allele numbers are n = 790 and n = 842, respectively; Original haplotype frequencies were obtained from table 3.8. Full haplotype structure containing htSNPs and non-genotyped SNPs obtained from figure 3.2**

**Table A3.22 Inferred allele frequencies and association testing for ICOS+1203, ICOS+1551, ICOS+1974, ICOS+4993, ICOS+5436**

| Allele | Hap | Pat chr % | Pat Total n (%) | Con chr % | Con Total n (%) | p-value OR(95%CI) |
|---|---|---|---|---|---|---|
| C | 1&4 | 33.28 | | 34.20 | | |
| | 2 | 40.49 | | 41.90 | | |
| | | | 583(73.77) | | 539(76.10) | |
| T | 3 | 17.80 | | 15.85 | | 0.6294 |
| | | | 141(17.80) | | 122(15.85) | 0.94(0.71,1.24) |

**hap = haplotype; Pat = patient; Con = control; n = number of haplotypes (chromosomes); OR(95%CI) = Odds ratio (95% confidence intervals);**

**Total number of patient and control chromosomes used to infer allele numbers are n = 790 and n = 708, respectively; Original haplotype frequencies were obtained from table 3.12. Full haplotype structure containing htSNPs and non-genotyped SNPs obtained from figure 3.3**

**Table A3.23 Inferred allele frequencies and association testing for ICOS+2658, ICOS+3168, ICOS+3396, ICOS+4888, ICOS+5053**

| Allele | Hap | Pat chr % | Pat Total n (%) | Con chr % | Con Total n (%) | p-value OR(95%CI) |
|---|---|---|---|---|---|---|
| C | 1&4 | 33.28 | | 34.20 | | |
| | 2 | 40.49 | | 41.90 | | |
| | 3 | 17.80 | | 15.85 | | |
| | | | 723(91.57) | | 651(91.95) | |
| T | | 0 | | 0 | | |
| | | | 0(0.0) | | 0(0.0) | N/A |

**hap = haplotype; Pat = patient; Con = control; n = number of haplotypes (chromosomes); OR(95%CI) = Odds ratio (95% confidence intervals);**

**Total number of patient and control chromosomes used to infer allele numbers are n = 790 and n = 708, respectively; Original haplotype frequencies were obtained from table 3.12. Full haplotype structure containing htSNPs and non-genotyped SNPs obtained from figure 3.3**

## Appendix V

## Reagents and Buffers

**Tris Acetate EDTA (TAE) Buffer (10X)**

| | |
|---|---|
| Tris base | 24.2 g |
| Glacial Acetic Acid | 5.71 ml |
| Ethylenediaminetetraacetic acid (EDTA), pH 8.0 (0.5 M) | 20 ml |

Make up to a final volume of 1 L with deionised water.

**1X TAE Running Buffer**

| | |
|---|---|
| TAE (10X) | 50 ml |
| Deionised water | 450 ml |

**3% Agarose Gel**

| | |
|---|---|
| Agarose | 6 g |
| TAE running buffer | 200 ml |

Boil solution to dissolve agarose.

**Ethidium Bromide (EB) stock (10 mg/ml)**

| | |
|---|---|
| Ethidium Bromide | 100 mg |
| Deionised water | 10 ml |

**Gel Loading Dye**

| | |
|---|---|
| Xylene cyanol | 50 mg |
| Bromophenol Blue | 75 mg |
| Glycerol | 25 ml |
| EDTA (0.5 M) | 5 ml |
| Deionised water | 20 ml |

Mix and store at room temperature.

**200 µM dNTP mix**

| | |
|---|---|
| ATP | 20 µl |
| CTP | 20 µl |
| GTP | 20 µl |

| | |
|---|---|
| TTP | 20 µl |
| Deionised water | 920 µl |

# REFERENCES

Abecasis, G. R. (2001). "Extent and distribution of linkage disequilibrium in three genomic regions." *Am J Hum Genet* **68**: 191-197.

Abecasis, G. R., W. O. Cookson, et al. (2001). "The power to detect linkage disequilibrium with traits in selected samples." *Am J Hum Genet* **68**: 1463-1474.

Ackerman, H., S. Usen, et al. (2003). "Haplotypic analysis of the TNF locus by association efficiency and entropy." *Genome Biol* **4**: R24.

Aizawa, Y., S. Sutoh, et al. (2005). "Association of interleukin-18 gene single-nucleotide polymorphisms with susceptibility to inflammatory bowel disease." *Tissue Antigens* **65**(1): 88-92.

Akey, J. and M. Xiong (2001). "Haplotypes vs single marker linkage disequilibrium tests: what do we gain?" *Eur J Hum Genet* **9**: 291-300.

Al-Tawaty, A. I. and S. M. Elbargathy (1998). "Coeliac disease in north-eastern Libya." *Ann Trop Paediatra* **18**: 27-28.

Allcock, R. J., L. Windsor, et al. (2004). "High-Density SNP genotyping defines 17 distinct haplotypes of the TNF block in the Caucasian population: Implications for haplotype tagging." *Hum Mutat* **24**(6): 517-525.

Altmuller, J., L. J. Palmer, et al. (2001). "Genomewide scans of complex human diseases: true linkage is hard to find." *Am J Hum Genet* **69**: 936-950.

Amundsen, S. S., A. T. Naluai, et al. (2004). "Genetic analysis of the CD28/CTLA4/ICOS (CELIAC3) region in coeliac disease." *Tissue Antigens* **64**(5): 593-599.

Babron, M. C., F. Truy, et al. (1997). "Behaviour of the maximun liklihood score when many affected sib pairs are issued from a few untyped parents." *Genet Epid* **14**: A4.

Bader, J. (2001). "The relative power of SNPs and haplotypes as genetic markers for association tests." *Pharmacogenomics* **2**: 11-24.

Barrett, J. C., B. Fry, et al. (2004). "Haploview: analysis and visualization of LD and haplotype maps." *Bioinformatics*.

Beckett, C. G., D. Dell'Olio, et al. (1998). "The detection and localisation of inducible nitric oxide synthase production in the small intestine of patients with coeliac disease." *Eur J Gastroenterol Hepatol* **33**: 939-.

Bertina, R. M. and E. al. (1994). "Mutation in blood coagulation factor V associated with resistance to activiated protein C
" *Nature* **369**: 64-67.

Bevan, S., S. Popat, et al. (1999). "Contribution of the MHC region to the familial risk of coeliac disease." *J Med Genet* **36**(9): 687-90.

Boudraa, G., W. Hachelaf, et al. (1996). "Prevalance of celiac disease in diabetic children and their relatives in west Algeria: screening with serological markers." *Acta Paediatra* **412**(Suppl): 58.

Cardon, L. R. and G. R. Abecasis (2003). "Using haplotype blocks to map human complex trait loci." *Trends Genet* **19**(3): 135-40.

Cardon, L. R. and J. I. Bell (2001). "Association study designs for complex diseases." *Nat Rev Genet* **2**: 91-99.

Cardon, L. R. and L. J. Palmer (2003). "Population stratification and spurious allelic association." *Lancet* **361**(9357): 598-604.

Carreno, B. M. and M. Collins (2002). "The B7 family of ligands and its receptors: new pathways for costimulation and inhibition of immune responses." *Annu Rev Immunol* **20**: 29-53.

Catassi, C., E. Fabiani, et al. (1996). "The coeliac iceberg in Italy. A multicentre antigliadin antibodies screening for coeliac disease in school-age subjects." *Acta Paediatr Suppl* **412**: 29-35.

Ciclitira, P. J., M. W. Johnson, et al. (2005). "The pathogenesis of coeliac disease." *Mol Asp Med* **In press**.

Clark, A. G. (1990). "Inference of haplotypes from PCR-amplified samples of diploid populations." *Mol Biol Evol* 7: 111-122.

Clark, A. G. (2004). "The role of haplotypes in candidate gene studies." *Genet Epid* **27**: 321-333.

Clot, F., M. C. Fulchignoni-Lataud, et al. (1999). "Linkage and association study of the CTLA-4 region in coeliac disease for Italian and Tunisian populations." *Tissue Antigens* **54**(5): 527-30.

Cohavy, O., J. Zhou, et al. (2004). "LIGHT expression by mucosal T cells may regulate IFN-gamma expression in the intestine." *J Immunol* **173**(1): 251-8.

Cohavy, O., J. Zhou, et al. (2005). "LIGHT is constitutively expressed on T and NK cells in the human gut and can be induced by CD2-mediated signaling." *J Immunol* **174**(2): 646-53.

Corder, E. H. and E. al. (1993). "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families." *Science* **261**: 921-923.

Couzin, J. (2002). "Genomics. New mapping project splits the community." *Science* **296**: 1391-1393.

Dadi, H. K., A. J. Simon, et al. (2003). "Effect of CD3-delta deficiency on maturation of alpha/beta and gamma/delta T-cell lineages in severe combined immunodeficiency." *N Engl J Med* **349**: 1821-1828.

Daly, M. J., J. D. Rioux, et al. (2001). "High-resolution haplotype structure in the human genome." *Nat Genet* **29**(2): 229-32.

Daum, S., U. Bauer, et al. (1999). "Increased expression of mRNA for matrix metalloproteinases-1 and -3 and tissue inhibitor of metalloproteinases-1 in intestinal biopsy specimens from patients with coeliac disease." *Gut* **44**(1): 17-25.

Dawkins, R., C. Leelayuwat, et al. (1999). "Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease." *Immunol Rev* **167**: 275-304.

Dawson, E. and E. al. (2002). "A first generation linkage disequilibrium map of human chromosome 22." *Nature* **418**: 1719-1723.

de Saint Basile, G. G., F.; Flori, E.; Uring-Lambert, B.; Soudais, C.; Cavazzana-Calvo, M.; Durandy, A.; Jabado, N.; Fischer, A.; Le Deist, F. (2004). "Severe combined immunodeficiency caused by deficiency in either the delta or the epsilon subunit of CD3." *J Clin Invest* **114**: 1512-1517.

Dean, A. G. (1999). "Epi Info and Epi Map: current status and plans for Epi Info 2000." *J Public Health Manag Pract* **5**(4): 54-7.

DeMorchi, M., I. Borelli, et al. (1979). "Two HLA-D and DR alleles are associated with coeliac disease." *Tissue Antigens* **14**: 309-316.

Devlin, B. and N. Risch (1995). "A comparison of linkage disequilibrium measures for fine-scale mapping." *Genomics* **29**: 311-322.

Dicke, W. K., H. A. Weijers, et al. (1953). "Coeliac Disease, The presence in wheat of a factor having a deliterious effect in cases of coeliac disease." *Acta Paediatrica* **42**: 34-42.

Ding, Y., L. Qin, et al. (2001). "Differential IL-10R1 expression plays a critical role in IL-10 mediated immune regulation." *J Immunol* **167**: 6884-6892.

Djilali-Saiah, I., J. Schmitz, et al. (1998). "CTLA-4 gene polymorphism is associated with predisposition to coeliac disease." *Gut* **43**(2): 187-9.

Ellis, H. J., E. L. Pollock, et al. (2003). "Investigation of the putative immunodominant T cell epitopes in coeliac disease." *Gut* **52**: 212-217.

Ewen, K. R., M. Bahlo, et al. (2000). "Identification and analysis of error types in high-throughput genotyping
" *Am J Hum Genet* **67**: 727-736.
Excoffier, L. and M. Slatkin (1995). "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population." *Mol Biol Evol* **12**(5): 921-7.
Falchunk, Z. M., G. N. Rogentine, et al. (1972). "Predominance of histocompatibility antigen HL-A8 in patients with gluten-sensitive enteropathy." *J Clin Invest* **51**: 1602-1605.
Fernandez, L., M. Fernandez-Arquero, et al. (2002). "Triplet repeat polymorphism in the transmembrane region of the MICA gene in celiac disease." *Tissue Antigens* **59**(3): 219-22.
Folwaczny, M., J. Glas, et al. (2005). "Polymorphisms of the interleukin-18 gene in periodontitis patients." *J Clin Periodontol* **32**(5): 530-4.
Forton, J., D. Kwiatkowski, et al. (2005). "Accuracy of haplotype reconstruction from haplotype-tagging single-nucleotide polymorphisms." *Am J Hum Genet* **76**: 438-448.
Freedman, M. L., D. Reich, et al. (2004). "Assessing the impact of population stratification on genetic association studies." *Nat Genet* **36**(4): 388-393.
Furugaki, K., S. Shirasawa, et al. (2004). "Association of the T-cell regulatory gene CTLA4 with Graves' disease and autoimmune thyroid disease in the Japanese." *J Hum Genet* **49**(3): 166-8.
Gabriel, S. B. and E. al. (2002). "The structure of haplotype blocks in the human genome." *Science* **296**: 2225-2229.
Gangolfi, L., R. Pratesi, et al. (2000). "Prevalance of celiac disease among blood donors in Brazil." *Am J Gastroenterol* **95**: 689.
Ghabanbasani, M. Z., I. Buyse, et al. (1994). "Possible association of CD3 and CD4 polymorphisms with insulin-dependent diabetes mellitus (IDDM)." *Clin Exp Immunol* **97**(3): 517-521.
Gierdraitis, V., B. He, et al. (2001). "Cloning and mutation analysis of the human IL-18 promoter: a possible role of polymorphisms in expression regulation." *J Neuroimmunol* **112**: 146-152.
Gonzalez, S., L. Rodrigo, et al. (2004). "Association of MHC class I related gene B (MICB) to celiac disease." *Am J Gastroenterol* **99**(4): 676-80.
Gordon, D., S. J. Finch, et al. (2002). "Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms." *Hum Hered* **54**: 22-33.
Gracie, J. A., N. Koyama, et al. (2005). "Disease association of two distinct interleukin-18 promoter polymorphisms in Caucasian rheumatoid arthritis patients." *Genes Immun* **6**(3): 211-6.
Greco, L., M. C. Babron, et al. (2001). "Existence of a genetic risk factor on chromosome 5q in Italian coeliac disease families." *Ann Hum Genet* **65**(Pt 1): 35-41.
Greco, L., G. Corazza, et al. (1998). "Genome search in celiac disease." *Am J Hum Genet* **62**(3): 669-75.
Greco, L., R. Romino, et al. (2002). "The first large population based twin study of coeliac disease." *Gut* **50**(5): 624-8.
Green, P. H. R. and B. Jabri (2003). "Coeliac disease." *Lancet* **362**: 383-391.
Green, P. H. R. and B. Jabri (2006). "Coeliac disease." *Annu Rev Med* **57**: 14.1-14.15.
Green, P. H. R., S. N. Stravropoulos, et al. (2001). "Charcteristics of celiac disease in the USA: results of a national survey." *Am J Gastroenterol* **96**: 126-131.

Green, P. M., S. Saad, et al. (1999). "Mutation rates in humans. Overall and sex-specific rates obtained from a population study of hemophilia B." *Am J Hum Genet* **65**: 1572-1579.

Greenberg, D. A., S. E. Hodge, et al. (1982). "Evidence for recessive and against dominant inheritance at the HLA-linked locus in coeliac disease." *Am J Hum Genet* **34**: 263-277.

Groh, V., A. Steinle, et al. (1998). "Recognition of stress-induced MHC molecules by intestinal epithelial gammadelta T cells." *Science* **279**: 1737-1740.

Haaning Andersen, A. D., M. Lange, et al. (2003). "Allelic variation of the inducible costimulator (ICOS) gene: detection of polymorphisms, analysis of the promoter region, and extended haplotype estimation." *Tissue Antigens* **61**(4): 276-85.

Hahn-Zoric, M., A. M. Hytonen, et al. (2003). "Association of -1087 *IL10* and -308 *TNFA* gene polymorphisms with serological markers of coeliac disease." *J Clin Immuno* **23**: 291-296.

Haimila, K., T. Smedberg, et al. (2004). "Genetic association of coeliac disease susceptibility to polymorphisms in the ICOS gene on chromosome 2q33." *Genes Immun.*

Harbo, H. F., E. G. Celius, et al. (1999). "CTLA4 promoter and exon 1 dimorphisms in multiple sclerosis." *Tissue Antigens* **53**(1): 106-10.

Harrop, J. A., M. Reddy, et al. (1998). "Antibodies to TR2 (herpevirus entry mediator), a new member of the TNF superfamily, block T cell proliferation, expression of activation markers and production of cytokines." *J Immunol* **167**: 5099-.

Hekkens, W. J. M., A. J. Haex, et al. (1970). Some aspects of gliadin fractionation and testing by histochemical methods. Edinburgh, Churchill Livingstone.

Hernadez, J. L., J. P. Michalski, et al. (1991). "Evidence for a dominant gene mechanism underlying coeliac disease in the west of Ireland." *Genet Epid* **8**: 13-27.

Higa, S., T. Hirano, et al. (2003). "Association between interleukin-18 gene polymorphism 105A/C and asthma." *Clin Exp Allergy* **33**: 1097-1102.

Hill, I., A. Fasano, et al. (2000). "The prevalance of celiac disease in at-risk groups of children in the United States." *J Pediatr* **136**: 86.

Holopainen, P., M. Arvas, et al. (1999). "CD28/CTLA4 gene region on chromosome 2q33 confers genetic susceptibility to celiac disease. A linkage and family-based association study." *Tissue Antigens* **53**(5): 470-5.

Holopainen, P., K. Mustalahti, et al. (2001). "Candidate gene regions and genetic heterogeneity in gluten sensitivity." *Gut* **48**(5): 696-701.

Holopainen, P., A. T. Naluai, et al. (2004). "Candidate gene region 2q33 in European families with coeliac disease." *Tissue Antigens* **63**(3): 212-22.

Horton, R., L. Wilming, et al. (2004). "Gene map of the extended human MHC." *Nat Rev Genet* **5**(12): 889-99.

Houlston, R. S., I. P. Tomlinson, et al. (1997). "Linkage analysis of candidate regions for coeliac disease genes." *Hum Mol Genet* **6**(8): 1335-9.

Hovdenak, N., E. Hovlid, et al. (1999). "High prevalence of asymptomatic coeliac disease in Norway: a study of blood donors." *Eur J Gastroenterol Hepatol* **11**(2): 185-7.

Hugot, J. P., M. Chamaillard, et al. (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." *Nature* **411**: 599-603.

Hunt, K. A., D. P. McGovern, et al. (2005). "A common CTLA4 haplotype associated with coeliac disease." *Eur J Hum Genet.*

Iles, M. M. (2004). "The effect of SNP marker density on the efficacy of haplotype tagging SNPs - a warning." *Ann Hum Genet* **00**: 209-215.

Jeffreys, A. J., L. Kauppi, et al. (2001). "Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex." *Nat Genet* **29**: 217-222.

Johnson, G. C., L. Esposito, et al. (2001). "Haplotype tagging for the identification of common disease genes." *Nat Genet* **29**(2): 233-7.

Johnston, S. D., R. G. Watson, et al. (1997). "Prevalence of coeliac disease in Northern Ireland." *Lancet* **350**(9088): 1370.

Ke, X. and L. R. Cardon (2003). "Efficient selection screening of haplotype tag SNPs." *Bioinformatics* **19**: 287-288.

Ke, X., A. Collins, et al. (2001). "PIRA PCR designer for restriction analysis of single nucleotide polymorphisms." *Bioinformatics* **17**(9): 838-9.

Kim, W. J., Y. J. Kang, et al. (2005). "LIGHT is involved in the pathogenesis of rheumatoid arthritis by inducing the expression of pro-inflammatory cytokines and MMP-9 in macrophages." *Immunology* **114**(2): 272-9.

King, A. L., S. J. Moodie, et al. (2003). "Coeliac disease: investigation of proposed causal variants in the CTLA4 gene region." *Eur J Immunogenet* **30**(6): 427-32.

King, A. L., S. J. Moodie, et al. (2002). "CTLA-4/CD28 gene region is associated with genetic susceptibility to coeliac disease in UK families." *J Med Genet* **39**(1): 51-4.

King, A. L., J. Y. Yiannakou, et al. (2000). "A genome-wide family-based linkage study of coeliac disease." *Ann Hum Genet* **64**(Pt 6): 479-90.

Kolho, K. L., M. A. Farkkila, et al. (1998). "Undiagnosed coeliac disease is common in Finnish adults." *Scand J Gastroenterol* **33**(12): 1280-3.

Kretowski, A., K. Mironczuk, et al. (2002). "Interleukin-18 promoter polymorphisms in type 1 diabetes." *Diabetes* **51**(11): 3347-9.

Kruglyak, L. (2005). "Power tools for human genetics." *Nat Genet* **37**(12): 1299-300.

Lander, E. S. and L. Kruglyak (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." *Nat Genet* **11**: 241-247.

Li, S. S., N. Khalid, et al. (2003). "Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms." *Biostatistics* **4**(4): 513-22.

Likken, D. T., A. Tellegen, et al. (1978). "Volunteer bias in twin research: the rule of two-thirds." *Soc Biol* **25**: 1-9.

Lin, S., D. J. Cutler, et al. (2002). "Haplotype inference in random population samples." *Am J Hum Genet* **71**: 1129-1137.

Lio, D., L. Scola, et al. (2005). "TNFalpha, IFNgamma and IL-10 gene polymorphisms in a sample of Sicilian patients with coeliac disease." *Dig Liver Dis.* **37**(10): 756-760.

Little, J., L. Bradley, et al. (2002). "Reporting , appraising, and integrating data on genotype prevalence and gene-disease associations." *Am J Epidemiol* **156**: 300-310.

Liu, J., S. H. Juo, et al. (2002). "Genomewide linkage analysis of celiac disease in Finnish families." *Am J Hum Genet* **70**(1): 51-9.

Lohmueller, K. E., C. L. Pearce, et al. (2003). "Meta-analysis of genetic variation association studies supports a contribution of common variants to susceptibility to common disease." *Nat Genet* **33**: 177-182.

Lopez-Vazquez, A., L. Rodrigo, et al. (2002). "MHC class I chain related gene A (MICA) modulates the development of coeliac disease in patients with the high risk heterodimer DQA1*0501/DQB1*0201." *Gut* **50**(3): 336-40.

Louka, A. S., S. J. Moodie, et al. (2003). "A collaborative European search for non-DQA1*05-DQB1*02 celiac disease loci on HLA-DR3 haplotypes: analysis of transmission from homozygous parents." *Hum Immunol* **64**(3): 350-8.

Louka, A. S., E. K. Stensby, et al. (2002). "Coeliac disease candidate genes: no association with functional polymorphisms in matrix metalloproteinase 1 and 3 gene promoters." *Scand J Gastroenterol* **37**(8): 931-5.

Louka, A. S., A. Torinsson Naluai, et al. (2002). "The IL12B gene does not confer susceptibility to coeliac disease." *Tissue Antigens* **59**(1): 70-2.

Macdonald, W. C., W. O. Dobbins, 3rd, et al. (1965). "Studies of the Familial Nature of Celiac Sprue Using Biopsy of the Small Intestine." *N Engl J Med* **272**: 448-56.

Maiuri, L., C. Ciacci, et al. (2001). "FAS engagement drives apoptosis of enterocytes of coeliac patients." *Gut* **48**: 418-424.

Marchini, J., L. R. Cardon, et al. (2004). "The effects of human population structure on large genetic association studies." *Nat Genet* **36**(5): 512-7.

Marsh, M. N. (1992). "Gluten, major histocompatibility complex, and the small intestine. A molecular and immunobiologic approach to the spectrum of gluten senstitivity ('celiac sprue')." *Gastroenterology* **102**: 330-354.

Martin-Pagola, A., G. Perez de Nanclares, et al. (2003). "No association of CTLA4 gene with celiac disease in the Basque population." *J Pediatr Gastroenterol Nutr* **37**(2): 142-5.

Martin, R. J., D. A. Savage, et al. (2005). "Interleukin 18 promoter polymorphisms are not strongly associated with type 1 diabetes in the UK population." *Genes Immun* **6**(2): 171-174.

McLoughlin, R., S. S. Sebastian, et al. (2003). "Coeliac disease in Europe." *Aliment Pharmacol Ther* **18 Suppl 3**: 45-8.

McManus, R., M. Moloney, et al. (1996). "Association of celiac disease with microsatellite polymorphisms close to the tumor necrosis factor genes." *Hum Immunol* **45**(1): 24-31.

McManus, R., A. G. Wilson, et al. (1996). "TNF2, a polymorphism of the tumour necrosis-alpha gene promoter, is a component of the celiac disease major histocompatibility complex haplotype." *Eur J Immunol* **26**(9): 2113-8.

Mein, C. A., B. J. Barrett, et al. (2000). "Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation." *Genome Res* **10**: 330-343.

Moodie, S. J., P. J. Norman, et al. (2002). "Analysis of candidate genes on chromosome 19 in coeliac disease: an association study of the KIR and LILR gene clusters." *Eur J Immunogenet* **29**(4): 287-91.

Moore KW, de Waal Malefyt R, et al. (2001). "Interleukin 10 and the interleukin 10 receptor." *Annu Rev Immunol* **19**: 683-765.

Mora, B., M. Bonamico, et al. (2003). "CTLA-4 +49 A/G dimorphism in Italian patients with celiac disease." *Hum Immunol* **64**(2): 297-301.

Morahan, G., D. Huang, et al. (2001). "Linkage disequilibrium of a type 1 diabetes susceptibility locus with a regulatory IL12B allele." *Nat Genet* **27**(2): 218-21.

Morrissey, P. J. and K. Charrier (1994). "Induction of wasting disease in SCID mice by the transfer of normal $CD4^+/CD45RB^{high}$ T cells and the regulation of this autoactivity by $CD4^+/CD45RB^{low}$ T cells." *Res Immunol* **145**: 357-.

Muller-Myhsok, B. and L. Abel (1997). "Genetic analysis of complex diseases." *Science* **275**: 1328-1329.

Mylotte, M., B. Egan-Mitchell, et al. (1974). "Family studies in coeliac disease." *Q J Med* **43**(171): 359-69.

Naluai, A. T., S. Nilsson, et al. (2001). "Genome-wide linkage analysis of Scandinavian affected sib-pairs supports presence of susceptibility loci for celiac disease on chromosomes 5 and 11." *Eur J Hum Genet* **9**(12): 938-44.

Naluai, A. T., S. Nilsson, et al. (2000). "The CTLA4/CD28 gene region on chromosome 2q33 confers susceptibility to celiac disease in a way possibly distinct from that of type 1 diabetes and other chronic inflammatory disorders." *Tissue Antigens* **56**(4): 350-5.

Neuhausen, S. L., M. Feolo, et al. (2001). "Linkage analysis of HLA and candidate genes for celiac disease in a North American family-based study." *BMC Med Genet* **2**(1): 12.

Nielsen, D. M., M. G. Ehm, et al. (1998). "Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus." *Am J Hum Genet* **63**(5): 1531-40.

Nilsen, E. M., F. L. Jahnsen, et al. (1998). "Gluten induces an intestinal cytokine response strongly dominated by interferon gamma in patients with celiac disease." *Gastroenterology* **115**: 551-563.

Nistico, L., R. Buzzetti, et al. (1996). "The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry." *Hum Mol Genet* **5**(7): 1075-80.

Niu, T., Z. S. Qin, et al. (2002). "Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms." *Am J Hum Genet* **70**: 157-169.

Novak, N., S. Kruse, et al. (2005). "Single nucleotide polymorphisms of the IL18 gene are associated with atopic eczema." *J Allergy Clin Immunol* **115**(4): 828-33.

Novota, P., K. Kolostova, et al. (2005). "Interleukin IL-18 gene promoter polymorphisms in adult patients with type 1 diabetes mellitus and latent autoimmune diabetes in adults."

Olmos, P., R. A'Hern, et al. (1988). "The significance of the concordance rate of type 1 (insulin-dependent) diabetes in identical twins." *Diabetologia* **31**: 747-750.

Ozaki, K., Y. Ohnishi, et al. (2002). "Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction." *Nat Genet* **32**: 650-654.

Patil, N., A. J. Berno, et al. (2001). "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21." *Science* **294**: 1719-1723.

Pender, S. L. F. and E. al. (1997). "A major role of matrix metalloproteinases in T cell injury in the gut." *J Immunol* **158**: 1582-1590.

Pennacchio, L. A., M. Olivier, et al. (2001). "An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing." *Science* **294**: 169-173.

Percopo, S., M. C. Babron, et al. (2003). "Saturation of the 5q31-q33 candidate region for coeliac disease." *Ann Hum Genet* **67**(Pt 3): 265-8.

Phillips, M. S., R. Lawrence, et al. (2003). "Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots." *Nat Genet* **33**(3): 382-7.

Polski, R., J. Ek, et al. (1993). "On the HLA-DQ(a1*0501,b10201) associated susceptibility in celiac disease: a possible gene dosage effect of DQ1*0201." *Tissue Antigens* **41**: 173-177.

Polui, A., M. Maki, et al. (1998). "TNF microsatellite alleles a2 and a3 are not primarily associated with celiac disease in the Finnish population." *Tissue Antigens* **51**: 553-555.

Popat, S., S. Bevan, et al. (2002). "Genome screening of coeliac disease." *J Med Genet* **39**: 328-331.

Popat, S., N. Hearle, et al. (2002). "Variation in the CTLA4/CD28 gene region confers an increased risk of coeliac disease." *Ann Hum Genet* **66**(Pt 2): 125-37.

Popat, S., N. Hearle, et al. (2002). "Analysis of the CTLA4 gene in Swedish coeliac disease patients." *Scand J Gastroenterol* **37**(1): 28-31.

Purcell, S., S. S. Cherny, et al. (2003). "Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits." *Bioinformatics* **19**(1): 149-150.

Reich, D. E. and E. al. (2001). "Linkage disequilibrium in the human genome." *Nature* **411**: 199-204.

Reich, D. E. and E. al. (2002). "Human genome sequence variation and the influence of gene history, mutation and recombination." *Nat Genet* **32**: 135-142.

Reich, D. E., S. B. Gabriel, et al. (2003). "Quality and completeness of SNP databases." *Nat Genet* **33**: 457-458.

Rioux, J. D., H. Karinen, et al. (2004). "Genomewide search and association studies in a Finnish celiac disease population: Identification of a novel locus and replication of the HLA and CTLA4 loci." *Am J Med Genet* **130A**(4): 345-50.

Rioux, J. D., M. S. Silverberg, et al. (2000). "Genomewide search in Canadian families with imflammatory bowel disease reveals two novel inflammatory loci." *Am J Hum Genet* **66**(1863-).

Roberts, A. I., L. Lee, et al. (2001). "NKG2D receptors induced by IL-15 costimulate CD28-negativeneffector CTL in the tissue microenvirnoment." *J Immunol* **167**: 5527-5530.

Roses, A. D. (1997). "A model for susceptibilty polymorphisms for complex diseases:apolipoprotein E and Alzheimer's disease." *Neurogenet* **1**: 3-11.

Rostom, A., C. Dube, et al. (2005). "The diagnostic accuracy of serological tests for coelaic disease: a systematic review." *Gastroenterology* **128**: S38-S46.

Ruben, C. E. (1960). "Some reflections on reversibility, gluten and the intestine." *Gastroenterology* **39**: 260-261.

Ruben, C. F., L. L. Brandborg, et al. (1962). Biopsy Studies on the pathogenesis of coeliac sprue in intestinal biopsy, Wolstenhome, G.E.W, Camera, M.P. (Eds.), London Ciba Foundation Study Book 14.

Rueda, B., M. A. Gonzalez-Gay, et al. (2005). "Interleukin-18-promoter polymorphisms are not relevant in rheumatoid arthritis." *Tissue Antigens* **65**: 544-548.

Rueda, B., M. Pascual, et al. (2003). "Association of MICA-A5.1 allele with susceptibility to celiac disease in a family study." *Am J Gastroenterol* **98**(2): 359-62.

Rueda, B., A. Zhernakova, et al. (2005). "Association study of functional genetic variants of innate immunity related genes in celiac disease." *BMC Med Genet* **6**: 29-35.

Ryan, A. W., J. M. Thornton, et al. (2004a). "Chromosome 5q candidate genes in coeliac disease: genetic variation at IL4, IL5, IL13, IL17B and NR3C1." *Tissue Antigens* **65**: 150-155.

Ryan, A. W., J. M. Thornton, et al. (2004). "Haplotype variation at the IBD5/SLC22A4 locus (5q31) in coeliac disease in the Irish population." *Tissue Antigens* **64**(2): 195-8.

Salvati, V. M., G. Mazzarella, et al. (2005). "Recombinant human interleukin 10 suppresses gliadin dependent T cell activation in ex vivo cultured coeliac intestinal mucosa." *Gut* **54**: 46-53.

Sblattero, D., I. Berti, et al. (2000). "Human recombinant tissue transglutaminase ELISA; an innovative diagnostic assay for celiac disease." *Am J Gastroenterol* **95**: 1253-1257.

Shaikh, R. B., S. Santee, et al. (2001). "Constitutive expression of LIGHT on T cells leads to lymphocyte activation, inflammation and tissue distruction." *J Immunol* **167**: 6330-.

Shidrawi, R. G., N. J. D. Parnell, et al. (1998). "Binding of gluten derived peptides to the HLA-DQ2 (α1*0501,β1*0201) molecule, assessed in a cellular assay." *Clin Exp Immunol* **111**: 158-165.

Shiina, T., H. Inoko, et al. (2004). "An update of the HLA genomic region, locus information and disease associations:2004." *Tissue Antigens* **64**: 631-649.

Shipman, R. T., A. L. Williams, et al. (1975). "A family study of coeliac disease." *Aust N Z J Med* **5**(3): 250-5.

Sivalingam, S. P., K. H. Yoon, et al. (2003). "Single-nucleotide polymorphisms of the interleukin-18 gene promoter region in rheumatoid arthritis patients: protective effect of AA genotype." *Tissue Antigens* **62**(6): 498-504.

Sollid, L. M. (2002). "Coeliac disease: dissecting a complex inflammatory disorder." *Nat Rev Immunol* **2**(9): 647-55.

Sollid, L. M., G. Markussen, et al. (1989). "Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer." *J Exp Med* **169**: 345-350.

Sollid, L. M., S. N. McAdam, et al. (2001). "Genes and environment in celiac disease." *Acta Odontol Scand* **59**(3): 183-6.

Sollid, L. M. and E. Thorsby (1993). "HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis." *Gastroenterology* **105**(3): 910-22.

Stenhammar, L., A. Brandt, et al. (1982). "A family study of coeliac disease." *Acta Paediatr Scand* **71**(4): 625-8.

Stephens, M. and P. Donnelly (2003). "A comparison of bayesian methods for haplotype reconstruction from population genotype data." *Am J Hum Genet* **73**(5): 1162-9.

Stephens, M., N. J. Smith, et al. (2001). "A new statistical method for haplotype reconstruction from population data." *Am J Hum Genet* **68**(4): 978-89.

Stewart, C. A., R. Horton, et al. (2004). "Complete MHC haplotype sequencing for common disease gene mapping." *Genome Res* **14**(6): 1176-87.

Stokes, P. L., R. Ferguson, et al. (1976). "Familial aspects of coeliac disease." *Q J Med* **45**(180): 567-82.

Stram, D. O., C. A. Haiman, et al. (2003). "Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study." *Hum Hered* **55**(1): 27-36.

Sulkanen, S., T. Halttunen, et al. (1998). "Tissue transglutaminase autoantibody enzyme-linked immunosorbent assay in detecting celiac disease." *Gastroenterology* **115**: 1322-1328.

Tamura, K., Y. Fukuda, et al. (2002). "IL18 polymorphism is associated with an increased risk of Crohn's disease." *J Gastroenterol* **37**(Suppl 14): 111-116.

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

TheMHCSequencingConsortium (1999). "Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium." *Nature* **401**(6756): 921-3.

Turner, D. M., D. M. Williams, et al. (1997). "An investigation of polymorphism in the interleukin-10 gene promoter." *Eur J Immunogenet* **24**: 1-8.

Ueda, H., J. M. Howson, et al. (2003). "Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease." *Nature* **423**(6939): 506-11.

Vader, W., Y. Kooy, et al. (2002). "The gluten response in children is directed against multiple gliadin and glutenin peptides." *Gastroenterology* **122**: 1729-1737.

Vader, W., D. Stepniak, et al. (2003). "The HLA-DQ2 gene dose effect in celiac disease is directly related to the
magnitude and breadth of gluten-specific T cell responses." *Proc Natl Acad Sci U S A* **21**: 12390-12395.

Vaidya, B., H. Imrie, et al. (1999). "The cytotoxic T lymphocyte antigen-4 is a major Graves' disease locus." *Hum Mol Genet* **8**(7): 1195-9.

Van Belzen, M. J., J. W. Meijer, et al. (2003). "A major non-HLA locus in celiac disease maps to chromosome 19." *Gastroenterology* **125**: 1032-1041.

van Belzen, M. J., C. J. Mulder, et al. (2004). "CTLA4 +49 A/G and CT60 polymorphisms in Dutch coeliac disease patients." *Eur J Hum Genet* **12**(9): 782-5.

Van Eerdewegh, P., R. D. Little, et al. (2002). "Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness." *Nature* **418**: 426-430.

van Heel, D. A., J. Satsangi, et al. (2000). "Inflammatory bowel disease: progress toward a gene." *Can J Gastroenterol* **14**(3): 207-18.

Vermeire, S., G. Wild, et al. (2002). "CARD15 genetic variation in a Quebec population: Prevalence, genotype-phenotype relationship, and haplotype structure." *Am J Hum Genet* **71**: 74-83.

Walsh, E. C., K. A. Mather, et al. (2003). "An integrated haplotype map of the human major histocompatibility complex." *Am J Hum Genet* **73**(3): 580-90.

Wang, J., J. C. Lo, et al. (2001). "The regulation of T cell homeostasis and autoimmunity by T cell-derived LIGHT." *J Clin Invest* **108**(12): 1771-80.

Weiss, K. M. and J. D. Terwilliger (2000). "How many diseases does it take to map a gene with SNPs?" *Nat Genet* **26**: 151-157.

West, J., R. F. Logan, et al. (2003). "Seroprevalence, correlates, and characteristics of undetected coeliac disease in England." *Gut* **52**(7): 960-5.

Wijmenga, C., M. J. van Belzen, et al. (2000). "The CTLA-4 gene is not associated with coeliac disease in the Dutch population." *J. Pediatr. Gastroenterol. Nutr.* **31**: S16-S16.

Willer, C. J., D. A. Dyment, et al. (2003). "Twin concordance and sibling recurrence rates in multiple sclerosis." *Proc Natl Acad Sci USA* **100**: 12877-12882.

Wood, J. P., M. A. Pani, et al. (2002). "A recently described polymorphism in the CD28 gene on chromosome 2q33 is not associated with susceptibility to type 1 diabetes." *Eur J Immunogenet* **29**(4): 347-9.

Woolley, N., P. Holopainen, et al. (2002). "A new locus for coeliac disease mapped to chromosome 15 in a population isolate." *Hum Genet* **111**: 40-45.

Wu, J., B. Z. Alizadeh, et al. (2004). "Association of FAS (TNFRSF6)-670 gene polymorphism with villous atrophy in coeliac disease." *World J Gastroenterol* **10**(5): 717-20.

Yamashita, T., K. Hamaguchi, et al. (2004). "IKBL promoter polymorphism is strongly associated with resistance to type 1 diabetes in Japanese." *Tissue Antigens* **63**(3): 223-30.

Zaykin, D. V., P. H. Westfall, et al. (2002). "Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals." *Hum Hered* **53**(2): 79-91.

Zeggini, E., W. Rayner, et al. (2005). "An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets." *Nat Genet* **37**(12): 1320-2.

Zhang, P. A., J. M. Wu, et al. (2005). "Association of polymorphisms of interleukin-18 gene promoter region with chronic hepatitis B in Chinese Han population." *World J Gastroenterol* **11**(11): 1594-8.

Zhong, F., C. C. McCombs, et al. (1996). "An autosomal screen for genes that predispose to celiac disease in the western counties of Ireland." *Nat Genet* **14**(3): 329-33.

Zondervan, K. T. and L. R. Cardon (2004). "The complex interplay among factors that influence allelic association." *Nat Rev Genet* **5**(2): 89-100.