HLA Disease Associations Methods Manual Version 0.1.0 (July 25, 2011)

HLA Disease Associations:
Detecting Primary and Secondary Disease Predisposing Genes

Available online at: The Immunology Database and Analysis Portal: https://www.immport.org

Glenys Thomson[1,*], Steven J. Mack[2], Ana M. Valdes[3], Lisa F. Barcellos[4], Pierre-Antoine Gourraud[5], Jill A. Hollenbach[2], Wolfgang Helmberg[6], Richard M. Single[7]

[1] Department of Integrative Biology, 3060 Valley Life Sciences Building MC #3140, University of California, Berkeley, CA 94720-3140, USA, glenys@berkeley.edu

[2] Center for Genetics, Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr Way, Oakland, CA 94609, USA, e-mails: sjmack@chori.org, jhollenbach@chori.org

[3]Twin Research Unit, King's College School of Medicine, London, UK, valdes.anam@gmail.com

[4]Division of Epidemiology, School of Public Health, University of California, Berkeley, CA 94720-7360, USA, barcello@genepi.berkeley.edu

[5]Department of Neurology, University of California, San Francisco, CA 94143, USA, pierreantoine.gourraud@ucsf.edu

[6]Department of Blood Group Serology and Transfusion Medicine, University of Graz, Graz Austria, wolfgang.helmberg@medunigraz.at

[7]Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05405, USA, richard.single@uvm.edu

Acknowledgments

This publication was supported in part by NIH/NIAID contract number HHSN266200400076C, ABD N01-AI-40076 (GT and RMS), and NIH AI67068 (JAH, SJM).

I. Overview

- ***A. Introduction***
- ***B. Primary disease genes: tests of linkage and association***
- ***C. Primary disease genes: modes of inheritance***
- ***D. Secondary disease genes***
- ***E. Detecting amino acids at classical HLA genes directly involved in disease risk***
- ***F. Analysis of KIR-HLA disease associations***

II. Primary Disease Genes: Tests of Linkage and Association

- ***A. Introduction to HLA associated diseases***
- ***B. HLA nomenclature, ambiguity reduction, and population data analyses***
- ***C. Family data and controls (AFBACs)***
- ***D. Tests of marker associations with disease***
- ***E. Stratified association tests, age of onset, and maternal-fetal effects***
- ***F. Linkage disequilibrium***

III. Primary Disease Genes: Modes of Inheritance

- ***A. The patient/control (P/C) ratio and relative penetrances***
- ***B. Detecting relative predispositional risk effects (RPEs) and T1D data***
- ***C. All pairwise relative risk comparisons and JIA-OP data***
- ***D. Analysis of subsets of the data and the single parent TDT***
- ***E. Genotype frequencies and tests of modes of inheritance***
- ***F. Affected sib pair identity by descent (IBD) values and modes of inheritance***
- ***G. The interrelationship of HLA associated diseases***

IV. Secondary Disease Genes

- ***A. Introduction***
- ***B. Matched cases and controls***
- ***C. Homozygous parent linkage and TDT tests***
- ***D. Conditional haplotype, genotype and logistic regression methods***
- ***E. Combined association and IBD data***

V. Detecting Amino Acids at Classical HLA Genes Involved in Disease Risk

- ***A. Introduction***
- ***B. Within serogroup and sequence alignment comparisons***
- ***C. The Unique Combinations Method***
- ***D. Sequence Feature Variant Type (SFVT) analysis***

Abbreviations used in the text

AFBACs—affected family based controls

AGFAP—antigen/allele genotype frequencies among patients

BISC—Bioinformatics Integration Support Contract

CGM—conditional genotype method

CHM—conditional haplotype method

CLR—conditional logistic regression

df—degrees of freedom (in, for example, a chi-square test of heterogeneity)

DR3—a haplotype of DRB1, DQA1, and DQB1 with the following allelic variants: DRB1*03:01- DQA1*05:01-DQB1*02:01

DR4— haplotypes of DRB1, DQA1, and DQB1 with the following allelic variants: DRB1*04:01- DQA1*03:01-DQB1*03:02 or DRB1*04:04-DQA1*03:01-DQB1*03:02, used in discussions of HLA associations with type 1 diabetes (T1D)

genoPDT—genotype pedigree disequilibrium test

GWAS—genome wide association study (GWASs—studies)

HLA genes—human leukocyte antigen presenting immune response genes

HLA class I genes—A, B, and C

HLA class II genes—DRB1, DQA1, DQB1, DPA1, DPB1 (DRA1 is much less polymorphic and rarely tested in disease association studies; DRB3, DRB4, and DRB5 have previously rarely been tested, but are now included in proposed high throughput sequencing studies)

HLA DR-DQ genes—shorthand for the HLA class II DRB1, DQA1, and DQB1 set of high LD genes: sometimes with DQA1 not typed

HPLT—homozygous parent linkage test

HPTDT—homozygous parent transmission disequilibrium test

HRR—haplotype relative risk

HSH—haplotype specific heterozygosity

HWP—Hardy Weinberg proportions

IBD—identity by descent (values of 2, 1, and 0 indicate sharing of parental alleles IBD for affected sib pairs, often denoted as X, Y, and Z)

IDAWG— Immunogenomics Data Analysis Working Group (IDAWG) (www.immunogenomics.org)

IDDM—insulin dependent diabetes mellitus (now called type 1 diabetes - T1D)

IDDM1—the HLA class II DRB1, DQA1, DQB1 genes that are the primary disease predisposing HLA region genes for type 1 diabetes (T1D), previously called IDDM, and are the major genetic contributors to T1D disease risk

ImmPort—Immunology Database and Analysis Portal (https://www.immport.org)

JIA—juvenile idiopathic arthritis

JIA-OP—juvenile idiopathic arthritis with the subphenotype oligoarticular-persistent

KIR—killer cell immunoglobulin-like receptor

LD—linkage disequilibrium

LR—logistic regression

MHC—major histocompatibility complex

MLE—maximum likelihood estimate

MPC pedigrees—multiplex parent-child family data (parents and offspring) ascertained based on the presence of at least one affected parent and one affected child

MPS pedigrees—multiplex parent-sib pair family data (parents and offspring) ascertained based on the presence of at least one affected parent and an affected sib pair

MSAT—microsatellite locus

MSP pedigrees—multiplex sib pair family data (parents and offspring) ascertained based on the presence of at least two affected sibs (previously referred to as MS (multiplex sibs) pedigrees, see e.g., Thomson (1995a, b)

NIDDM—non-insulin dependent diabetes mellitus (now called type 2 diabetes – T2D)

NK—natural killer

OCGM—overall conditional genotype method

OCHM—overall conditional haplotype method

OR—odds ratio

PCA—principal components analysis

P/C ratio—ratio of patient to control population level *frequencies* for an allele, haplotype, or genotype

PDT—pedigree disequilibrium test

PyPop—*Py*thon for *Pop*ulation Genomics analyses (an analysis software package specifically designed to handle HLA data - www.pypop.org)

QC—quality control

RPEs—relative predispositional effects

RR—relative risk

S pedigrees—simplex family data (parents and offspring) ascertained based on the presence of at least one affected child

SNP—single nucleotide polymorphism

SPTDT—single parent TDT

SFVT—sequence feature variant type

T1D—type 1 diabetes (previously referred to as IDDM – insulin dependent diabetes mellitus)

T2D—type 2 diabetes (previously referred to as NIDDM – non-insulin dependent diabetes mellitus)

TDT—transmission disequilibrium test

VNTR—variable number of tandem repeats

I. Overview

Note that, except as necessary, references are not given in Section I; detailed references are listed later in the relevant sections.

A. Introduction

Our aim is to review methods to optimize detection of *all* disease genes in a genetic region that are directly involved in differential risk to a *complex disease*. A number of factors make the study of complex diseases difficult, including the following:

a. Incomplete penetrance: not all susceptible individuals are affected.

b. Disease heterogeneity: e.g., maturity onset diabetes of the young (MODY), type 1 diabetes (T1D), and type 2 diabetes (T2D) were initially regarded as different manifestations of the same disease. We now know that even within these three separate categories there is additional genetic heterogeneity. Most likely, many complex diseases have genetic heterogeneity, e.g., schizophrenia and alcoholism; although informative subdivisions of the data for genetic analyses based on for example, clinical features, have so far not been successful.

c. Low disease prevalence: e.g., T1D and multiple sclerosis in China are so rare that affected sib pair families are not available for linkage screens. In this situation, disease genes can only be investigated using association studies, e.g., looking at genes identified in Caucasians or other ethnic groups as involved in disease, candidate genes, associated single nucleotide polymorphisms (SNPs) identified by genome wide association studies (GWASs), and genes identified in animal models.

d. High disease prevalence: e.g., heart disease; the high prevalence can cause problems with linkage analyses as disease genes may be inherited from both sides of the family.

e. Late-onset diseases: e.g., T2D, heart disease, Alzheimer's and Parkinson's disease are difficult to study by linkage methods as parents of the affected patient are usually deceased, as may be some siblings.

f. Infectious diseases: e.g., AIDS and tuberculosis are particularly difficult to study due to difficulties in obtaining accurate knowledge of exposure.

g. The involvement of several (many) disease-predisposing loci: apart from HLA (human leukocyte antigen) associations, with many complex diseases the disease predisposing genes identified by GWASs generally have small odds ratios (ORs).

h. Environmental factors, age of onset, gender-specific, maternal-fetal, and imprinting effects have all been shown to apply for many complex diseases and can alter the power of a study, and also result in heterogeneity in results between studies.

i. Interaction effects: the genetic and biochemical pathways involved in disease risk are poorly understood at this point, and similarly interaction effects between disease predisposing loci.

j. Strong linkage disequilibrium (LD) between closely linked loci: once a genetic region has been implicated in disease risk, localization of the locus or loci or amino acids directly

involved is complicated by the high LD which is often seen between closely linked markers. Cross-ethnic studies can sometimes be very important in showing consistency in risk heterogeneity at a putative primary disease predisposing gene, and also in identification of primary genes when there is a breakdown in LD.

These features make it difficult to localize disease genes, ascertain the number and relationship of disease loci involved, understand modes of inheritance and interaction effects, determine the molecular basis of disease, and understand the mechanism(s) by which these genetic changes give rise to disease. This latter point unfortunately holds even when strong associations with primary disease genes have been identified, e.g., the many diseases with strong and verified associations and direct differential risk effects for the classical HLA immune response genes.

There has been much discussion of the role of common versus rare genetic variation in differential risk for complex diseases (see e.g., Loehmuller et al. 2003). A combination of these two extreme models may very well apply. Different individuals and families will often have disease predisposition due to a different set of disease predisposing genes, making their identification even more difficult. Further, to always keep in mind, all these genes are *susceptibility* rather than *necessary* loci. Also, every disease may show its own unique features of genetic predisposition, making it difficult to predict which methods may be optimal for detecting disease genes. Comparisons between familial and "sporadic" cases of disease may be informative, and again patterns of inheritance may differ greatly between diseases.

Despite extensive efforts by many groups, until recently only a few genes and some genetic regions involved in complex diseases had been identified. The general picture was one of difficulty in locating disease genes and replication of reported linkages and associations. The major exception was the many well established HLA disease associations. In a meta-analysis by Hirschhorn et al. (2002), only six non-HLA genes were consistently replicated as associated with complex diseases, including apolipoprotein E (APOE) with Alzheimer's disease and CCR5 with AIDS. These six genes were identified from literature data on 166 putative associations which had been studied three or more times.

Since this meta-analysis, the pace of discovery of complex disease genes identified via linkage and association studies has increased due to the availability of new technologies. GWASs are finding new disease predisposing genes or genetic regions; although the SNPs verified as associated with disease appear to explain only a small fraction of the genetic heritability. GWASs have also identified many additional diseases with validated HLA region associations. The exact number of verified HLA disease associations is not known, but was estimated to be > 300, and certainly > 100, even before more recent GWAS findings (which identify the HLA region as involved in disease, but do not identify the locus or loci directly involved in differential disease risk).

In this review, our discussions focus on the HLA region (human ch. 6p21), although the analysis methods and issues discussed are generic and relate to all genomic regions and complex diseases in general. Note however, that many standard analysis programs are designed for the study of bi-allelic SNP data and cannot accommodate the highly polymorphic HLA data (see Table I.A.1). The HLA classical antigen presenting genes usually typed are: class I (A, B, and C) and class II (DRB1, DQA1, DQB1, DPA1, and DPB1). DRA is not routinely typed as the single known amino acid is not known to be functionally significant: also DRB3/4/5 are usually not typed due to high LD with DRB1. As seen in Table I.A.1, there are thousands of possible alleles

for class I loci and hundreds for some class II loci. The number of alleles seen in any individual study will be substantially less than these total numbers of variants; however it is not uncommon to see 40 to 50 protein-level alleles at DRB1, 15 or more at DQB1, and 50-80 DRB1-DQB1 haplotypes. See www.ebi.ac.uk/imgt/hla/stats.html for data on additional HLA and non-HLA loci.

Table I.A.1: Polymorphism of the HLA class I and II genes[a]

Gene	Alleles	Proteins	Nulls[b]
Class I			
A	1698	1243	83
B	2271	1737	73
C	1213	884	33
Class II			
DRA	7	2	0
DRB1	975	736	11
DRB3	57	46	0
DRB4	15	8	3
DRB5	19	16	2
DQA1	44	27	1
DQB1	158	109	1
DPA1	32	16	0
DPB1	149	129	3

[a] From www.ebi.ac.uk/imgt/hla/stats.html (IMGT/HLA database Release 3.5.0, July 14, 2011)
[b] Null alleles are not expressed on the cell surface

As our starting point, we discuss methods to test for evidence of the involvement in disease risk of a genetic region under study, based on *significance levels* from linkage and/or association studies, *replication studies*, or *meta-analyses* (Section I.B and Section II). Our emphasis is on the HLA region including its inherent complexities of nomenclature, typing techniques evolving through time, and the continual discovery of new alleles. For closely linked markers, there will often be multiple associations with disease, and linkage analyses identify a region rather than the specific disease predisposing gene. Hence, the first task is to identify the *primary* (major) disease predisposing gene or genes in a genetic region, that is, to distinguish "true" associations from those due to LD with the actual disease predisposing variants. (We remind the reader that these predisposing loci directly influence differential disease risk heterogeneity within the context of incomplete penetrance.)

Throughout, we will follow the somewhat vague definition that has evolved with progress in disease gene discovery (see Thomson et al. 2008): *primary* disease genes are those that "stand out" in initial association studies, and *for the most part* associations of other markers in the genetic region can be explained via their LD patterns with the primary disease gene(s). A number

of genes or SNPs (not necessarily in the same gene) in a region may be included in the primary disease gene category. Additional (*secondary*) disease predisposing genes are detected once we take account of (condition on) these *primary* disease genes, usually after study of additional marker loci with an increased sample size; secondary disease genes are expected to have more subtle, weaker effects. Also, these secondary effects may be restricted to a subset of alleles, haplotypes, or genotypes at the *primary* disease gene, and some may only be seen in specific populations or ethnic groups.

Initially, only the classical HLA class I and II genes were studied (first via serological typing and more recently via molecular typing) with a few other genes in the HLA region studied occasionally. Identification of the primary gene(s) at the classical loci was based mainly on higher ORs for the class II DR-DQ (DRB1, DQA1, DQB1) associations than for class I (in most cases), with informal, as well as formal, applications of conditional analyses taking account of LD patterns. Cross-population and cross-ethnic studies were, and continue to be, very important in showing consistency in risk heterogeneity at a putative primary gene, and also in identification of primary genes when there was a breakdown in LD, e.g., with the closely linked and high LD HLA DR-DQ (DRB1, DQA1, DQB1) genes. With more recent typing of microsatellites (MSATs) and then large numbers of SNPs in the HLA region, the direct primary role of one or more classical HLA genes has continued to be validated.

Under the assumption that a primary HLA region gene (or genes) has been identified, we discuss analyses that can be applied to the data to detect significant differences in relative risk effects and modes of inheritance (Section I.C and Section III). A full understanding of the heterogeneity of genetic risk at a primary locus is necessary to avoid spurious results when testing for secondary disease genes.

How do we detect *secondary* (additional) disease genes? The stratification (conditional) analyses described to detect *secondary* disease genes (Section I.D and Section IV) are also of course the basis of determining if a gene is *primary* in disease risk heterogeneity. Novel methods have been developed in study of both the primary and secondary disease genes in the HLA region. There is evidence of the role in disease risk of secondary genes in the HLA region, including other classical HLA genes, as well as effects associated with MSATs and SNPs. However, identification of MSAT and SNP associations that are not due to LD with primary disease predisposing genes has been as difficult as study of the non-HLA disease genes for a complex disease; in both cases considerable heterogeneity is seen across studies.

With both primary and secondary disease genes that involve one or more classical HLA antigen presenting genes, an additional aim is to detect the specific *amino acids, or combinations of amino acids,* that are directly involved in disease risk heterogeneity (Section I.E and Section V). As with conditional analyses of HLA allele, haplotype and genotype data, identifying specific amino acid effects is also difficult, again due to high LD and extensive polymorphism at functionally important sites.

HLA class I proteins are also ligands for the KIR (killer cell immunoglobulin-like receptor - human ch. 19q13.4) inhibitory and activating receptors that are expressed on natural killer (NK) cells, and a small percentage of cytotoxic T-cells, and regulate cell killing and cytokine response. Disease associations with variation at KIR loci has also been demonstrated, generally in combination with their HLA class I ligands. A summary of specific issues relating to study of KIR associations with disease is given in Section I.F; more specific details will be given at a later date in a separate Methods Manual.

No *existing* programs, either individually, or in aggregate, can handle the breadth and complexity of the analyses that are needed to detect primary and secondary disease genes in the HLA region using currently available methods. Further, even with modern computers one cannot study, let alone *interpret*, every possible combination of genetic markers and their haplotypes across the genome, or even within a genetic region. The complex LD patterns of HLA region genes, SNPs, amino acids at the classical HLA genes, etc., are a major complicating factor in deciphering specific genetic variants directly involved in differential disease risk. Further, many analysis software packages do not work with the high degree of polymorphism found at HLA, while others only allow a limited number of highly polymorphic loci to be evaluated.

The effects of environmental factors, age of onset and gender-specific effects, undetected heterogeneity of disease, gene-gene and gene-environment interactions further complicate our attempts to detect predisposing genes (*primary* and *secondary*) for complex diseases. Our survey of methods below particularly highlights the importance of a *complementary, multi-strategy* array of methods to uncover all the different facets of complex genetic diseases.

We recommend that the data and analysis results be studied carefully to understand differences in significance from different methods. There is no "best practice" for the types of analyses needed for these projects. The power of different methods will vary depending on the specific genetic and environmental features of the disease under consideration; for most complex diseases this involves many unknown factors. Although we must rely heavily on computers, in the final analysis of multiple effects in a genetic region and/or interaction or independent effects between unlinked genes, manipulation and scrutiny of the data by the individual investigator must play a crucial role. Our aim is to look for consistency of results across studies, and to take note of effects which may be seen in one or more analyses but may be missed in other analyses.

From our combined experience in analyses of data on complex diseases, we emphasize that there is a required interplay between studying individual SNP, MSAT, and HLA allele, haplotype, and genotype effects, and that all putative significant effects must be scrutinized in individual detail by the researcher. Also, fit of the data to a particular model, for example, that a particular disease gene can explain all the linkage and association data in a genetic region, *does not validate* that model. It may merely mean there is insufficient power to detect additional effects, or that appropriate stratification of the data to detect additional effects was not carried out, or that additional markers need to be typed. However, *rejection* of a model, apart from type 1 error or breaking of assumptions of the model, for example, random sampling of a homogenous population, allows one to *unequivocally* state that additional factors must be incorporated into the model of disease, and guides further investigation of such factors.

In the following, we will not discuss the issue of correcting for multiple testing. From our experiences with HLA-associated diseases with regard to detecting the effects of non-HLA region genes as well as additional genes in the HLA region, the issue is the difficulty in finding let alone replicating effects, rather than a deluge of type 1 errors. Many initial findings of primary HLA disease associations would not have survived correction for multiple tests, yet they have been extensively replicated and verified. Although we are very sensitive to the issue of type 1 errors, our emphasis throughout will be on detecting effects to be followed up in independent studies, ranked of course by relative p-values. We strongly encourage use of *resampling* methods in all analyses of multiple markers in a genetic region. One can then adjust for multiple testing of non-independent results due to LD and take account of the fact that haplotypes are estimated rather than known as assumed in most analyses, and to give an empiric p-value to guide interpretation of results. However, in many cases the computing needs of resampling techniques

may be prohibitive. Other approaches for multiple testing include sequential methods (adjusting first for a number of loci and then a number of alleles at a given locus at a second stage), and adjusting for a number of effectively independent comparisons based on LD among loci/alleles/sequence features.

B. Primary disease genes: tests of linkage and association

Examples of diseases with well established primary HLA associations include (see Dausset and Svejgaard 1977, Tiwari and Terasaki 1985, Lechler 1994, Thorsby 1997, Thorsby et al. 2007):

class II: HLA DR-DQ (DRB1, DQA1, DQB1) for type 1 diabetes (T1D); HLA-DRB1 for rheumatoid arthritis, juvenile idiopathic arthritis (JIA), multiple sclerosis, and systemic lupus erythematosus; HLA-DQB1 for narcolepsy; and HLA-DPB1 for chronic berrylium disease; and

class I: HLA-B for ankylosing spondilitis and HLA-C for psoriasis.

Hemochromatosis is an example where the initial association was with a classical HLA gene (the A3 allele of the HLA-A gene), but the *primary* gene mapped to the extended HLA region. Many of the HLA-associated diseases are of autoimmune or inflammatory origin, but Hodgkin's disease and other cancers, and infectious diseases such as malaria, tuberculosis, and AIDS also show HLA associations.

With the many GWAS results implicating the HLA region, the classical HLA genes are the most likely candidates. In examples where GWAS and HLA typing results are both available, the p-value and OR from the strongest SNP association is in some cases close to the results from HLA typing, e.g., multiple sclerosis (IMSGC 2007, Ramagopalan et al. 2009), and systemic lupus erythematosis (Barcellos et al. 2009). In many other cases, the HLA region association for GWAS data is much smaller than that derived from HLA typing data for the primary disease gene. For T1D, ORs as high as ~22 have been observed for the most common predisposing three-locus HLA DR-DQ genotype in Caucasians (DRB1-DQA1-DQB1: 03:01-05:01-02:01 / 04:01-03:01-03:02), and an OR of ~0.03 for the most protective (dominant) haplotype (15:01-01:02-06:02), compared to the GWAS strongest effect with an OR of 0.28 (or 3.57 for the alternate SNP allele) (Cooper et al. 2008). When only GWAS data are available giving strong evidence of HLA region involvement, as in schizophrenia (Shi et al. 2009) and Parkinson's disease (Hamza et al. 2010), one cannot predict if the association(s) will be similar, or much stronger, with HLA typing. The latter is probably more likely the more complex the hierarchy of relative predisposing or protective risk categories, as with T1D, or if appropriate tagging SNPs for the classical HLA loci (most of which are unknown) were not included in the GWAS SNPS.

As additional markers in the HLA region are typed, the conditional analysis methods described in Section IV to detect secondary disease genes, must first be applied to the *primary* disease gene, to show continued *validation*, or not, of its primary role. When secondary genetic effects *are identified* in a region, one would then again apply the methods of Section III to understand the genetic basis and heterogeneity of this additional effect, in combination with the primary disease gene.

The following is a summary of the topics and methods described in Section II (Primary Disease Genes: Tests of Linkage and Association):

1. **HLA Nomenclature and STREIS Reporting Guidelines**: HLA data should be recorded using the current nomenclature version (http://www.ebi.ac.uk/imgt/hla/nomenclature). Further, the principles of the STrengthening the REporting of Immunogenomic Studies (STREIS) guidelines developed by the Immunogenomics Data Analysis Working Group (IDAWG) (www.immunogenomics.org) are strongly recommended (also see Hollenbach et al. 2011b, Gourraud et al. 2011). Quality control (QC) of the data should include: validation of allele names against a specifc IMGT/HLA database release, appropriate binning of alleles for meta-analyses, testing of Hardy Weinberg proportions (HWP) in controls (significant deviations may indicate errors in allele calls), and study by the investigator of multilocus haplotype patterns, as previously unobserved high LD HLA B-C, DR-DQ and DPA1-DPB1 haplotypes *may* flag errors. Modules in the PyPop (Python for Population Genomics) software package (www.pypop.org) (Lancaster et al. 2003, 2007a, 2007b) can be used for all these analyses. All family genotypes should be examined for Mendelian inconsistencies using for example PEDCHECK (O'Connell and Weeks 1998).

2. **Tests of nuclear family AFBAC data:** For family based data, the AFBACs (affected family based controls) are the non- or never-transmitted parental alleles as appropriate to the ascertainment scheme [S (simplex) ascertainment refers to nuclear family samples selected based on at least one affected child, other ascertainment schemes considered are MPC (multiplex parent-child), MSP (multiplex sib pairs), and MPS (multiplex parent-sib pairs) Section II.C]. Appropriate tests are:
   a. **Equality of AFBAC maternal and paternal frequencies** (a significant difference may imply a maternal/fetal interaction effect, in which case only the paternal AFBACs should be used).
   b. **Test for equality of AFBAC frequencies with control marker frequencies available from any other sources,** e.g., population data, other disease studies, and AFBACs ascertained under different criteria. With case/control data, one must be careful to understand the ascertainment scheme for the controls, i.e., whether it is from a random *or* non-diseased population. For rare diseases there will be little difference, but for common diseases this is not the case.
   c. **Test for multiplicative structure of the paternal transmitted and non-transmitted alleles for S ascertainment**, with appropriate modifications for other ascertainment schemes, and similarly for the maternal alleles. A significant difference in the latter case may reflect a maternal/fetal interaction effect.
   d. **Test for a multiplicative structure of the paternal non-transmitted alleles and the child's genotype for S ascertainment**, with appropriate modifications for other ascertainment schemes, and similarly for the maternal alleles. A significant difference in the latter case may reflect a maternal/fetal interaction effect.
   e. **Maternal offspring compatibility** of the mother's and affected child's genotypes can also be tested using conditional logistic regression (CLR) (Bronson et al. 2009), with compatibility or not of the child to the father used as a control.
   f. **Test for fit to HWP of the AFBAC genotypes created, with S ascertainment**, from the maternal and paternal non-transmitted alleles in each family. For MSP

ascertainment (nuclear families with at least two affected siblings), only the families where both sibs share the same two parental alleles (AC) (Figure II.C.1c) are used in this test. A significant deviation from HWP may reflect a maternal/fetal interaction effect, as well as population stratification effects.

3. **Tests of association and linkage:** A number of methods to test the data for association or linkage or both are available; some of course are specific to particular ascertainment schemes. In all tests, the effect of the sex of the affected child, and age of onset of disease should be examined and used as a covariate in CLR tests or appropriate subdivisions of the data in chi-square heterogeneity tests (Section II.D). Standard disease association and or linkage analyses of the classical HLA genes are usually the first step in any study. Given the strong LD between the class II DR-DQ genes, the class I B-C genes, the class II DPA1-DPB1 genes, and the DRB3/4/5 loci and DRB1, these are usually analyzed as haplotypes, and locus specific effects may be determined when there is sufficient breakdown in the LD with heterogeneity of risk effects. Genotype effects are also considered, with the caveat that genotype frequencies rapidly become very small in patients, and are always small in controls due to the high level of polymorphism of HLA alleles. For this reason most analyses of classical HLA loci focus on allele or haplotype data, and genotype data analyses are often restricted to comparisons of specific, more frequent, genotype subsets.

   a. **For case/control data, test for association** of marker genes linked to a disease predisposing gene, or a gene directly involved in disease risk heterogeneity, using a contingency table test of heterogeneity, CLR, or other tests. The individual allele, haplotype or genotype contributions to the overall test can also be examined (see Section II.D). Age of onset effects should always be considered when possible, and likewise heterogeneity in allele frequencies across the study population identified for example by principal components analysis (PCA), both can be accommodated as covariates in CLR analyses.
   b. **For nuclear family based data, test for association** of marker genes linked to a disease predisposing gene, or a gene directly involved in disease risk heterogeneity, using the transmission disequilibrium test (TDT) (Spielman et al. 1993).
   c. **For nuclear family based data, test for equality of paternal versus maternal transmission rates** of marker alleles within each ascertainment scheme
   d. **For nuclear family based data, test for paternal versus maternal effects which may be genotype dependent**, i.e., with S ascertainment test for equality of A_1A_k genotypes in the affected child versus A_kA_1, where the first listed allele is that transmitted from the father, k = 2, 3, …m (Thomson 1995b).
   e. **For nuclear family based data ascertained for the presence of at least two affected sibs (MSP pedigrees)** deviations from the Mendelian random expectations of 25%, 50%, and 25% that two affected sibs will on average share 2, 1, and 0 parental chromosomes in common identical by descent (IBD) implicate a disease predisposing gene in the region.

f. **For nuclear family based MSP data**, use of allele sharing IBD values can increase the power of association studies, and vice versa, partitioning linkage analyses by genotypes of associated alleles can also increase power (Section II.E)

g. **Test for heterogeneity of risk effects** would be the next step for genes showing an overall association with disease (see Section I.C and Section III).

4. **Linkage disequilibrium (LD) values** should be plotted for all markers in the region where significant associations with disease are found. Informal inspection of the data across loci can target the possibility of the highest overall chi-square values (with heterogeneity testing of patient and control allele frequencies), or highest ORs for specific alleles, implicating the primary disease gene or genes, and the associations of other genes being *for the most part* explained by the known patterns of LD of the HLA genes. Formal analyses would involve use of conditional methods described in Section I.D and Section IV.

   As mentioned above, many standard population and disease analysis software do not allow for the high level of polymorphism of the HLA genes, including programs for estimation of haplotype frequencies and LD. As mentioned in point 1 above, appropriate algorithms have been developed for HLA data using the haplotype and LD estimation module in PyPop (www.pypop.org) but the precompiled version of the software for haplotype estimation is currently limited to a total of 7 loci and sample size n = 5,000 individuals at a time. Users can modify these constraints in the source code. The Estihaplo algorithm of Gourraud et al. (2007) (http://birl.supbiotech.fr/hla-estihaplo.html) does not have this restriction and will be incorporated into PyPop at a later date. Similarly, the iHap web application (www.immunogenomics.org/ software.html), is capable of estimating haplotypes for much larger datasets than PyPop. Standard measures of individual allele and overall LD for a pair of loci, as well as normalized values, should be calculated, as well as haplotype frequencies across a number of loci at a time after the QC steps outlined above (see Single et al. 2011, Mack et al. 2011, Hollenbach et al. 2011b).

C. Primary disease genes: modes of inheritance

How do we determine if a primary disease predisposing gene has been identified? Then, how do we detect any additional disease predisposing genes in this genetic region? The methods we use to detect, or not, secondary disease-predisposing genes in a region (Section IV) are also the base of the formal analyses that must be performed to detect a primary disease predisposing gene or genes versus marker genes in LD with the primary disease gene. Detailed study of this primary disease gene, or combination of genes would then follow (Section III).

Historically, the classical HLA genes were studied first with serological typing of class I and later class II loci and more recently via molecular typing of class II followed later by class I. As mentioned above, identification of the primary gene(s) was based mainly on higher ORs for the class II DR-DQ associations than for class I (in most cases), with informal, as well as formal, applications of conditional analyses taking account of LD patterns. With typing of additional genes, as well as MSAT and SNP typing in the HLA region, the primary role of HLA classical

genes in disease has been demonstrated in many, but not all, cases. Secondary roles of other classical HLA genes, as well as other HLA region loci have been demonstrated.

Once a primary disease gene (or genes) in a genetic region has been identified, the emphasis is then on understanding all genetic aspects of its mode of inheritance and heterogeneity in risk: these topics are covered below in Section III (Primary Disease Genes: Modes of Inheritance). Detecting differential risk between alleles, haplotypes, and genotypes at the primary disease gene is a major consideration as these influence the categories to be considered with the stratification analyses described below in Section IV (Secondary Disease Genes). We may often need to combine sets of HLA alleles, haplotypes, and genotypes at the primary disease gene that have similar relative risk effects into a "homogenous" class, in order to avoid sparseness of cells and resulting possible bias and inaccuracy in parameter estimates, e.g., with CLR and conditional haplotype and genotype methods (CHM and CGM). Note however, that we could miss heterogeneity with some analyses, and must constantly be on the alert for this possibility. As additional disease genes are identified, interplay with *re-analysis* of the relative risk effects at the primary and secondary disease genes is required.

The most complex pattern of an HLA disease association is seen with T1D, the HLA DRB1-DQB1 haplotypes and genotypes show a hierarchy from highly predisposing, predisposing, intermediate ("neutral"), protective to highly protective effects. The relative risk patterns are seen consistently across all ethnic groups (summarized in the meta-analysis of Thomson et al. 2007a). In other cases, such as multiple sclerosis and narcolepsy, there is one major predisposing HLA DR-DQ haplotype, with additional smaller effects of other alleles at these loci (Barcellos et al. 2006, Mignot et al. 2001, 2007). Studies in non-Caucasian populations were also able to tease apart the primary effects of DRB1 in multiple sclerosis (Oksenberg et al. 2004) and DQB1 in narcolepsy (Mignot et al. 2001) due to a breakdown in the strong LD.

The following is a summary of the topics and methods described in Section III (Primary Disease Genes: Modes of Inheritance):

1. **Relative risk effects are ranked by the ORs of alleles, haplotypes and genotypes** with the aim of defining specific risk categories of predisposing through protective effects, including subsets with *homogenous risk.* These can be defined with some confidence for the more common variants using the methods described below.
2. **Significantly different relative predispositional effects (RPEs)** of alleles, haplotypes or genotypes are determined by sequential analyses of the data for heterogeneity, with removal of the most significant effects at each stage.
3. **All pairwise relative risk comparisons** of alleles, haplotypes or genotypes extend and complement the above analysis in terms of delineating groups of alleles with homogeneity of risk within the group and heterogeneity between. The analysis can consider more common alleles where the boundaries between differential risk effects are clearer, as well as rarer alleles where the boundaries are usually less clear. These results are also very pertinent to application of the Unique Combinations Method of Salamon et al. (1996) (Section I.E and Section V) to detect *amino acids* that distinguish between heterogeneous risk categories of HLA alleles or haplotypes.
4. **Modes of inheritance** studies can be carried out on genotype data using the AGFAP [antigen (allele) genotype frequencies among patients] method and CLR models, affected

sib pair IBD values, and combined analyses thereof. With family data selected for the presence of at least one affected parent and one or two affected offspring (MPC and MPS pedigrees respectively), test for equality (recessive) or not (additive) of the parental contributions from the affected versus unaffected parent. Ultimately, all features of the data must be explained if the primary gene has been identified, or additional secondary genes remain to be identified.

5. **Cross-population and cross-ethnic studies and meta-analyses:** If a primary disease gene has been identified, then *a consistent hierarchy* of ORs and differential risk effects should be seen across studies and ethnic groups, and *disease prevalence* should correlate with the relative frequencies in ethnic groups of the predisposing, intermediate, and low risk alleles, haplotypes and genotypes (see for example Valdes et al. 1997, Thomson et al. 2007a). Some ethnic groups and not others may show a breakdown in a strong LD pattern allowing discrimination of which gene (or genes) is directly involved in disease risk.
6. **Interrelationship of HLA associated diseases:** the same alleles or haplotypes may be associated with differential risk for a number of diseases, e.g., DRB1*15:01 DQB1*06:02 is predisposing for multiple sclerosis and narcolepsy while it is very protective for T1D; DRB1*03:01 DQB1*02:01 is predisposing for T1D and celiac disease; DRB1*04:01 DQB1*03:02 is predisposing for T1D and rheumatoid arthritis. It is of particular interest to study the genetic interrelationship of these and other HLA associated diseases.

D. Secondary disease genes

We stress again that we must identify *all* heterogeneity in disease risk at the *primary* disease predisposing gene, *including* relatively weak effects, before proceeding with conditional analyses to detect *secondary* genetic effects. Otherwise, spurious results in identifying secondary disease predisposing genes may be seen resulting from unaccounted for heterogeneity in relative risk effects at the primary disease gene at the allele, haplotype, or genotype levels.

Taking account of these precautions, with many HLA-associated diseases there is evidence of the role in disease risk of additional genes in the HLA region; both other classical HLA genes, as well as additional genes across the HLA region. Especially in the latter case, as mentioned above, identification of these genes has been as difficult as study of non-HLA disease genes, with considerable heterogeneity seen across studies. Many reports of other HLA region gene, MSAT, and SNP associations have appeared in the literature. In many of these studies it has been difficult to determine if an additional HLA region gene is involved in disease, versus the associations reflecting LD with the antigen presenting HLA molecules directly involved in disease. However, a number of analytic strategies have been developed to remove the effects of LD with the antigen presenting HLA genes directly involved in the disease (Section IV).

Notwithstanding, there are many challenges in trying to identify appropriate techniques for the analysis of the extensive data generated with detailed study of specific genetic regions. Some markers that do not show a single locus association with disease may nevertheless be directly involved in disease predisposition; their effects may only become obvious once the data are stratified by the effects of a primary disease gene or interaction effects are taken into account. Hence, as *primary* and *secondary* disease genes are identified, subsequent analyses should consider *all* remaining marker genes, or minimally tagging SNPs thereof, rather than, for

example, focusing on a set of markers which initially showed a significant association as a single SNP (see Thomson et al. 2008). We emphasize that categories classified as homogenous with respect to risk should be continually *reevaluated* as additional genes involved in disease are discovered.

Further, associations between markers and disease loci that are not evident with a single marker locus may be identified in multi-locus marker analyses using known (from family data, although it is rare they will be known with 100% accuracy unless the loci are highly polymorphic such as most of the HLA classical loci) or estimated haplotype frequencies (see Single et al. 2011). With linkage analyses, suspicion of additional disease genes in a region can initially present as a secondary peak in an area not in LD with the primary linkage peak.

Even with sample sizes in the thousands, with stratification analyses, many cells in the data matrix will have very small or zero values, particularly as *secondary*, in addition to *primary*, disease genes are identified and multi-locus haplotypes are considered. In addition to a lack of statistical power in such cases, significance levels reported from application of a software package to these data may be spurious due to inappropriate use of a statistical test on rare alleles, haplotypes, or genotypes.

When secondary genes are identified for diseases with strong primary associations with a classical HLA gene(s), they are often reported as *independent* effects. In most cases, the authors are referring to *additional* (secondary) effects, which is a different issue than a test for independence or not of risk effects. In fact, in some cases the effect is only seen on specific high risk, or vice versa non-high risk, alleles, haplotypes or genotypes at the primary HLA classical gene(s). Care should thus be taken to be very specific about what has been shown as a statistically significant effect.

Methods to detect genes or markers additional to a primary predisposing gene in a genetic region all rely on stratification analyses to take account of the effects of LD with the primary predisposing gene(s). With all the methods discussed below, power is an issue even with sample sizes in the hundreds or thousands, and as mentioned above, stratification approaches can quickly result in small cell sizes. All of the methods described below have been successful in detecting the role of additional disease genes in the HLA region, however, the actual genes involved have not always been identified with MSAT and SNP typing, and there is considerable heterogeneity between studies.

The following is a summary of the topics and methods described in Section IV (Secondary Disease Genes):

1. **Matched cases and controls and the homozygous parent linkage and TDT tests:** These complementary tests each use only a restricted subset of the data. In some cases power may be increased by combining data across studies and even ethnic groups; however, this gain in power may be countered by loss if there are population or ethnic group specific effects.
2. **Conditional haplotype, genotype and logistic regression methods (CHM, CGM, and CLR):** The basis of specific and overall haplotype or genotype tests, is that if all disease risk in a genetic region is explained by a primary gene denoted by locus A, then with conditioning on locus A, variation at locus B (which under the null hypothesis has no effect on disease) will have the same expected relative frequencies in patients and

controls. Deviation from these expectations implicates additional disease predisposing genes in the genetic region. Step-wise and CLR can also be used to account for known covariates with disease, e.g., age of onset, principle components etc.

3. **Combined association and linkage data:** Extending the conditional analyses above, it is informative in tests of fit of a putative primary disease gene(s) to all aspects of the data to include both association and linkage data, e.g., using the MASC method of Clerget-Darpoux et al. (1988).

E. Detecting amino acids at classical HLA genes directly involved in disease risk

When a classical HLA gene has been identified as primary in a disease association, our interest then focuses on identifying *specific amino acids* and *combinations thereof* that are potentially involved in differential disease risk. This is difficult for a number of reasons: the extensive polymorphism of the amino acid sites, with most of the variation occurring at functionally important sites, and with up to six "alleles" seen at some amino acid sites; the nature of HLA polymorphism which is the result of mutation and also gene conversion events which results in a patch-work pattern of amino acid variation; and the high LD between many alleles and amino acid sites.

Peptide motifs important for binding to HLA molecules, including critical residues, have been defined. Determining the HLA amino acid residues directly involved in a specific disease can facilitate predictions about peptide epitopes that are more, or less, likely to be presented by a particular HLA allele. Such knowledge can of course be very important in the design of vaccines and our understanding of autoimmunity (see Karp et al. 2010).

The following is a summary of the topics and methods described in Section V (Detecting Amino Acids at Classical HLA Genes Directly Involved in Disease Risk):

1. **Within serogroup analyses**: comparisons of the amino acid sequences of alleles within the same serogroup with differential risk effects versus those within the same risk category can implicate amino acids directly involved in disease risk heterogeneity. This approach focuses on a smaller number of amino acids to compare and led to for example identification of amino acid position 57 of DQB1 in T1D risk. Also it may pick up interaction effects that are missed in other analyses, for example amino acid position 86 of DRB1 in systemic sclerosis and JIA (Karp et al. 2010, Thomson et al. 2010).
2. **Sequence alignment of alleles stratified by risk category**: this is most likely to be informative when there are only a few risk categories of alleles, and it is not known how often the rheumatoid arthritis "shared epitope" phenomenon will be observed; also note that there is disease risk heterogeneity within the "shared epitope" set of alleles.
3. **Salamon's Unique Combinations Method:** identifies amino acids that distinguish for example a set of high disease risk sequences from *all other* sequences, or from the set of sequences of intermediate, or separately low, disease risk alleles. (The original algorithm of Salamon et al. (1996) has been slightly modified (Thomson et al. 2010) and see Section V.) Preliminary analyses indicate that this method may be particularly useful in more directly detecting amino acids involved in differential disease risk versus those

associations due to LD, and also interaction effects. To apply the Unique Combinations Method one needs to accurately define sets of alleles with homogenous risk within a set and heterogeneity of risk between sets.

4. **Sequence Feature Variant Type (SFVT) analysis:** a set of association tests are systematically applied focusing on variation (termed variant types - VTs) at biologically relevant SFs, which are based on structural and functional features of the protein (Karp et al. 2010).
5. **Conditional Haplotype Method (CHM):** as described above for detecting primary and secondary disease predisposing genes, a series of stratified analyses can be applied to detect amino acids directly involved in differential disease risk versus those associations with disease due to LD (Karp et al. 2010, Thomson et al. 2010); the difficulty lies in the extensive amino acid LD at the HLA loci. Even starting with all pairwise comparisons and building up from that, it is difficult to pinpoint residues directly involved in disease risk heterogeneity due to the complex patterns of LD. Further, a small number of highly polymorphic amino acids often define most allele level variation, excluding the possibility of further conditional tests.

F. Analysis of KIR-HLA disease associations

A role for the KIR loci and their HLA ligands in autoimmune diseases (Khakoo et al. 2004, Li et al. 2004, Williams et al. 2005, Khakoo and Carrington 2006, Hollenbach et al. 2009), and infectious diseases such as HIV and Hepatitis C (Gaudieri et al. 2005, Khakoo and Carrington 2006), as well as solid organ and hematopoietic stem cell transplant (Sun et al. 2005, Kunert et al. 2007, Gedil et al. 2007) and pregnancy (Lanier 1999, Moffett and Hiby 2007) is now established (see Hollenbach et al. 2011a). As with HLA, the KIR loci also have high levels of polymorphism and LD, and a clear history of extensive recombination and gene conversion events, and evidence of selection acting at the population level (Single et al. 2007a). However, the KIR typing systems are not as well developed as those for HLA; they currently often type only for presence or absence of a gene and in many cases investigators have to deal with large amounts of missing data leading to specific issues in data analyses (Single et al. 2008). Likewise, most typing techniques cannot detect or distinguish between loci that may be duplicated on a KIR haplotype. In addition, the relationship between the KIR and their HLA ligands dictates consideration of interaction effects between these two important polymorphic regions when analyzing for disease associations. Details of analysis of KIR data will be provided in a separate Methods Manual at a later date (also see Hollenbach et al. 2011a, b).

II. Primary Disease Genes: Tests of Linkage and Association

A. Introduction to HLA associated diseases

The search for associations between genes in the HLA region and specific diseases in humans was stimulated by studies in mice showing associations of the mouse major histocompatibility (MHC) system (termed H-2) and oncogenic viruses (reviewed in Klein 1975). Initial studies in humans showed weak associations with Hodgkin's (Amiel et al. 1967) and some other diseases. Subsequent studies showed more striking and consistent associations (see Dausset and Svejgaard 1977). While ABO blood group associations with disease were well known and replicated, the odds ratios (ORs) were all small and a mechanism implicating them in disease risk was not clear. In contrast, many of the HLA disease associations were quite striking and consistently found (some representative examples are given below in Table II.A.1 for some of the earlier studies with serological data) and a functional role of immune response region genes was feasible.

Table II.A.1: HLA associated diseases*

HLA	PATIENTS	CONTROLS	OR
Ankylosing spondylitis			
(a) A2	64%	50%	1.8
B27	**94%**	**9%**	**158.4**
Cw1	42%	7%	9.6
Cw2	44%	13%	5.3
(b) **B27**	**90%**	**8%**	**87.8**[a]
(c) **B27**	**92%**	**10%**	**103.5**
Type 1 diabetes			
(a) B8	37%	22%	2.1[a]
B15	23%	15%	2.1[a]
(b) **DR3**	**52%**	**23%**	**3.6**
DR4	**74%**	**24%**	**9.0**
DR3 or DR4	**93%**	**43%**	**17.6**
DR2	**4%**	**29%**	**0.1**
Multiple sclerosis			
(a) **DR2**	**67%**	**25%**	**6.1**
Rheumatoid arthritis			
(a) **DR4**	**81%**	**33%**	**8.7**
Narcolepsy			
(a) **DR2**	**95%**	**33%**	**38.6**

Hemochromatosis			
(a) A3	72%	21%	9.7
Celiac disease			
(a) B8	79%	31%	8.4
DR3	**96%**	**27%**	**64.9**

* For each disease, the frequency of the presence of an associated HLA allele in homozygotes or heterozygotes is given in patients and controls. The letter designation denotes the HLA gene, while the number is assigned to a specific allele at the gene. The data shown are older serological level HLA typing, rather than more recent molecular typing. Recent data are usually reported as allele, haplotype or genotype frequencies, rather than presence versus absence of an allele as above. The data are from various sources, including summaries and references in Thomson (1981), Thomson (1983), Thomson et al. (1988). The allele with the strongest association, based on the odds ratio (OR), is indicated in bold. Type 1 diabetes (T1D) was previously referred to as juvenile diabetes, and then insulin dependent diabetes mellitus (IDDM), before the present designation.

[a] Based on multiple studies (see Thomson 1981).

With HLA disease associations, the high level of LD between many of the classical HLA genes means that multiple disease associations are often observed, some of which may indicate a genetic factor directly implicated in disease risk heterogeneity, and others may be due to LD of a marker with this locus. This is illustrated by the original associations of the serologically defined class I alleles A3, B7, and later the class II DR2 allele with multiple sclerosis in samples of European origin. With T1D a similar picture due to LD is seen, with initial A1, B8, and B15, and later DR3 and DR4, associations. As class II typing became available (originally DR and later DQ and then DP), and later high resolution molecular typing (class II and then class I genes), these and many other diseases showed a stronger association with the HLA class II genes (usually DRB1 and/or DQB1); mostly based on higher ORs, and also that the class I associations could *mostly* be explained by LD with the class II genes (Tiwari and Terasaki 1985, Lechler 1994, Thorsby 1997, Thorsby et al. 2007). Exceptions are for example, ankylosing spondilitis and its primary B27 association, psoriasis and C6, chronic berrylium disease and its association with glutamic acid at amino acid position 69 of HLA DPB1 (Snyder et al. 2008), and hemochromatosis, where the initial association was with A3, but the *primary* gene mapped to the extended HLA region.

Affected sib pair methods were also used early in the study of HLA associated diseases. Deviations from the Mendelian random expectations of 25%, 50%, and 25% that two affected sibs will on average share 2, 1, and 0 parental chromosomes in common identical by descent (IBD) implicate a disease predisposing gene in the region. Linkage of the HLA region with T1D was initially demonstrated with 15 affected sib pairs ($p<0.001$) (Cudworth and Woodrow 1975) with a mean sharing IBD value of 0.81 (compared to the expectation of 0.5 in the absence of a disease predisposing gene in the region). These values were validated in a number of studies, including the meta-analysis of Payami et al. (1985) (see Table II.A.2). Other HLA-associated diseases such as rheumatoid arthritis and multiple sclerosis required much larger sample sizes, in the hundreds, to get statistically significant linkage. However, even these numbers are quite moderate. Use of MSAT and SNP markers *in lieu of* classical HLA typing in genome-wide linkage scans can make it more difficult to detect linkages.

Table II.A.2: Affected sib pair identity by descent values for type 1 diabetes (T1D)

A. Cudworth and Woodrow 1975

IBD Sharing:	**2**	**1**	**0**	
Observed	67%	27%	6%	Total 15

χ^2_2 = 14.07, $p < 0.001$, mean sharing = 0.805

B. Payami et al. 1985, 538 families

IBD Sharing:	**2**	**1**	**0**	
Observed:	373	283	55	Total 711
	52%	40%	8%	

χ^2_2 = 314.03, $p < 10^{-5}$, mean sharing = 0.724

Initial modeling of HLA disease associations assumed that the serologically HLA defined alleles were markers for closely linked disease predisposing genes, which were often assumed to have recessive or dominant modes of inheritance (see e.g., Thomson and Bodmer 1977a, b). This continued even after elucidation of the role of MHC molecules in the adaptive immune response via MHC restriction whereby T-cell recognition of infected cells requires a combined signal from both MHC molecules and pathogen peptides (Zinkernagel and Doherty 1974). MHC restriction allowed the high polymorphism of MHC molecules to be interpreted in a functional sense. Doherty and Zinkernagel (1975) reasoned that individuals heterozygous at MHC loci would be at a selective advantage as they could mount an immune response across a broader range of pathogens. Many lines of evidence support a role of selection as well as reproductive mechanisms in shaping MHC variation (reviewed in Meyer and Thomson 2001, also see Solberg et al. 2008).

The demonstration in the 1980's that MHC molecules directly bind peptides that are then presented to T-cells (Guillet et al. 1986), combined with increasing disease data, focused attention more on a direct role of the HLA antigen presenting molecules in disease, including auto-immune responses. It turns out that the "simple" disease models described above were nonetheless very instructive in the development of our understanding of HLA disease associations, and in some instances continue to be so today.

With molecular typing and examination of LD patterns and conditional haplotype analyses, the association of the haplotype DRB1*15:01 DQA1*01:02 DQB1*06:02 (the serological designation DR2 was later split into DR15 and DR16) was shown to be the primary multiple sclerosis association in Europeans. With T1D, the data not only showed evidence of a stronger, hence possibly direct, role of DR3 and DR4 in disease risk, but also excess risk with the heterozygous combination B8/B15 and later DR3/DR4 over either homozygote (Svejgaard et al. 1980, Svejgaard and Ryder 1981, Rotter et al. 1983, Louis and Thomson 1986, Thomson et al. 1988).

Cross ethnic studies in particular can be very important in distinguishing between the effects of the DRB1, DQA1, and DQB1 genes, which are in very strong LD. For T1D both DRB1

and DQB1 are directly involved in disease with a complex hierarchy of highly predisposing, predisposing, intermediate, protective, and highly protective effects of both haplotypes and genotypes (see for example, Thomson et al. 2007a, Erlich et al. 2008 and references therein). In Caucasian populations, the three most common T1D predisposing haplotypes are DRB1*03:01 DQB1*02:01, DRB1*04:01 DQB1*03:02, and DRB1*04:04 DQB1*03:02, and a strong protective effect is seen for DRB1*15:01 DQB1*06:02 (see summary Table III.B.1 given later). Further, the DR-DQ haplotypes show consistency of the *hierarchy* of ORs and relative risk effects, even in ethnic groups where for example the high risk DR3 and/or DR4 haplotypes are missing (see e.g., Thomson et al. 2007a). The *role of DQB1* in T1D risk was demonstrated in many studies via differential risk of DRB1*04:01 on DQB1*03:02 (predisposing) versus DQB1*03:01 (protective) haplotypes (Todd et al. 1987, Horn et al. 1988). The role of *DRB1 in T1D risk* was demonstrated by a hierarchy for DR4 DQB1*03:02 haplotypes in various studies. Using a large cross ethnic meta-analysis, Thomson et al. (2007a) showed the following significant risk effects DRB1*04:05 = *04:01 = *04:02 > *04:04 > *04:03 (= denotes no significant risk difference detected, > denotes a significant risk difference; these are all on DQB1*03:02 haplotypes).

Studies in non-Caucasian populations were also able to tease apart the primary effects of DRB1 in multiple sclerosis (Oksenberg et al. 2004) and DQB1 in narcolepsy (Mignot et al. 2001); in both multiple sclerosis and narcolepsy, the DRB1*15:01 DQB1*06:02 haplotype is predisposing in Caucasians, and of interest, as discussed above, this is a highly protective haplotype in T1D.

B. HLA nomenclature, ambiguity reduction, and population data analyses

HLA typing methods and nomenclature have evolved over time: the recent revised nomenclature uses four colon-delimited fields, allowing more than 2 digits of variation (i.e., removing the limitation of only up to 99 variants) per field (IMGT/HLA database release 3.0.0 April 1, 2010, and release 3.5.0 July 14, 2011: http://www.ebi.ac.uk/imgt/hla/). Different typing methods may not detect the same sets of alleles, nor might they resolve *allele and genotype ambiguities* in the same way, and hence could give different results for the same sample. Because of this the Immunogenomics Data Analysis Working Group (IDAWG) (www.immunogenomics.org),), an international collaboration of investigators working in various aspects of immunogenomics, has been formed. The goals of IDAWG are to facilitate the sharing of immunogenomic data, e.g., HLA and KIR, and to foster the consistent analysis and interpretation of those data by the immunogenomics community and the larger genomics communities (see www. .immunogenomics.org, the "Proposal for HLA Data Validation" at https://www.immport.org, and Gourraud et al. (2011) and Hollenbach et al. (2011b). We encourage researchers to format and archive their data in accordance with the standards proposed by the IDAWG.

IDAWG has partnered with the BISC (Bioinformatics Integration Support Contract) to develop the "HLA Silver Standard" for HLA genotype data collection. While any data analyzed are still subject to errors based on the set of rules used to make "calls," the process used to make the calls (and any inherent biases) would now be transparent. This is particularly important for developing appropriate *binning rules* for studies with data that are heterogeneous with respect to time, and/or meta-analyses of data. An algorithm for ambiguity reduction using a set of rules developed by Steven J. Mack and Richard Single, based on "common and well documented" (CWD) alleles (Cano et al. 2007) is part of the BISC HLA data submission (see "HLA

Guidelines" and "Proposal for HLA Data Validation") (https://www.immport.org, www.immunogenomics.org).

Spurred by the recent new HLA nomenclature adoption, current software developed by IDAWG includes two allele name translation applications available either for download (The Allele Name Translation Tool, or ANTT) or use over the internet (Update NomenCLature, or UNCL), both of which translate the alleles in an entire dataset. Current IDAWG projects include the development of a biostatistical framework for the integrative analyses of HLA and KIR data (as well as data from any highly polymorphic genomic region). This framework will eventually succeed the existing PyPop framework described below. However, until then, the PyPop modules will be appropriately modified to take account of larger data set sizes if necessary (e.g. PyPop is currently limited to 5,000 individuals for the estimation of haplotypes and calculation of LD values), as well as the new nomenclature.

PyPop (Python for Population Genomics) (www.pypop.org, current release version 0.7.0.) (Lancaster et al. 2003, 2007a, b) is a software framework for analyzing large-scale multi-locus genomic population level data. It can also be applied to patient data with the caveat that the haplotype estimation module assumes HWP; when there is a strong disease association this assumption may not apply for patient data, except for a recessive disease model. Although applicable to any genomic region, PyPop was developed specifically for the analysis of highly polymorphic HLA data starting with the 13th International Histocompatibility Workshop in 2003 and continuing, as available software packages were at that time, and many still are, inadequate for HLA (and KIR) data analyses. PyPop is also integrated into ImmPort, the Immunology Database and Analysis Portal (https://www.immport.org).

PyPop is open source, cross platform, modular, and facilitates meta-analyses. The modules include: (1) Allele counts (including a filter module to ensure that allele names are valid, and "binning" rules, that can be modified by the user, to allow across population analyses), (2) Hardy Weinberg overall testing of genotype data and individual genotype level tests [in an upcoming release tests of all heterozygotes/homozygotes, and all heterozygotes for a specific allele will be included, and in future work, the issue of how to handle cut-offs for rare alleles and genotypes (e.g., identifying or excluding the p-values of tests that are not biologically reasonable), will be addressed], (3) Neutrality tests of the allele frequency distributions, (4) Haplotype and LD estimation, and measures and testing of significance of LD, for specified locus combinations (pairwise, triples, and other combinations of up to eight loci at a time), (5) Allele to amino acid translation and amino acid level analyses of LD, Hardy Weinberg proportions and neutrality. (Note that PyPop also accepts sequence data as input.)

HLA allele and haplotype frequencies vary across populations, geographic regions and ethnic groups (Meyer and Thomson 2001, Meyer et al. 2006, Meyer et al. 2007, Single et al. 2007b, Solberg et al. 2008). These data are important regarding potential heterogeneity in disease studies. Population level allele-frequency data and HLA allele frequency maps are available from a recent comprehensive meta-analysis of HLA population data (Solberg et al. 2008) is available at www.pypop.org/popdata).

Since common HLA alleles are directly involved in disease, study of population level variation complements disease studies. Differences in disease prevalence for T1D between ethnic groups correlate strongly with their known differential frequencies of high, intermediate, and low risk HLA DR-DQ haplotypes and genotypes (Valdes et al. 1997). Studying evidence of selection and other features of the evolutionary history of a genetic locus and region involved in disease is an important corollary to all HLA disease association studies.

C. Family data and controls (AFBACs)

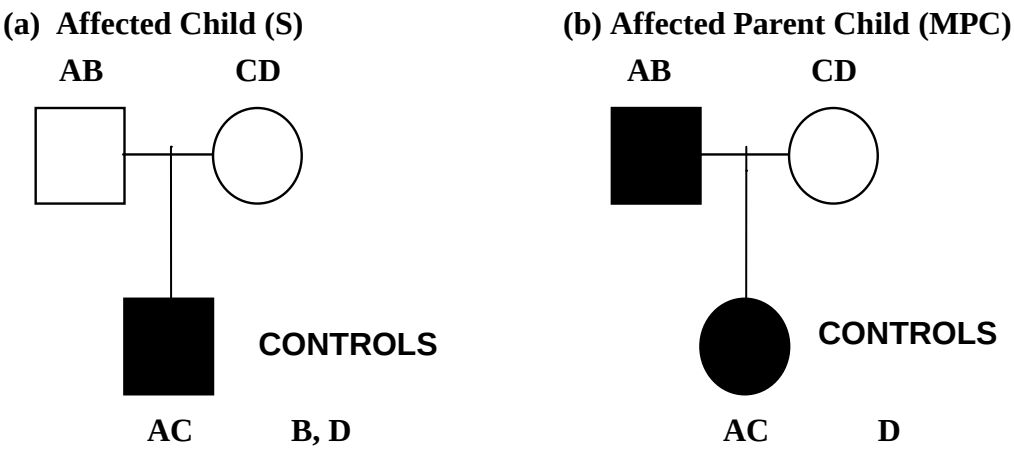
Control allele frequencies with nuclear family data

The use of nuclear family data (two parents and their children) to estimate control marker allele frequencies (and similarly haplotype frequencies) was introduced, with application to HLA data, by Rubenstein et al. (1981), Field et al. (1986), and Falk and Rubenstein (1987). In families ascertained for the presence of at least one affected child, termed S (simplex) pedigrees (also referred to as *trio* families), the two parental marker alleles *not* transmitted to the affected child (the proband) are used as population (control) alleles (see Figure II.C.1a). These are referred to as AFBACs (affected family based controls). This matched design for patient (parental transmitted) and "control" (parental non-transmitted) marker alleles avoids ethnic confounding in the case of a stratified population (Khoury 1994, Schaid and Sommer 1993, 1994).

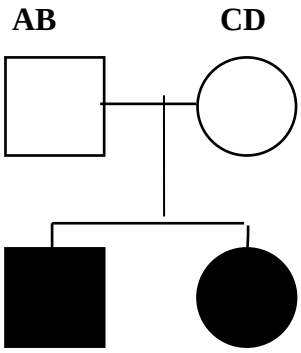
Field (1989) and Thomson et al. (1989) extended this AFBAC approach to nuclear pedigrees ascertained for the presence of at least two affected sibs, termed MSP (multiplex sib pairs) pedigrees (although referred to then as MS pedigrees, for multiplex sibs); using the alleles from both parents that are *never* transmitted to either sib in the affected sib pair as the "control" population (Figure II.C.1c). Using a novel ordered notation, Thomson (1995a, b) provided a complete theoretical analysis of nuclear family based marker allele distributions. Using a single gene disease model with random mating, the validity of the results outlined above for the AFBAC method for S (simplex) and MSP (multiplex sib pairs) pedigrees were confirmed.

Additionally MPC (multiplex parent-child) and MPS (multiplex parent-sib pairs) ascertainment schemes were considered: these extend the S and MSP ascertainment schemes to also include at least one affected parent. The non- or never-transmitted alleles respectively from the *unaffected* parent form the respective AFBAC populations (Figures II.C.1b and d). Similar extensions apply to other ascertainment schemes, e.g., with ascertainment based on the presence of at least three affected sibs in a nuclear family, the AFBACs are the parental alleles *never* transmitted to any of the three affected sibs (Payami et al. 1985).

Figure II.C.1: Affected Family Based Controls (AFBACs) for Four Ascertainment Schemes

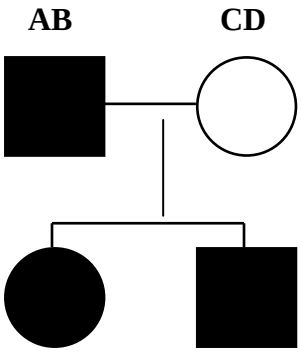


(c) Affected Sib Pairs (MSP)



		CONTROLS
AC	AC	B, D
	AD	B -
	BC	- D
	BD	- -

(d) Affected Parent Sibs (MPS)



		CONTROLS
AC	AC	D
	AD	-
	BC	D
	BD	-

* The four family based ascertainment schemes are selected on the presence of at least: (a) one affected child (S), (b) one affected child and one affected parent (MPC), (c) two affected sibs (MSP), and (d) two affected sibs and one affected parent (MPS). As noted in the text, additional family members (parent or sib) may be affected within each of these schemes; they occur with frequencies expected under the disease model and are retained in analyses. For the MSP and MPS ascertainment schemes, using the ordered notation of Thomson (1995a) the first affected sib ascertained is always denoted by genotype AC, with the four possible genotypes for the second affected sib listed.

AFBAC equations

The equations below are from Thomson (1995a, b). They apply to the case of a marker allele sufficiently closely linked to the disease predisposing allele that we can assume the recombination fraction between them is zero ($\theta = 0$). (The equations in the case of a non-zero recombination value are also given in these papers.) The subscripts i and j refer to the two paternal alleles (AB), ordered such that allele i (A) is transmitted to the affected sib, or first affected sib, in each pedigree type (see Figure II.C.1a), and similarly for k and l for the two maternal alleles (CD). For affected sib pair pedigrees (MSP and MPS), we denote the four possible genotypes (AC, AD, BC, BD) of the second affected sib in relation to that (AC) of the first affected sib, with parameters that reflect their degree of sharing identical by descent (IBD) of the two parental alleles with the first affected sib: sharing both parental alleles IBD (denoted δ_{1010}) (both affected sibs share the genotype AC), sharing the paternal, but not maternal alleles IBD (δ_{1001}) (the first affected sib has genotype AC, and the 2nd affected sib has the genotype AD), sharing the maternal, but not paternal alleles IBD (δ_{0110}) (AC and BC), and sharing no parental alleles IBD (δ_{0101}) (AC and BD). The more common notation combines the two share 1 categories, and denotes share 2 parental alleles IBD by X, share 1 by Y, and share 0 by Z. In the absence of a disease association: $\delta_{1010} = \delta_{1001} = \delta_{0110} = \delta_{0101} = ¼$, i.e., X = ¼, Y = ½, Z = ¼.

For simplex (S), multiplex parent-child (MPC), multiplex sib pairs (MSP), and multiplex parent-sib pairs (MPS) pedigrees the following equations hold in the general case, assuming a random mating population and $\theta = 0$. The equations for the MPC and MPS pedigrees relate to the

cases where the father is the affected parent, with appropriate modification if the mother is affected.

$S_{ijkl} = p_j\, p_l\, S_{i+k+}$,

$MPC_{ijkl} = p_l\, S_{ijk+}$,

$MSP_{ijkl}\, \delta_{1010} = p_j\, p_l\, MSP_{i+k+}\, \delta_{1010}$,
$MSP_{ijkl}\, \delta_{1001} = p_j\, MSP_{i+kl}\, \delta_{1001}$,
$MSP_{ijkl}\, \delta_{0110} = p_l\, MSP_{ijk+}\, \delta_{0110}$,
$MSP_{ijkl}\, \delta_{0101}$ does not simplify,

$MPS_{ijkl}\, \delta_{1010} = p_l\, MPS_{ijk+}\, \delta_{1010}$,
$MPS_{ijkl}\, \delta_{1001}$ does not simplify,
$MPS_{ijkl}\, \delta_{0110} = p_l\, MPS_{ijk+}\, \delta_{0110}$,
$MPS_{ijkl}\, \delta_{0101}$ does not simplify.

In S (simplex) pedigrees, with $\theta = 0$, the two AFBAC non-transmitted parental alleles from each pedigree thus give unbiased estimates of *population* (control) marker-allele frequencies. Similarly, with appropriate modifications, for the other pedigree types: for MSP (affected sib pair) pedigrees the AFBACs are the parental alleles *never*-transmitted to either affected sib. For MPC and MPS pedigrees (multiplex parent-child and multiplex parent-sib pairs respectively), with at least one affected parent, the same principles apply, i.e., the non- or never-transmitted alleles respectively from the *unaffected* parent form the AFBACs.

Application of the AFBAC method

For both S (simplex) and MSP (affected sib pair) ascertainment schemes there will be some pedigrees with a parent affected, and, although the pedigrees will then look like those obtained through MPC (at least one affected parent and one affected child) and MPS (at least one affected parent and two affected sibs) ascertainment criteria, respectively, they are still used in the S and MSP categories for analyses (provided the ascertainment scheme was strictly adhered to), i.e., all parents contribute appropriately to the AFBACs. Similarly, with pedigrees from S and MPC ascertainment, there will in some families be a second affected sib (or third or more), again, if the ascertainment was strictly based on the presence of at least one affected child (the proband), then such pedigrees are retained under the S and MPC ascertainment schemes, and these additional affected children (or unaffected children) *do not enter* into the AFBAC calculations. This is because the affected parent, or additional affected child, or both, occur in the frequencies expected under the ascertainment scheme applied, and are accounted for in the expectations.

Data will rarely be specifically sampled for MPC and even less so for MPS pedigrees. However, these data can be extracted without bias from S and MSP data sets for additional

specific analyses, including unique information. Specific tests using the different pedigree types are detailed in Thomson (1995a, b).

Control genotype frequencies for S and MS pedigrees can be obtained in two ways. The first is by constructing genotypes using the paternal and maternal non-transmitted alleles in each S pedigree (2n total), i.e., the genotypes denoted BD in Figure II.C.1a, and similarly in MSP pedigrees but now only using those pedigrees where both affected sibs share the same two parental alleles, i.e., BD from the sibs in row 1 who share both parental alleles IBD (δ_{1010}) (Figure II.C.1c). The usual Hardy Weinberg proportions (HWP) test can be applied to these control genotypes. A maternal-fetal interaction effect could cause deviations from HWP. *Alternatively*, one can take allele (or haplotype) AFBACs and form a control population assuming HWP (with a sample size of half the number of AFBACs).

Note that for transmitted alleles in MSP pedigrees, the two affected sibs are *not independent*, and cannot be included as independent data points in any analyses. The average of the two sibling's genotypes is often used in analyses.

For S pedigrees, application of the AFBAC method is simple even in cases where the markers are *not* highly polymorphic; subtracting the allele counts (2n, where n is the number of families) of the proband (affected child) from the total allele counts from the two parents (4n) gives the AFBAC (2n) allele counts (Barcellos et al. 1997). In other ascertainment schemes, the highly polymorphic nature of HLA data makes their application feasible to calculate the AFBACs, and the more HLA genes typed the less any bias. If some parental typings are missing, the non-transmitted alleles from this parent can often be calculated if additional children have been typed, this however will be a *biased estimate* resulting from the fact that homozygous parents cannot be distinguished. Extreme caution must be applied in such applications, and it is not necessarily recommended at all.

The AFBAC method has often mistakenly been equated to a heterogeneity test of association of patient and AFBAC frequencies. The AFBAC method per se is for obtaining an unbiased estimate of marker allele frequencies, and does not refer to any specific test of association/linkage. The transmission disequilibrium test (TDT) of Spielman et al. (1993) is the appropriate statistical test to apply, since it tests for deviations from the null hypothesis expected 50:50 ratio of transmitted and non-transmitted alleles from parents heterozygous for the marker alleles, and theoretically is robust to stratification effects. However, note that in the case of a homogenous population sample, the results from a TDT and contingency table analysis will be very similar.

For extended pedigree data, the pedigree disequilibrium test (PDT) can be applied which also includes disease discordant sibships (Martin et al. 2000, 2003, and see Barcellos et al. 2006 for application to multiple sclerosis data). All family genotypes should be examined for Mendelian inconsistencies using for example PEDCHECK (O'Connell and Weeks 1998).

D. Tests of marker associations with disease

Case/control association studies

A major concern with all disease association studies is heterogeneity between cases and controls due to allele frequency differences related to population stratification. If such heterogeneity is not taken into account, it may be mistakenly interpreted as association of a marker locus (loci) with

disease. This was less problematic with previous studies of classical HLA locus associations with disease, where studies were usually limited to an ethnic group within a geographic location. But it is particularly pertinent to large scale studies, e.g., GWASs, given the known heterogeneity in for example Caucasians across Europe for HLA frequencies (see Solberg et al. 2008) and also some SNPs. Population substructure, e.g. for Caucasians north/south European, can be detected using principal components analysis (PCA) in for example EIGENSTRAT (Price et al. 2006) or using ancestry informative markers (AIMs) (see for example Pritchard et al. 2000, Seldin and Price 2008), and outliers removed from further study. The major PCs from the remaining data can be incorporated as covariates in the association analyses, potentially increasing power and avoiding stratification effects.

Analyses of HLA association case/control data have until recently used mostly standard chi-square tests of heterogeneity. Initial studies, with serological typing and a high frequency of blank alleles at the classical HLA loci, were analyzed based on presence versus absence of a specific allele, since homozygotes versus heterozygotes for a blank allele could not be distinguished. With molecular typing, this issue was mainly removed, and the first test applied would be an overall *test of heterogeneity of all alleles* between cases and controls (with appropriate binning of rare alleles). Individual allele contributions to the significance of an overall test, combined with consideration of ORs, and the frequencies of the alleles (which influence significance levels) are the first step in detecting all disease risk heterogeneity at the associated locus. See Tables III.B.1 and III.C.1 later for examples. (Note that the test of each individual allele's contribution to the overall significance are based on a chi-square with 1 df; these individual p-values are biased as the assumption of a 1 df chi-square is incorrect, and conservative, the p-values can be used however for a relative ranking of the allelic effects.)

Association analyses using logistic regression (LR) are now also used in case/control studies. Logistic regression is a form of *generalized linear modeling* for data with a dichotomous, or binary, outcome variable, such as case/control data where the outcome is either 'affected' or 'unaffected.' Among its advantages, logistic regression provides a means to develop association models that include the contribution of both quantitative and qualitative covariates, which can involve cumbersome stratification procedures in contingency table heterogeneity testing; logistic regression is a critical tool in the analysis of complex multivariate datasets.

However, attention must be paid to how both quantitative and qualitative variables are coded, because interpretation of the ORs can greatly vary (Hollenbach et al. 2011a). A disadvantage of logistic regression analysis is a tendency to overestimate ORs (Nemes et al. 2009) in sample sizes <500, possibly leading to erroneous conclusions. Therefore, as with any statistical method, care must be taken in both the application and interpretation of results in logistic regression analyses. In addition, it is important that the data be reported in a manner, for example a table, which allows readers to reanalyze the data.

Logistic regression analysis is always performed using computer software. While many commercial (e.g., SAS, SPSS, STATA) and free (e.g. PLINK, see Purcell et al. 2007) software packages can be used for logistic regression, many are not suitable for polymorphic HLA (or KIR) data. For example PLINK performs logistic regression for SNP or CNV (copy number variant) data and will not handle the level of polymorphism typical of HLA data. While HLA data may be recoded for conditional analysis of SNP or CNV data in PLINK, this will depend on the specific research question, and except for cases with only one HLA allele associated with disease (versus the rest of the alleles at that locus having homogeneous relative risk effects), this is not recommended. There are also several packages and standard functions for the R language

(e.g., the 'glm' function in the base package) that can perform logistic regression, but extreme care must again be taken in the coding of HLA data, as many of these functions are not designed to handle high levels of polymorphism.

A regression analysis produces a model that includes all of the variables that are useful in predicting the (dichotomous) outcome variable. Logistic regression provides an OR for each variable involved in predicting outcome, and can be particularly useful in HLA studies that must account for a number of cofactors in assessing associations with complex diseases. Several options are available in building the most parsimonious model, i.e., the one explaining the maximum of the variance with the minimal number of variables. A *backward stepwise approach* can be particularly productive in an exploratory analysis; the initial (full model) starts with inclusion of all candidate variables; these are tested one by one for statistical significance, deleting any that are not significant. This procedure allows the investigator to eliminate noncontributing variables and build a more parsimonious model. A *forward stepwise approach* can be more robust when large numbers of variables are studied. In this procedure, the most significant variables are successively added to the model, until the addition of variables does not contribute to a significant increase in the variance explained by the model.

Family based association studies

Ott (1989), Knapp et al. (1993), Spielman et al. (1993), and Thomson (1995a, b) demonstrated that, with nuclear-family based data, tests of association are confounded with tests of linkage. No matter which test is used, e.g., a contingency table test of heterogeneity, the haplotype relative risk (HRR) (Ott 1989), the TDT (Spielman et al. 1993), logistic regression (LR) (using probands and AFBACs), the property that only associations of marker genes linked to a disease gene ($\theta < 0.5$) will be detected holds for all family-based tests. This is in contrast to case/control association studies where population stratification can lead to false association results even with unlinked genes. However, an assumption of random mating is required for all tests except the TDT; hence it is the preferred test.

E. Stratified association tests, age of onset, and maternal-fetal effects

Use of allele sharing IBD values can increase the power of association studies (Fingerlin et al. 2004, Thomson 1995b). Vice versa, partitioning linkage analyses by genotypes of associated alleles can also increase power (Clerget-Darpoux et al. 1995; Greenberg 1993; Greenberg and Doneshka 1996; Hodge 1993; Li et al. 2004). With T1D, IDDM2 (the VNTR 5' to the insulin gene), and study of affected sib pair data, Dizier et al. (1994) show that without stratification, the IBD frequencies are 0.24, 0.53, and 0.23 and these are not significant from random expectations. However, the distribution is quite different when stratified by the genotype of the index sib: 0.27, 0.60, and 0.13 for genotype 11, and 0.07, 0.53, and 0.40 for the combined other genotypes. Li et al. (2004) have also shown that there can be large variability in linkage scores when families are stratified by the genotype of a single, randomly selected sib.

Age of onset effects may be important for both primary and secondary disease gene effects, both for HLA and non-HLA regions [for HLA and T1D see e.g., Valdes et al. (1999, 2005a); for Alzheimer's disease and APOE see Blacker et al. (1997); for Alzheimer's disease and HLA, see Zareparsi et al. (2002)]. The power of the TDT and case/control association tests can be greatly affected by the age of onset of patients in the study (Li and Hsu 2000). Further, age-

related fluctuations in allele and genotype frequencies in controls can lead to loss of power and increase in type 1 error rates; Payami et al. (2005) show that APOE frequencies in controls can vary by twofold in the extreme cases due to age and gender. For cytochrome P450 2D6 (CYP2D6) the age and gender-specific fluctuations in controls were less pronounced than for APOE and non significant; however, there were significant age-dependent departures from HW, even for the youngest cohort. Further, gender differences should also be investigated.

As for association studies, age of onset effects which are not incorporated into a linkage analysis can drastically reduce power (Hsu et al. 2002; Li 1999; Li and Hsu 2000). The true disease model (unknown for complex diseases), especially common variant versus rare variants with extensive allelic heterogeneity, will influence whether association or linkage studies are more powerful (Pritchard 2001; Risch and Merikangas 1996)

For family data, we have discussed above (Section II.D) comparison of non-transmitted alleles (AFBACs) from mothers (termed NIMAs—non-inherited maternal antigens) versus fathers (NIPAs—non-inherited paternal antigens) overall and based on specific genotypes of the proband. Maternal offspring compatibility (see Bronson et al. 2009), parent of origin and NIMA effects in autoimmune diseases have been seen, but are not well replicated. The exposure to NIMA via several different mechanisms may shape the immune system of the offspring and either predispose or protect against immune reactions including in *utero* exposure to NIMA, as well as postpartum exposure to NIMA mediated by breast-feeding and/or long-term persistence of maternal cells in the offspring. With HLA associated diseases NIMA effects should be tested for in the overall data as well as in subsets of patients with and without high risk genotypes, and for specific genotypes. NIMA effects were first reported in rheumatoid arthritis in mothers of DR4-negative patients (ten Wolde et al. 1993), were not seen in the study of Silman et al. (1995), but were again reported by Van der Horst-Bruinsma et al. (1998) and Harney et al. (2003). In T1D the results are not consistent, with some reports of NIMA effects, e.g., see Pani et al. (2002), and other studies not showing this effect (Hermann et al. 2003, Lambert et al. 2003, Bronson et al. 2009).

Genomic imprinting has been implicated in susceptibility to complex diseases such as systemic lupus erythematosus and other HLA-associated autoimmune conditions. Genomic imprinting is defined as a phenomenon in which the disease phenotype depends on which parent passed on the disease predisposing gene. While much remains to be learned about the underlying epigenetic mechanisms involved in imprinting, it has been shown to play a role in several birth defects, certain genetic diseases, and cancers, and possibly autoimmunity. Differences in maternal and paternal transmission rates of predisposing alleles have been seen in several studies of autoimmune diseases including T1D and systemic lupus erythematosus (Bennett et al. 1996, Fajardy et al. 2002, Sasaki et al. 1999, Sekigawa et al. 2003).

F. Linkage disequilibrium

If a *primary* disease gene in the region under study has been identified, the first step in study of additional disease genes in the region would be to plot the LD of all marker genes *with* the *primary* disease gene. Informal inspection would show in most cases a strong correlation of this LD pattern with the strength of association of markers with disease (before any stratification analyses have been applied and assuming the primary disease gene has been correctly identified). Obvious deviations from this pattern would indicate potential *additional* disease genes (or

markers in strong LD with *additional* disease genes). (If multiple genes have a *primary* effect, these genes are combined together as one "super-locus" combination; haplotypes would need to be estimated for some analyses to form the "alleles" at this "super-locus.")

What measure of LD should we use when studying marker associations with each other? There is no one measure which is perfect, as we are trying to depict multidimensional variables (allele frequencies and numbers of alleles at each locus, and the associations (LD) between all pairwise combinations of alleles at the two loci) by one overall LD measure. For multiallelic markers appropriate extensions of two biallelic measures are often used: the D' statistic is a weighted average of the normalized LD statistic between each pair of alleles (Hedrick 1987), and the W_n statistic which is a re-expression of the X^2 statistic normalized to be between zero and one (see e.g., Meyer et al. 2006, Single et al. 2007c); the latter measure is more informative in most cases with respect to the analyses in this section. When there are only two alleles per locus, W_n is equivalent to the correlation coefficient between the two loci, defined as $r = D/[p_A p_a p_B p_b]^{1/2}$, where p_A,p_a and p_B,p_b are the allele frequencies at the two loci (*A* and *B*) respectively and *D* is the standard measure of LD between the two loci.

A measure called haplotype specific heterozygosity captures the informativeness with respect to association studies of markers on specific haplotypes of the primary disease predisposing gene (Malkki et al. 2005). The LD of all markers with each other should also be studied and will guide single association versus haplotype analyses. LD and haplotype blocks in the HLA region vary based on specific HLA haplotypes (Ahmad et al. 2003, Blomhoff et al. 2006).

We have recently developed a complementary pair of *asymmetric* measures of the strength of pairwise LD for multi-allelic data: these are called conditional linkage disequilibrium (CLD) measures. These more accurately reflect the independence or lack of independence for genetic variation at two loci than do standard LD measures. For the bi-allelic case they are symmetric and equivalent to the correlation coefficient *r* (most often reported as r^2 as described above). These new CLD measures are particularly relevant to disease association studies: to more accurately determine when stratification analyses can be applied to detect primary (major) disease predisposing genes, as well as to identify additional disease genes in a genetic region. They are also applicable to the study of evolutionary forces such as selection acting on individual amino acids of specific genes, or other loci in high LD. The measures can be applied to variation at any pair of loci (HLA and other genes, SNP data, MSAT data, and haplotypes thereof, as well as biologically relevant sequence features (SFs) (Karp et al. 2010) based on structural and functional features of a protein). With SNPs it is recommended for analysis of haplotype block data, both for block-block comparisons of LD patterns, and for block to HLA (or other primary disease locus) data. A manuscript on this work is in preparation, and also see Single et al. (2011).

No LD measure completely captures all pertinent features of the data. Thus, we always recommend consideration of other complementary summary measures of the strength and structure of LD in multi-allelic data, and also visualization of the LD structure (Barrett et al. 2005). For SNP data, methods have been developed to consider single marker versus haplotype associations with disease (Browning 2006, Browning et al. 2005, Morris 2006, Purcell et al. 2007).

Study of the break down of LD structure, e.g., using the extended haplotype homozygosity (EHH) (Sabeti et al. 2002, Voight et al. 2006) can be used to detect selection across the genome, and is of particular interest with respect to variants implicated in disease. A number of other tests listed in Sabeti et al. (2002) can also be applied to detect selection; power

will vary as well as sensitivity to misspecification of the parameters, e.g., the frequency of undetected alleles. Tests of selection using LD patterns are also available (Thomson and Klitz 1987, Klitz and Thomson 1987, Robinson et al. 1991a, 1991b, Grote et al. 1988). We refer readers to three recent reviews: Meyer and Thomson (2001), Harris and Meyer (2006), and Hedrick (2006).

III. Primary Disease Genes: Modes of Inheritance

A. The patient/control (P/C) ratio and estimating relative penetrances

In this section, we assume that a primary disease gene has been identified, and introduce and discuss the patient/control (P/C) ratio (Thomson et al. 2007a, 2008). The P/C ratio is defined as follows, under the assumption for example that the HLA DR-DQ genes are directly involved in disease, as in T1D:

$$P/C = [\text{freq(DR-DQ)}_{\text{patients}}] / [\text{freq(DR-DQ)}_{\text{controls}}],$$

where freq(.) denotes the frequency of the haplotype or genotype under consideration (or allele or SNP or HLA amino acid as appropriate, and multiple combinations thereof) also see Clerget-Darpoux et al. (1988). The control population is assumed to be a random population level sample as this simplifies the parameter estimates. Although this is usually *not* the case, i.e., controls are usually selected based on absence of disease, nonetheless this will not drastically alter any results for diseases which are relatively rare, which is the case with most HLA associated diseases.

Under the assumption that the primary disease gene has been identified, the P/C ratios give a maximum likelihood estimate (MLE) of the relative penetrance values for each genotype ($f_{ij} = w_{ij} /T$, where w_{ij} is the absolute penetrance value and T is the disease prevalence), and allele (or haplotype as appropriate) ($f_{i.} = w_{i.} /T$) (Thomson et al. 2007a, 2008) (Table III.A.1). Unfortunately, the P/C ratios per se cannot be directly compared across populations, and particularly across ethnic groups, given they are a function of disease prevalence. However, within a population, the ratio for two different genotypes or haplotypes of their P/C ratios is a function only of the absolute penetrance values, allowing comparisons across studies and ethnic groups, i.e.,

$$[P/C\ A_iA_j] / [P/C\ A_kA_l] = f_{ij} / f_{kl} = w_{ij} / w_{kl},\ ij \neq kl, \text{ and}$$

$$[P/C\ A_i] / [P/C\ A_j] = f_{i.} / f_{j.} = w_{i.} / w_{j.},\ i \neq j.$$

The population prevalence parameter T *cancels* from the relative penetrance estimates in this case (see Table III.A.1).

Table III.A.1: Patient and control frequencies and P/C ratios

	Genotypes	Alleles or Haplotypes
	A_iA_j	A_i
Controls:	$f(A_iA_j)$	$f(A_i)$
Patient population [a]:	$w_{ij}\ f(A_iA_j)\ /T$	$w_{i.}\ f(A_i)\ /T$
Patient/control (P/C):	$w_{ij}\ /T$	$w_{i.}\ /T$

[a] $T = \Sigma w_{ij}\ f(A_iA_j)$ is the disease prevalence.

Under the assumption that the primary disease gene has been identified, the P/C ratios give a maximum likelihood estimate of the *relative* penetrance values for each genotype (these are a function of disease prevalence) (Thomson et al. 2007a, 2008). However, the ratio of two P/C ratios estimates the ratio of the *absolute* penetrance values for the two haplotypes or genotypes under consideration. Further, it is equivalent to the OR for the comparison of these two alleles, haplotypes or genotypes. These ratios of absolute penetrance values may *vary between* populations because they include the averaged effects of non-HLA genes, environmental factors, and other HLA genes.

In the simplest case the ratio of the absolute penetrance values for the DR-DQ genotypes and haplotypes, or other primary disease gene(s), would be the same across all studies. Further, the relative rankings based on P/C ratios for a set of alleles, haplotypes or genotypes can be compared across populations (see Thomson et al. 2007a, 2008). One reason for consideration of the P/C ratio is that when there is a complex hierarchy of predisposing through protective effects of alleles, haplotypes, or genotypes, the ORs are then much more complex to interpret than the P/C ratios, in that the denominator (comparison group) of the OR calculation contains a mixture of alleles with differential risk effects. This is avoided with use of the P/C ratio within a study.

As emphasized in Section I, unless we fully understand, and stratify by, *all* heterogeneity in disease risk at the *primary* disease gene, including relatively weak effects, then application of methods to detect *secondary* disease genes in the region may give spurious results. Also, we need to detect *all* heterogeneity in disease risk at the allele, haplotype, and genotype levels, to optimally apply methods to detect the amino acids directly involved in disease risk heterogeneity when one or more classical HLA genes is (are) primary. For a primary gene, estimation of relative penetrance values from the P/C ratios for alleles, haplotypes, genotypes, and amino acid variation, form the base of our study of modes of inheritance and identification of significant differential risk categories. Various analysis methods are listed below regarding appropriate, and in some cases, complementary statistical tests to consider.

B. Detecting relative predispositional risk effects (RPEs) and T1D data

In this Section we discuss the *relative predispositional effects* (RPE) method and in the next Section (III.C) we consider tests of risk heterogeneity using all pairwise comparisons, with

examples from T1D (Noble et al. 1996), and a subset of JIA patients classified as oligoarticular-persistent (OP) (Hollenbach et al. 2010, Thomson et al. 2010). Both methods identify significant risk heterogeneity of alleles at a *primary* disease gene, and we recommend application of both methods, along with considerations of sample size (hence significance of effects - less frequent classes even with stronger effects can contribute less to the overall chi-square (see Thomson et al. 2008)) combined with use of ranking by OR and P/C ratios, and meta-analysis results.

The RPE method (Payami et al. 1989) identifies heterogeneity in disease risk at the *primary* disease gene; common alleles, haplotypes, or genotypes with the strongest predisposing or protective effects are *sequentially* removed from the analysis until no further heterogeneity in risk effects is seen. When, for example, a disease has a strong association (predisposing) with one allele, this allows us to determine if the decrease in frequency of other alleles is the *expected consequence* of the increased frequency of the first allele, or if there is a true negative (protective) association of some alleles, and vice versa if one allele shows a strong protective effect. Payami et al. (1989) showed that after taking account of the DR3 association ($p < 0.00001$) with Graves disease (which was well established also from other studies), a significant negative (protective) effect was seen for DR5 ($p < 0.0001$) after removal of the DR3 effect, and the remaining alleles showed no significant risk heterogeneity. Similarly, a second positive association may be masked by a major strong association with one allele. Also with application of RPE analysis, so-called "neutral" (intermediate) effects which are characterized by similar allele frequencies in patients and controls, may show significant risk differences between predisposing and protective sets of alleles. All these effects are identified with sequential application of the RPE method.

Determining the order in which alleles, haplotypes or genotypes are sequentially removed is not trivial, and requires interplay between the individual contribution to the Chi-square heterogeneity test, the ORs or Patient/Control (P/C) ratio (as mentioned above, the OR and P/C ratio are often close in value), and the control frequencies. The sample size for each allele is a factor in this analysis; more common alleles (predisposing, neutral, and protective) are identified with this method. Also, for the same control frequency and equivalent strength of effects, a positive association will contribute more to the overall Chi-square than a negative association. Also, less frequent classes even with much stronger predisposing or protective effects can contribute less to the overall Chi-square. The investigator must weigh these different issues in application of the RPE method.

As previously mentioned, T1D shows the most complex hierarchy of RPEs at the allele, haplotype, genotype, and amino acid level, ranging through very predisposing, predisposing, neutral (intermediate), protective, and very protective (for results from a large meta-analysis see Thomson et al. 2007a). A small subset of the most common HLA DR-DQ haplotypes, ranked by their ORs, are given below in Table III.B.1, with the last three columns showing the overall and individual chi-square values in a test of heterogeneity (the third last column), and then after sequential removal of the most significant effects (the last two columns). The predisposing DR3 (for the definition of HLA DR-DQ abbreviations see the footnote to Table III.B.1), and DRB1*04:01 and *04:04 with DQA1*03:01 DQB1*03:02 haplotypes, and the very protective DR15 haplotype are removed first. (The same results apply with separate removal of the predisposing and the protective haplotypes.) The neutral effect of the DR1 haplotype then shows as significant relative to the remaining haplotypes (second last column), and after removal of DR1, the relatively homogenous protective category remains, with slight evidence of heterogeneity (last column). (A large meta-analysis in Thomson et al. (2007a) verifies heterogeneity of this category.) Note that the ranking neutral (intermediate) is relative and

dependent on the presence of predisposing and protective haplotypes (or genotypes) in the population under study. If the predisposing haplotypes listed here are missing or rare in an ethnic group, e.g., Asian, then haplotypes that appear neutral in a Caucasian population with similar frequencies in patients and controls, will appear to be predisposing (relatively) in the Asian population.

Table III.B.1: RPE Analysis of HLA DR-DQ Haplotype Frequencies in Type 1 Diabetes

DRB1	DQA1	DQB1[a]	T1D(%)[b]	Cont's(%)[c]	OR[d]	P/C ratio[e]	X^2 [f]	X^2[g]	X^2[h]
Predisposing									
04:01	03:01	03:02	91 (25.3)	11 (4.0)	8.21***	6.33	29.7****	-	-
04:04	03:01	03:02	38.5 (10.7)	5 (1.8)	6.54***	5.94	12.0***	-	-
03:01	05:01	02:01	115 (31.9)	26 (9.4)	4.55***	3.39	20.4****	-	-
Neutral									
01:01	01:01	05:01	27.5 (7.6)	17 (6.1)	1.27	1.24	0.0	11.2***	-
Protective									
13:02	01:02	06:04	10.5 (2.9)	12 (4.3)	0.67	0.67	2.6	0.8	4.7*
04:01	03:01	03:01	6 (1.7)	12 (4.3)	0.38*	0.40	6.9**	0.1	0.4
07:01	02:01	02:01	6.5 (1.8)	30 (10.8)	0.15***	0.17	32.3****	6.1	1.4
13:01	01:03	06:03	3 (0.8)	18 (6.5)	0.12***	0.12	21.6****	4.8*	1.6
Very protective									
15:01	01:02	06:02	1 (0.3)	43 (15.5)	0.02***	0.02	70.3****	-	
Others			61 (17.0)	104 (37.3)	-	-	-	-	
Total			**360**	**278**			**195.9******	**23.0******	**8.1**

Data from Noble *et al.* (1996) of 180 affected sib pairs with type 1 diabetes.

[a] HLA DRB1 DQA1 DQB1 haplotypes—the following abbreviations are often used: DR3 (DRB1*03:01 DQA1*05:01 DQB1*02:01), DR1 (DRB1*01:01 DQA1*01:01 DQB1*05:01), DR7 (DRB1*07:01 DQA1*02:01 DQB1*02:01), and DR15 (DRB1*15:01 DQA1*01:02 DQB1*06:02); for the three DR4 haplotypes there is heterogeneity at the DRB1 and DQB1 loci, similarly for the two DR13 haplotypes.

[b] Patient counts and (%), patient counts are the average of the two affected sibs; haplotypes listed have a frequency >4% in at least one of the patient or control haplotypes.

[c] Controls are AFBACs (Thomson 1995a, b).

[d] Odds ratio (OR) of this haplotype versus all others: * ($p<0.05$), ** ($p<0.01$), *** ($p<0.001$), **** ($p<0.0001$).

[e] Patient/control (P/C) ratio using frequency data (see Section III.A above)

[f] X^2 contribution of the individual haplotype in the heterogeneity test of patients versus control haplotypes (1 df) (these individual p-values are biased as the assumption of a 1 df chi-square is incorrect, and conservative; the p-values can be used however for a relative ranking of the allelic effects), and the total chi-square (df = number of classes -1) (the 'other' category is not included in the calculations)

[g] X^2 contribution of the individual haplotype in the RPE heterogeneity test of patients versus control haplotypes after removal of the highly protective DR15 haplotype, and the three predisposing haplotypes DRB1*04:01 and 04:04 with DQA1*03:01 DQB1*03:02 and DR3; the overall test and the individual contribution of DR1 are both highly significant.

[h] As for previous column, except DR1 is now additionally removed; the overall test is non-significant, with only marginal significance for one individual haplotype, so no further rounds of testing are carried out.

The P/C ratios for the HLA DR-DQ haplotypes in T1D are given in column 7. In this case, the relative rankings of the HLA DR-DQ haplotypes from most predisposing through most protective are the same based on ORs and P/C ratios. As above, this is not always the case, although there is usually a very strong correlation between the two values. For haplotypes (or genotypes) with a P/C ratio <1, the inverse of this number is used in comparison of relative strength with positively associated haplotypes (P/C ratio >1); hence the DR15 very protective effect is much stronger than the three haplotypes listed with predisposing effects.

Genotype frequencies are given below in Table III.E.3. Note that genotype frequencies rapidly become very small in patients, and are always small in controls. For this reason many analyses focus on DR-DQ haplotype data, and for genotype data often consider comparisons of specific subsets.

In the next Section, we consider all pairwise allele risk comparisons using JIA-OP data. We give the RPE analysis results for JIA-OP in the next Section, and also for illustration include the details of the step by step RPE analysis in Appendix A.

C. All pairwise relative risk comparisons and JIA-OP data

In addition to application of the RPE method, we strongly urge statistical testing of all pairwise allele (or haplotype or genotype) comparisons for heterogeneity. This complements the RPE analyses above (Section III.B), and clearly indicates the extent, and possible complexity, of disease risk heterogeneity. Some HLA associated diseases show considerable heterogeneity in disease risk at the primary gene, for example, T1D as shown above, which at present is the most extreme example. Other diseases also show a hierarchy of predisposing, intermediate, and protective risk effects, for example JIA-OP as shown in Table III.C.1 below and Appendix A. In contrast, some diseases show only one major association, e.g., B27 and ankylosing spondilitis, and the DRB1 "shared epitope" and rheumatoid arthritis (Gregerson et al. 1987)—although there is risk heterogeneity within the "shared epitope" category of DRB1 alleles— and additional risk heterogeneity may be found with analyses of larger sample sizes and high resolution molecular typing.

Multiple sclerosis has a strong DRB1*15:01 association in Caucasians and initially the remaining DRB1 alleles were thought to be homogenous in risk. However, a large study of ~1300 multiple sclerosis families by Barcellos et al. (2006) validated the role of additional DRB1 alleles and genotypes in disease risk using a variety of methods including CLR. DRB1*15 was strongly associated with disease (p = 7.8E-31), and a dose effect was shown with an OR of 9.8 for DRB1*15/15. The heterozygote DRB1*15/08 was high risk (OR = 7.7) while DRB1*15/14 was low risk (OR = 1.9); the ORs for the remaining heterozygotes ranged from 3.5 to 5.0. A modest dose effect (recessive) was detected for DRB1*03.

We use JIA-OP data (Hollenbach et al. 2010, Thomson et al. 2010) and show the RPE analysis results of the common DRB1 alleles (Table III.C.1) and also application of all pairwise risk comparisons (Table III.C.2, and Appendix B). The details of each round of application of the RPE method to JIA-OP data (Hollenbach et al. 2010, Thomson et al. 2010) are given in Appendix A, and the results summarized in column 1 of Table III.C.1 below. (Columns 2 and 3 are described later with respect to the all pairwise comparisons risk analysis.) The color code for the alleles is: predisposing (shaded teal), neutral (intermediate) (shaded light grey), protective

(shaded bright pink). Rare alleles which cannot be placed into one of these categories either by the RPE method or all pairwise comparisons are shaded green.

The RPE analysis of the JIA-OP data set shows considerable overall heterogeneity in risk at DRB1 ($p < 1.1E-27$) (see Appendix A, and Table A.1: DRB1*08:01 (predisposing) and DRB1*15:01 and DRB1*07:01 (protective) are the strongest effects (the set of alleles labeled Category 1 (removed after the first round of analyses) in column 1). Removal of these alleles still gives a highly significant result ($p < 4.1E-10$), with DRB1*11:04 (predisposing) and DRB1*04:01 (protective) as the strongest effects (set labeled 2 in column 1). Note that a strong argument could be made for deleting these alleles also at the previous round, and separately for only removing DRB1*08:01 (predisposing) at the first round, but neither alters the outcome. With removal of these strong effects, there is only minimal evidence of remaining risk heterogeneity ($p < 0.02$), with DRB1*11:03 ($p < 0.01$) (predisposing) and DRB1*01:03 ($p < 0.02$) (protective) the strongest effects (set labeled 3 in column 1 of Table III.C.1). Note that these latter p-values would not be significant with corrections for multiple comparisons of either the overall tests, or those for individual alleles. Notwithstanding, we mention these results, since our principal aim is to detect heterogeneity that may be relevant to detecting additional disease genes in a genetic region or identifying the amino acids in the *primary* disease gene that are directly responsible for the disease risk heterogeneity. These results are all compatible with the heterogeneity testing of all pairwise allele comparisons (see below), and consideration of ORs and P/C ratios (which are similar in these cases).

Table III.C.1: JIA-OP HLA DRB1 allele data ranked by Odds Ratio (OR)

RPE[a]	A[b]	B[c]	DRB1	Pat	Con	Chi-sq.	p-value[d]	OR	CI[e]	CI[e]
		Ix	*11:03	12	1	6.80	0.009	9.40	1.22	72.49
1	I	Ix	*08:01	102	13	48.61	3.1E-12	6.90	3.83	12.43
2	I	Ix	*11:04	57	11	20.71	5.3E-06	4.26	2.21	8.20
			*04:03	9	3	1.68	0.20	2.33	0.63	8.65
3	II	IIx	*13:01	90	38	9.99	0.002	1.95	1.31	2.90
		IIx	*01:02	9	5	0.35	0.55	1.39	0.46	4.18
	II	IIx	*11:01	60	36	1.42	0.23	1.31	0.85	2.02
		IIx	*09:01	9	6	0.08	0.78	1.16	0.41	3.28
	II	IIx	*01:01	74	50	0.52	0.47	1.16	0.79	1.69
	II	IIx	*03:01	89	61	0.50	0.48	1.14	0.81	1.62
		IIx	*12:01	10	8	0.006	0.94	0.96	0.38	2.46
	II	IIx	*13:02	28	23	0.05	0.82	0.94	0.53	1.64
			*13:03	10	9	0.11	0.74	0.86	0.34	2.12
			binned[f]	27	27	0.92	0.34	0.76	0.44	1.31
			*16:01	6	8	1.05	0.30	0.58	0.20	1.67
			*14:01	11	18	4.05	0.04	0.46	0.22	0.99
			*15:02	5	10	3.26	0.07	0.38	0.13	1.12
	III	IIIx	*04:04	7	16	6.34	0.01	0.33	0.14	0.81
1	III	IIIx	*15:01	38	80	28.24	1.1E-07	0.33	0.22	0.49
1	III	IIIx	*07:01	30	65	23.92	1.0E-06	0.33	0.21	0.51
2	III	IIIx	*04:01	21	47	18.10	2.1E-05	0.33	0.19	0.55

3		IIIx	*01:03	4	11	5.42	0.02	0.28	0.09	0.87
			TOTAL	**708**	**546**	**182.1**	**1.1E-27**			

[a] Numbers denote the order of removal due to largest effect(s) in the RPE analysis
[b] Set A: Based on pairwise allele comparisons, the common alleles are divided into mutually exclusive, and significantly different, predisposing (I), neutral (intermediate) (II), and protective (III) categories for use later in amino acid comparisons (described in detail below and see Table II.H.2)
[c] Set B: The sets I, II, and III above are expanded (indicated by Ix, IIx, and IIIx) to include rare alleles, while excluding those alleles which do not clearly fall into one of the 3 risk categories (see below and Appendix B)
[d] The individual p-values are biased as the assumption of a 1 df chi-square is incorrect, and conservative; the p-values can be used however for a relative ranking of the allelic effects
[e] The upper and lower 95% confidence intervals (CIs) for the Odds Ratio (OR) are given
[f] The binned category consists of all alleles with an expected value < 5 under the chi-square test of heterogeneity of patient and control allele counts

The all pairwise risk comparisons is straightforward: the p-values (uncorrected) from a chi-square test of heterogeneity of all pairwise DRB1 allele comparisons is given in Table III.C.2 for alleles that are *relatively common* in patients or controls in the JIA-OP data. These show a most striking pattern of predisposing (shaded teal), neutral (intermediate) (shaded light grey), and protective (shaded pink) alleles (these are referred to as categories I, II, and III in column 2 of Table II.H.1 above) with risk homogeneity within and risk heterogeneity between the three categories. That is, we have strong evidence that we can confidentially assign these alleles into *nearly* mutually exclusive risk categories.

Table III.C.2: Pairwise risk heterogeneity comparison p-values for common HLA DRB1 alleles and JIA-OP

Reduced data set A - categories I, II, and III

DRB1	*08:01	*11:04	*13:01	*11:01	*01:01	*03:01	*13:02	*04:04	*15:01	*07:01	*04:01
*08:01		0.3454	0.0004	7E-06	4E-07	1E-07	1E-06	4E-10	1E-18	2E-17	8E-16
*11:04	0.3454		0.0376	0.0029	0.0006	0.0004	0.0005	1E-06	1E-11	4E-11	4E-10
*13:01	0.0004	0.0376		0.2186	0.0766	0.0567	0.0496	0.0002	2E-09	1E-08	1E-07
*11:01	7E-06	0.0029	0.2186		0.6705	0.6201	0.371	0.0054	1E-05	2E-05	7E-05
*01:01	4E-07	0.0006	0.0766	0.6705		0.9539	0.5604	0.0096	2E-05	4E-05	0.0001
*03:01	1E-07	0.0004	0.0567	0.6201	0.9539		0.5794	0.0094	1E-05	2E-05	1E-04
*13:02	1E-06	0.0005	0.0496	0.371	0.5604	0.5794		0.051	0.0055	0.006	0.0084
*04:04	4E-10	1E-06	0.0002	0.0054	0.0096	0.0094	0.051		0.8678	0.9155	0.9679
*15:01	1E-18	1E-11	2E-09	1E-05	2E-05	1E-05	0.0055	0.8678		0.9226	0.8521
*07:01	2E-17	4E-11	1E-08	2E-05	4E-05	2E-05	0.006	0.9155	0.9226		0.9246
*04:01	8E-16	4E-10	1E-07	7E-05	0.0001	1E-04	0.0084	0.9679	0.8521	0.9246	

A perfect risk discrimination pattern is broken only by the comparisons of DRB1*13:02 with *13:01 and *04:04. Note that there is an ~ 2 fold difference in the OR for the alleles DRB1*13:01 and *13:02 which flank the range of the alleles listed as neutral (intermediate) in Table III.C.1 so we anticipate that larger sample sizes may show significant risk heterogeneity

within this category. Table B.2 of Appendix B extends the alleles considered, such that rarer alleles within the bounds of the three categories I, II and II above are now included, and this extended set is referred to as categories Ix, IIx, and IIIx, see column 3 of Table III.C.1 above. Table B.3 considers all alleles.

These data show not only the possibility of heterogeneity within a risk category (predisposing, neutral, and protective in this case) but also the fuzziness surrounding these categories. The nearly perfect mutually exclusive block structure of the common alleles in Table III.C.1 above is illusory for many diseases when looking at all alleles. But, there is value in the exercise of building up boundaries in the risk profiles of a subset of alleles. In application of some tests, for example the Unique Combinations Method of Salamon et al. (1996) (see Section V.C), accurate classification is essential to identify amino acids and combinations thereof unique to specific risk categories. Any inaccuracy in risk category assignment may invalidate results. However, while many diseases may show a continuum of risk categories, others may not. Whichever situation applies to a particular disease, detailed consideration of RPE analysis and pairwise risk comparison results are very beneficial.

While our examples have emphasized alleles of DRB1 in JIA-OP, and haplotypes of DRB1-DQB1 in T1D analyses (unpublished data), we emphasize again that genotype analyses are also important, as heterogeneous risk effects may often manifest at this level. Given sample size considerations, these may often be restricted to comparisons of specific genotype combinations.

D. Analysis of subsets of the data and the single parent TDT

In the case, for example, of a single allele showing a strong association with disease, e.g., DRB1*15:01 and multiple sclerosis and DQB1*06:02 and narcolepsy, it may be beneficial to consider subsets of the data to possibly increase the power to detect RPEs of other alleles (and also to detect other gene effects). This may be particularly so if the strongly associated allele shows a more dominant than recessive mode of inheritance (allowing for incomplete penetrance and also so-called sporadic cases of disease, i.e., a base risk of disease regardless of genotype at the primary gene). The rationale for this approach is to concentrate on families (or appropriate subsets of case/control data) where the effects are not overridden by the predominant risk of one or more alleles (or haplotypes or genotypes). With case/control data, one can analyze the RPE's and all pairwise allele risk comparisons as follows: (1) using all patient data; and then when a strong association is found, e.g., DQB1*06:02 and narcolepsy, (2) looking at the distribution of DQB1 non-*06:02 alleles in heterozygous patients, and (3) similarly in patients homozygous for DQB1 non-*06:02 alleles. The control population is the same in each case, the DQB1 non-*06:02 alleles: a stratified subset of the controls is not needed, avoiding small sample sizes in the controls.

Mignot et al. (2007) developed and applied this approach to narcolepsy families where *neither* parent had the DQB1*06:02 allele. In addition, they developed the *single parent* TDT: in simplex (trio) families with a DQB1*06:02 allele transmitted to the proband, the TDT is applied to the other parent using only families where this other parent does not have the DQB1*06:02 allele. Overall, the results indicated that in addition to the well-known strong effect of DQB1*06:02, there were additional RPEs of DQB1*03:01 (susceptibility) and DQB1*05:01 (resistance), replicating previously reported results in case/control studies where similar stratifications were applied (Mignot et al. 2001).

E. Genotype frequencies and tests of modes of inheritance

The AGFAP method

When a marker allele is *strongly* associated with a disease predisposing gene the so-called antigen (allele) genotype frequencies among patients (AGFAP) method has been informative with respect to mode of inheritance - recessive versus additive (dominant expectations are very close to those for an additive model; the additive expectations are less cumbersome to express and are used throughout) (Thomson and Bodmer 1977a, b, Thomson 1983, 1993, 1995a, b). In the case of a linked recessive disease predisposing gene, the genotype frequencies in patients are expected to be in HWPs based on the allele frequencies in the patients. However, for other disease models, lack of fit to HWPs in patients may signal an associated disease gene rather than typing errors.

We consider a bi-allelic marker gene denoted A, with alleles denoted A and a, and respective frequencies p_A and p_a. Allele A is positively associated with disease, and we define a parameter k such that if D denotes the disease predisposing allele at locus D, then the frequency of the haplotype AD is given by kp_D. The theoretical expectations under recessive and additive models are then given in Table II.E.1. Expectations for multiple alleles are easily obtained, as well as the estimates of the parameters k_i for each marker allele A_i allowing tests of mode of inheritance (Thomson 1993, 1995a).

Table III.E.1: Theoretical recessive and additive AGFAP expectations

	AA	**Aa**	**aa**
Recessive	k^2	$2k(1-k)$	$(1-k)^2$
Additive	$k p_A$	$k(1-p_A) + (1-k) p_A$	$(1-k)(1-p_A)$

For a recessive disease model, the expectations are HWPs based on the marker allele frequencies in *patients*, which in this case are k, and 1-k respectively for A and a. Before discovery of the hemochromatosis gene in the extended HLA region, allele A3 of the HLA class I A locus was known to be increased in patients over controls (see Table II.A.1: in Caucasians ~ 72% of patients had at least one copy of the allele A3 compared to ~ 21% of controls – OR = 9.7). Application of the AGFAP method correctly indicated a very close fit to recessive expectations for hemochromatosis (Thomson 1983, see Table III.E.2A below), and rejected an additive model ($p<0.001$). The hallmark of recessive inheritance is that many more patients will be homozygous for the associated allele than under an additive model.

For an additive model, most patients are expected to be heterozygous rather than homozygous for the associated marker allele: the associated A allele frequency in patients is $(k+p_A)/2$. The association of HLA-B27 and ankylosing spondilitis was found early in the study of HLA associated disease (see Table II.A.1: in Caucasians ~ 94% of patients had at least one copy of the allele B27 compared to ~ 9% of controls – OR = 87.8). Application of the AGFAP method (Thomson 1983, see Table III.E.2B below) rejected a recessive model ($p<0.001$), and showed close fit to an additive model.

Table III.E.2: Application of the AGFAP method to hemochromatosis and ankylosing spondilitis data [a]

A. Hemochromatosis

	A3A3	**A3Ax**	**AxAx**	
Observed	20	43	21	Total: 84 individuals
Recessive	20.5	42.0	21.5	ns
Additive	8.8	55.0	20.2	p<0.001

B. Ankylosing spondilitis

	B27B27	**B27Bx**	**BxBx**	
Observed	3	70	6	Total: 79 individuals
Recessive	18.3	39.4	21.3	p<0.001
Additive	3.5	69.1	6.0	ns

[a] More details are given in Thomson (1983, 1993) (also source references for the data). Ax and Bx denote the non-A3 and non-B27 alleles respectively. The control allele frequency for A3 is 0.146, and for B27 is 0.048. The expectations for the observed data are given for recessive and additive models: for the additive model the maximum likelihood estimate (MLE) is used.

As stated in Section II.A, initial modeling of HLA disease associations assumed that the serologically defined HLA alleles were markers for closely linked disease predisposing genes, as is the case with our modeling for the AGFAP method. It turns out that in fact many (most) HLA disease associations are due to a primary gene at one or more of the classical HLA antigen presenting genes. However, the "simple" disease models described above were nonetheless very instructive in the development of our understanding of HLA disease associations, and in many instances continue to be so today when there are strong disease associations. Also, additive AGFAP expectations continue to hold if we additionally allow for sporadic cases when the marker locus is itself directly involved in differential disease risk.

Original application of this model to T1D favored a recessive model, further investigation of multi-allelic class II HLA DR-DQ locus associations led to discovery of heterogeneity beyond this simple model at the genotype and haplotype levels with excess risk of heterozygotes DR3/DR4 over both homozygotes (see e.g., Louis and Thomson 1986 and references therein). These results highlight the fact that fit to a particular model does not provide verification of that model; a more complicated model may apply and it is just that the simpler model is not rejected.

For the T1D data in Table III.B.1 common HLA DR-DQ genotype counts and frequencies are given in Table III.E.3 for the three common predisposing haplotypes and the "neutral" DR1 haplotype. Compared to recessive AGFAP expectations (HWPs *within* the patient group), there are excess heterozygotes for DR3 and the two predisposing DR4 haplotypes in this data set (DRB1*04:01 and *04:04 with DQA1*03:01 DQB1*03:02) and deficiency of the homozygous classes and the heterozygote for these two DR4 haplotypes. Note, as stated above (Section III.E), that for multi-allelic systems, the control frequencies for many genotypes immediately become small unless the sample size is very large.

Table III.E.3: AGFAP recessive analysis of HLA DRB1-DQB1 genotype frequencies in type 1 diabetes

DRB1	DQB1	/	DRB1	DQB1[a]	T1D(%)[b]	AFBACs[c]	Recessive[d]	X^{2e}
03:01	02:01	/	04:01	03:02	44 (24.4)	1.0 (0.7)	29.1 (16.1)	7.6**
03:01	02:01	/	04:04	03:02	21.5 (11.9)	0.5 (0.3)	12.3 (6.8)	6.9**
03:01	02:01	/	03:01	02:01	10 (5.6)	1.2 (0.9)	18.3 (10.2)	3.8*
01:01	05:01	/	04:01	03:02	9 (5.0)	0.7 (0.5)	6.9 (3.8)	0.6
01:01	05:01	/	03:01	02:01	8 (4.4)	1.6 (1.2)	8.7 (4.9)	0.1
04:01	03:02	/	04:01	03:02	5.5 (3.1)	0.2 (0.2)	11.5 (6.4)	3.1
04:01	03:02	/	04:04	03:02	4.5 (2.5)	0.2 (0.1)	9.7 (5.4)	2.8
others					77.5	133.6	83.5 (46.4)	0.4
Total						**180**	**139**	**180**

Data from Noble et al. (1996) and personal communication.

[a] HLA DRB1 DQB1 genotypes, DQA1 can be inferred from Table III.B.1
[b] Patient genotype counts and (%), patient counts are the average of the two affected sibs; only genotypes common in patients are listed.
[c] Controls are estimated from AFBACs assuming HWP (Thomson 1995a, b)
[d] Recessive expectations (Thomson 1983) for cases under the AGFAP analysis; these are HW expectations based on the allele frequencies in the patients (from Table III.B.1), p<<0.0001
[e] X^2 (df=1) of observed versus recessive expectations for individual patient genotypes: * (p<0.05), ** (p<0.01), *** (p<0.001), **** (p<0.0001)

The AGFAP method, and especially its application and similar comparative analyses with T1D HLA DR-DQ data, underscores the information content of genotype risk comparisons. In a large meta-analysis, Thomson et al. (2007a) confirmed additional specific risk effects that are genotype dependent. For example, the well known increased association of DRB1*04:01 DQB1*03:02 versus DRB1*04:01 DQB1*03:01 with T1D is relatively much stronger in combination with DR3 heterozygotes than with DR1 heterozygotes. Analyses of data are thus encouraged at the allele, haplotype, and genotype levels as appropriate, including consideration of specific appropriate subsets of the data.

Conditional logistic regression

While application of the AGFAP method and other comparative analyses of genotype data can be very informative, as illustrated above, in some situations CLR analyses may be more informative. For example, application of the AGFAP method to the PTPN22 RA data of Begovich et al. (2004) cannot distinguish between recessive and additive models (assuming a gene in LD with the marker) (see Thomson et al. 2008). However, use of the likelihood ratio test and CLR analysis significantly excluded a recessive model (p<0.0001) (Begovich et al. 2004); the data were consistent with both additive and dominant modes of inheritance. We *emphasize* this in light of our suggestion in Section I.A (Introduction) that all data sets be analyzed by a variety of methods and results compared. The difference in results in this case most likely arise from the

different models considered, a linked gene in LD with the marker in the AGFAP analysis, and a direct role of the marker in the CLR analysis. With this data, considering the P/C ratios and RPE analysis, we see consistency with the CLR results with the difference in the genotype risks (see Thomson et al. 2008: Table11.3, $p<<0.0001$). In analyses of genotype data and multiple sclerosis, Barcellos et al. (2006) showed that CLR modeling had more power to detect less common genotype effects compared with genoPDT (genotype PDT) analyses.

Differential parental transmission expectations when a parent is affected

For family based data ascertained for the presence of at least one affected parent and at least one or two affected offspring (MPC and MPS pedigrees, see Figures III.C.1b and d respectively), there are diagnostic differential transmission ratios from the affected and unaffected parents based on a recessive (symmetric) versus additive (asymmetric) model. The transmission expectations, as well as that of the non-transmitted allele from the affected parent are given in Table 1 of Thomson (1995a), and also the expected distributions for affected sib pair IBD distributions are given in Table 2.

F. Affected sib pair identity by descent (IBD) values and mode of inheritance

The distribution of the number of HLA haplotypes, or other marker variation, shared by affected sib pairs can also be used to obtain information on the mode of inheritance of the disease: recessive and additive as well as intermediate models (Thomson and Bodmer 1977a, b, Thomson 1980, Louis et al. 1983, Motro and Thomson 1985, Payami et al. 1985, Thomson 1995a, b). (The test is carried out on marker genes close to, or which may include, the disease-predisposing gene.) For recessive and additive models (allowing for incomplete penetrance but not for sporadic cases of disease), the expected IBD values do not involve the disease penetrance values, and are functions only of the disease predisposing allele frequency, which for a two-disease allele model is denoted by p_D.

For the recessive model, the share 2 (X), 1 (Y) and 0 (Z) parental alleles (haplotypes identical by descent (IBD) values in randomly ascertained affected sib pairs (MSP pedigrees) (Figure II.C.1c) are:

$$X = 1 / (1 + p_D)^2, \quad Y = p_D / (1 + p_D)^2, \quad Z = p_D^2 / (1 + p_D)^2.$$

For the additive model, they are:

$$X = (1 + p_D) / [2(1 + 3p_D), \quad Y = 1/2, \quad Z = p_D / (1 + 3p_D).$$

In the limit, for very small disease allele frequencies, i.e., as p_D tends to zero, for the recessive model the expectations tend to X = 1, Y = 0, Z = 0, while the expectations for the additive model are X = ½, Y = ½, Z = 0. Both recessive and additive expectations tend to X = ¼, Y= ½, Z = ¼ as p_D tends to 1, as expected.

Ironically, as above, given we now know the extent of heterogeneity of the HLA DR-DQ contribution to T1D, the results for T1D show incredibly close fit to a recessive model. Data and analyses from 538 families with 711 T1D affected sib pair comparisons show IBD share 2, 1, and 0 values of 373, 283, and 55 compared to recessive expectations of 372.3, 284.4, and 54.3 (Payami et al. 1985). The observed IBD frequencies of 52%, 40%, and 8% are very highly significantly different from random 25%, 50%, and 25% expectations ($p< 10^{-5}$) and reject an additive model ($p<<0.0001$), a hallmark of which is the share 1 class has an expectation of 50%, and of course the share 2 class is greater in frequency than the share 0 class.

The expected IBD values for *affected sib trios* for the recessive and additive models have also been estimated, and applied to T1D HLA DR-DQ data (Payami et al. 1985). Again the data are compatible with recessive expectations and reject an additive model. The sib pair and sib trio data both give high estimates of the disease allele frequency under a recessive model (0.38 and 0.34 respectively); given known monozygotic twin concordance rates, this indicates that the model of one HLA region gene with recessive inheritance is incompatible with the overall data. A model incorporating another (hypothetical) unlinked predisposing gene of risk magnitude around that of HLA was shown to be compatible with observed data (Thomson 1980). We of course now know that HLA contributes the major genetic contribution to T1D, with many other loci contributing much weaker effects to disease risk (Cooper at al. 2008).

The affected sib pair IBD sharing expectations for *MPS pedigrees* (families ascertained for the presence of at least two affected sibs and *one affected parent*) have been obtained for the recessive and additive expectations, as above for a marker gene linked to the actual predisposing disease gene (Thomson 1995a, c). In this case the share 1 (Y) values are subdivided based on whether the affected sibs share the allele transmitted by the affected parent (Y_1), or the unaffected parent (Y_2). For the recessive model:

$$X = 1 / [2 (1 + p_D)],\ Y_1 = p_D / (1 + p_D)^2,\ Y_2 = p_D / (1 + p_D)^2,\ Z = p_D^2 / (1 + p_D)^2.$$

For the additive model:

$$X = [1 + 5p_D + 2p_D^2] / [2 (1 + 9p_D + 6p_D^2)],\quad Y_1 = [1 + 3p_D + 4p_D^2] / [2 (1 + 9p_D + 6p_D^2)],$$

$$Y_2 = [2p_D (3 + p_D)] / [2 (1 + 9p_D + 6p_D^2)],\quad Z = [4p_D (1 + p_D)] / [2 (1 + 9p_D + 6p_D^2)].$$

As for MSP pedigrees, the contrast between these expectations is large; in the limit as p_D tends to zero, the recessive model expectations tend to $X = 1/2$, $Y_1 = 0$, $Y_2 = 1/2$, $Z = 0$, while the additive model expectations tend to $X = 1/2$, $Y_1 = 1/2$, $Y_2 = 0$, $Z = 0$. Both recessive and additive expectations converge on random expectations of $X = ¼$, $Y_1 = ¼$, $Y_2 = ¼$, $Z = ¼$ as p_D tends to 1, as expected. The asymmetry in both recessive and additive expectations for the Y_1 and Y_2 expectations provide a powerful additional contrast in analyses of the data.

G. The interrelationship of HLA associated diseases

There is strong evidence for the interrelationship of disease risk with a number of HLA associated diseases: the same alleles or haplotypes are found strongly associated with differential risk for a number of diseases, e.g., DRB1*15:01 DQB1*06:02 is predisposing for multiple sclerosis and narcolepsy while it is very protective for T1D; DRB1*03:01 DQB1*02:01 is predisposing for T1D and celiac disease; DRB1*04:01 DQB1*03:02 is predisposing for T1D and rheumatoid arthritis. It is of particular interest to study the genetic interrelationship of these and other HLA associated diseases.

Payami et al. (1987) developed the appropriate affected sib pair IBD distributions for various combinations of two diseases, e.g., one sib has both diseases designated A and B, the other sib only has disease A, etc., for various combinations of the same genes contributing to disease risk, versus different genes contributing to each disease risk. Data with three diseases, T1D, rheumatoid arthritis, and autoimmune thyroid disease were analyzed and showed evidence for commonality of effects for T1D subsets (DR3 and DR4) separately with rheumatoid arthritis and autoimmune thyroid disease, and probably separate alleles contributing to rheumatoid arthritis and autoimmune thyroid disease. These results are logical given the known HLA disease associations. With high level resolution typing, and the fact that known high risk (predisposing and protective) HLA DR-DQ alleles and haplotypes are found in a number of diseases, it is time to reinvestigate this issue.

IV. Secondary Disease Genes

A. Introduction

With extensive SNP typing in a genetic region, the HLA region for example, before stratification analyses are performed, the number of markers with significant associations with disease may be quite large. While the strongest effects are expected to be seen around the *primary* disease gene, nevertheless very strong associations are seen at the many SNPs in high LD with the *primary* disease gene(s) for many HLA associated diseases (see for example Valdes et al. 2009, 2010). With application of stratification analyses, when the true *primary* disease gene has been identified, the number of markers showing significant associations is expected to decrease quite a bit, and further, the strength of these associations may be much weaker. This is demonstrated with class III markers and T1D in Valdes et al. (2009).

But as mentioned above, all marker loci should continue to be studied, as markers which do not show a significant association before stratification analyses, may nevertheless show significant associations with stratification analyses (see Thomson et al. 2008). Further, constant reassessment of *primary* disease genes must be made as further markers are studied and *additional* disease genes are identified. It may be that some SNPs are also identified as *primary* effects, either additional to the classical HLA genes, and possibly with stronger effects, or interaction effects, etc.

As *additional* disease genes are identified, they are also subject to the same analyses as *primary* disease genes, again all heterogeneity in disease risk must be identified, including allele, haplotype, or genotype specific effects, and interaction effects (see Section III). Future stratification analyses *must* also take account of all known *primary* and *additional* disease gene effects. As mentioned above, cell sizes may become small. We may need to combine classes of homogenous effects, but again, these must be continually reassessed as *additional* disease genes are identified. Tests for modes of inheritance, maternal-fetal effects, and imprinting should also be carried out on all *primary* disease genes including *additional* disease gene effects as they are identified (see Thomson et al. 2008 for review).

The methods described below to detect *additional* disease genes are used to formally define *primary* disease genes. In effect, all stratification analyses should be applied in both directions. As illustration, for T1D, we would first condition on HLA DR-DQ haplotypes and genotypes and consider marker genes, e.g., an individual SNP, MSAT, or specific gene, or haplotype or genotype combinations thereof. For all significant effects, the data should also be analyzed *vice versa*, conditioning on this marker and studying the HLA DR-DQ variation. We repeat here, that we define *primary* disease genes as those that "stand out" in initial association studies, and *for the most part* associations of other markers in the genetic region can be explained via their LD patterns with the primary disease gene(s). As our analyses progress, a number of genes or SNPs (not necessarily in the same gene) in a region may be included in the *primary* disease gene category.

Reports of other HLA region gene and MSAT associations with diseases have appeared in the literature. In many of these studies it has been difficult to determine if an additional HLA region gene is involved in disease, versus the associations reflecting LD with the antigen presenting HLA molecules directly involved in disease. However, a number of analytic strategies have been developed to remove the effects of LD with the antigen presenting HLA genes directly involved in the disease and these are described below.

Controlling for the influence of class II DR-DQ haplotype and genotype effects, a role in T1D has been shown of additional HLA class II (DPB1) (Noble et al. 2000, Valdes et al. 2001) and class I genes (including age of onset effects and rapid disease progression) (Fujisawa et al. 1995, Nakanishi et al. 1995 , Noble et al. 2002, Steenkiste et al. 2007, Tait et al. 1995, Valdes et al. 1999, 2005a).

Also, various analyses have shown the presence of additional disease predisposing MSATs on specific high risk DR-DQ haplotypes and genotypes, that is, DR3, and DRB1*0401 and DRB1*0404 with DQB1*0302 (see e.g., Hanifi Moghaddam et al. 1998, Johansson et al. 2003, Lie et al. 1999a, 1999b, Pugliese et al. 2007, Steenkiste et al. 2007, Valdes et al. 2005c, Zavattari et al. 2001). Possible heterogeneity of the DR15 haplotype and T1D has been shown, with significant reduction in the diabetes-protective effect typically associated with this haplotype (Valdes et al. 2005b). The results from study of the effects of additional disease genes using different populations and even different samples from similar populations are very heterogenous and show weak effects, reminiscent of non-HLA gene effects. This is clearly demonstrated in analyses of the 13th International Histocompatibility Workshop T1D HLA data, a large worldwide collection typed for the classical HLA genes and eight msats in the HLA region (Pugliese et al. 2007 , Steenkiste et al. 2007).

Extensive SNP typing in the HLA region in combination with high resolution HLA typing in the T1DGC (Type 1 Diabetes Genetic Consortium) samples, has allowed more detailed analyses of the secondary roles of other classical HLA genes, as well as discovery of several

other secondary loci in the HLA region involved in T1D risk after conditioning for DRB1-DQB1 (see e.g., Valdes et al. 2009, 2010).

B. Matched cases and controls

A modification of the classic case/control design is to match cases and controls at the genotype level for the HLA class I or class II genes known to increase risk of disease, or other primary disease predisposing gene(s). This eliminates the effects of LD between the primary disease gene and other marker genes under study (Hanifi Moghaddam et al. 1998). Significant case/control differences in the distribution of marker allele, haplotype and genotype frequencies reflect the effect of additional disease susceptibility genes in the region. The major disadvantage of this matched approach is that it reduces the number of cases and controls available for analysis. An advantage is that effects can be summed over populations: the ratio of matched cases/controls must be the same in all data sets if they are to be combined. This approach is powerful if an effect is specific to or more easily detected in a specific subset of the data, either high risk or low risk. Note that the matched cases and controls approach is a specific case of the conditional genotype method (CGM) discussed below. This approach and the CGM can also be applied to family based data using, for example, matched genotypes from the proband (or average of affected sib pairs) and AFBACs (see Section II.C).

With matched cases and controls, the usual case, especially with T1D and its hierarchy of HLA DR-DQ disease effects, will be to restrict analyses to the genotypes common in patients. Hanifi-Moghaddam et al. (1998) analyzed data on T1D patients and controls matched for DR3/DRB1*0401 DQB1*0302, the most common high-risk genotype in Caucasians. Two HLA regions showed significant MSAT associations. Some analyses of the 13th IHW disease data sets used matched cases and controls (Thomson et al. 2007b, Pugliese et al. 2007, Thorsby et al. 2007, Steenkiste et al. 2007) and found evidence of additional disease predisposing loci for T1D and other diseases. Note however that the results are heterogenous between studies and replication of effects was difficult.

With diseases with a dominant predisposing effect, for example, narcolepsy and DQB1*0602, the decision must be made whether to match specifically on the non-DQB1*0602 alleles in heterozygous individuals. As we have discussed above, it is *essential* to have performed an RPE analysis of allele, haplotype and genotype effects at the primary disease locus before proceeding in this way. With narcolepsy, additional DQB1 effects are seen as described above and must be taken into account, similarly with multiple sclerosis and DRB1. The difficulty with dominant protection, for example, T1D and DR15/DRX (X=non-DR15), is that the sample size in patients will be small, but nevertheless in samples where there are sufficient numbers in this category they should be investigated. Valdes et al. (2005b) have found preliminary evidence of a marker which modifies the protective effect of DR15 in a Swedish population.

Additional genetic effects may be specific to a high risk category. On the other hand, they may be restricted to, or be more easily detected in, a subgroup of cases and controls or families lacking the high risk factors at the primary disease gene(s). Removing multiple sclerosis trio families with DRB1*1501, and also DRB1*03 and DRB1*0103, from consideration, Yeo et al. (2007) found a protective effect for the allele C*05 of the HLA class I C locus. The effect could not be distinguished in the high risk DRB1*1501 set of families.

C. Homozygous parent linkage and TDT tests

Homozygous parent linkage test (HPLT)

Using affected sib pair data, the role of genes additional to HLA DR-DQ in T1D was demonstrated by Robinson et al. (1993) using the homozygous parent linkage test (HPLT). Affected sib pairs (MSP pedigrees) with a *parent homozygous* for the DR3 haplotype were examined; a marker gene was used to distinguish between the DR3 haplotypes; in this case it was the highly polymorphic B locus, but it could be any combination of marker genes. Under the null hypothesis that no HLA region variation additional to that defined by DR3 is involved in T1D, the affected sib pairs should share the two parental DR3 haplotypes equally frequently. Significant deviation from 50% sharing was observed. Since DR3 haplotypes can for the most part be assumed to be homogenous for their DR-DQ alleles (DRB1*03:01 DQA1*05:01 DQB1*02:01), this test implicated other HLA region genes in T1D on DR3 haplotypes. This result is consistent with more recent studies using matched case/control data (above), the homozygous parent TDT (HPTDT), and the conditional haplotype method (CHM) (see below).

How many parents can be expected to be homozygous for an allele (haplotype) at the primary disease predisposing gene? This is relevant not only to the current method, but also the HPTDT described below. For simplex (S) trio families, and approximately for multiplex affected sib pair (MSP) families, the expected number of homozygous parents is given by 2 (APi) (ACi), where APi is the observed frequency of the allele or haplotype (say DR3) in patients, and ACi is that in the AFBACs (determined from the results in Thomson 1995a, b). In the T1D affected sib pair families in Noble et al. (1996) shown in Table III.B.1, the DR3 haplotype had a frequency of 0.319 in patients and 0.094 in controls; the observed number of homozygous DR3 parents was 21 (0.058) (Table III.E.3), which agrees well with the estimated expected of 0.06. Obviously the more polymorphic the marker locus the greater the power of this method (more informative DR3 parents); haplotypes of markers can be used to increase the power. If the sample size is sufficient, specific marker alleles and haplotypes can be studied to narrow down significant effects. The robustness of this approach is tempered by the fact that only a subset of all available data is used. Note that a significant effect with the HPLT does not mean a significant effect will be seen with the HPTDT (see below), and vice versa. As with the matched cases and controls approach, similarly with the HPTDT below, the robustness of this approach to across population analyses may be countered by the fact that only a limited subset of the data is used.

Homozygous parent TDT (HPTDT)

A modification of the TDT utilizing trio family data— S pedigrees (Figure II.C.1a) (affected sib pairs can also be used from MSP pedigrees)—with as above a parent homozygous for DR3 (the HPTDT), has also shown heterogeneity of DR3 haplotypes for T1D risk (Johansson et al. 2003 , Lie et al. 1999a, b). Families were examined to determine which haplotypes, defined by DR3 and an MSAT marker in this case, were transmitted (T) and not transmitted (NT) from the homozygous DR3 parent to the affected child. As above, families are informative only if they are heterozygous at the marker locus.

Issues of sample size are dramatically illustrated in application of the HPTDT to the 13th IHW data on T1D (Pugliese et al. 2007). Of the 307 families with a homozygous DR-DQ parent, 179, 83, and 14, respectively, were homozygous for DR3, DRB1*0401 DQB1*0302, and DRB1*0404 DQB1*0302 (these are the three predisposing haplotypes shown in Table III.B.1and

common in Caucasian T1D patients). These numbers are exceedingly impressive and highlight the importance of collaboration and sharing. Nevertheless, to apply the HPTDT, the MSAT marker (of eight studied) had to be heterozygous in the homozygous DR parent. Further, because there are multiple MSAT alleles at each locus, the actual numbers in any one MSAT allele category is often small, even with the impressive sample size of the 13th IHW. (If one used biallelic SNPs, then many of the parents homozygous for the primary disease gene would not be heterozygous for the marker, hence haplotypes of multiple SNPs should be used.)

Parents homozygous for the three predisposing haplotypes listed above were analyzed both in individual populations and combined across populations (Pugliese et al. 2007, Steenkiste et al. 2007). The 13th IHW HPTDT results showed some significant effects but these were very heterogenous; different markers were significant in different populations, and the marker implicated in Lie et al. (1999a, b) was not replicated. However, as mentioned above, there is combined evidence from a number of studies for heterogeneity of DR3 haplotypes and T1D.

D. Conditional haplotype, genotype and logistic regression methods

The conditional haplotype method (CHM)

The logic of the CHM is as follows: if all HLA region genes directly involved in disease susceptibility have been identified, for example, HLA DR-DQ in T1D as the null, then the *relative* frequencies of alleles at polymorphic marker loci on high-risk haplotypes containing, for example, DR3 should be the same in cases and controls; similarly for other high risk haplotypes, as well as neutral, and protective haplotypes (in the latter case sample size in patients may be an issue). Denote the primary disease locus by A (or the primary plus secondary loci that have been identified), that is, all putative disease predisposing loci in the region, and alleles or haplotypes thereof by A_i, i = 1,2,.., k_A, and a linked marker locus by B, with alleles B_k, k = 1,2,..,k_B, then, under the null that the *A* locus defines all disease predisposition in the region:

$$f_{pat}(A_i\text{-}B_k) / f_{pat}(A_i\text{-}B_l) = f_{con}(A_i\text{-}B_k) / f_{con}(A_i\text{-}B_l),$$

where $f_{pat}(.)$ and $f_{con}(.)$ represent patient and control frequencies. That is, although the frequencies of the haplotypes A_i-B_k and A_i-B_l will differ between patients and controls, the relative frequency of their ratios is expected to be the same in patients and controls, for each A_i.

Inequality of these relative frequencies in patients and controls is expected if the allele or SNP or haplotypes thereof under study *does not include all* genes involved in the disease process and in LD with the marker loci. While fit to these expectations does not exclude the possibility that other genes in the HLA complex are involved in disease, lack of fit unequivocally shows that all disease-predisposing genes in the region have *not* been identified (provided that stratification effects have not produced spurious results). With a CETDT analysis one similarly tests for heterogeneity of a specific haplotype using the TDT statistic (Koeleman et al. 2000a, b).

The CHM method initially showed heterogeneity of serologically defined DR3 haplotypes (Thomson et al. 1988) that has also been verified with a number of other methods of analysis as described above. The method was later applied, with development of an appropriate statistical test, to amino acid sites in the HLA DR-DQ genes to detect combinations of amino acids

involved in disease risk (Valdes and Thomson 1997, Valdes et al. 1997). As noted above, fit to the model does not imply that all amino acids have been identified, but lack of fit indicates that all genetic variation has not been accounted for.

The overall conditional haplotype method (OCHM)

Thomson (1984) developed a method to test for additional genetic effects over all haplotypes, henceforth referred to as the overall conditional haplotype method (OCHM). In that application of the test, haplotypes needed to be estimated. This test can be applied without resort to haplotype estimation; however the test statistic with both these approaches is not straightforward. One can also take the observed (obs) and expected (exp) $f_{pat}(B_kB_l)$ values from above and consider the observed and expected allele frequencies at the B locus under the null, obtained simply by the method of allele (gene) counting. This is in effect the same as the overall conditional haplotype method (OCHM) of Thomson (1984) but obviates the need to estimate LD values in controls between the A and B genes. Note also, however, that it uses ratios of A locus genotype frequencies in patients versus controls rather than ratios of allele frequencies as in Thomson (1984), which may lead to a larger variance for the estimated values. Our results re statistical testing for OCGM data also apply; further work is required to determine the appropriate test statistic, again resampling is a solution.

One advantage of the CHM, OCHM, and CETDT, and the CGMs described below, is that more of the data is used than with the matched genotype approach and the HPTDT and HPLT methods. Results across studies cannot be directly combined, although combining of, for example, *p* values can be carried out (Fisher 1970). Care must be taken in all analyses and their interpretations with rare haplotypes and sparse cells.

The conditional genotype method (CGM)

Similar to the CHM, one can consider genotype frequencies (the CGM): if all disease-predisposing genes in the region have been identified and represented by the locus A, then under this null hypothesis, the genotype frequencies at a linked marker locus B (not involved in disease nor in LD with additional genes involved in disease) are expected to satisfy the relationship:

$$f_{pat}(A_iA_j\ B_kB_l) / f_{pat}(A_iA_j\ B_mB_n) = f_{con}(A_iA_j\ B_kB_l) / f_{con}(A_iA_j\ B_mB_n).$$

The genotypes at the B locus include all homozygotes and heterozygotes, and similarly all homozygotes and heterozygotes can be considered at the A locus (except for those that are too rare), although each is analyzed individually with the CGM (note again the equivalence with the matched case/control method (Section III.D)).

The overall conditional genotype method (OCGM)

If the effect of an additional disease predisposing gene in the region, for example, the B gene or one in LD with it, is not specific to a particular haplotype or genotype at the primary disease locus A, then more power should be available by considering B locus genotypes combined over all A locus genotypes - the overall conditional genotype method (OCGM). In this case, we use

the following expectation for each genotype combination in patients (Thomson and Valdes 2007):

$$\exp f_{pat}(A_iA_j\ B_kB_l) = [f_{con}(A_iA_j\ B_kB_l)]\ [f_{pat}(A_iA_j) / f_{con}(A_iA_j)].$$

For each B locus genotype, B_kB_l, we add over the A locus effects above, and expected patient values are compared to the observed. The question of statistical testing then arises. Application of a standard test of homogeneity of the *B* genotype observed (obs) and expected (exp) genotype numbers does not give a chi-square distribution, in fact the distribution is exponential. This is because of use of the ratio of the A_iA_j genotype frequencies in the estimation of expected values. Note also that the use of low frequency control genotype frequencies will be problematic notwithstanding; these genotypes can be left out of all studies. An appropriate test statistic has been developed (Thomson and Valdes 2007); a resampling approach is also an option.

Conditional logistic regression (CLR)

With CLR modeling, as described above, various aspects of the data can be analyzed: modes of inheritance of specific markers or proposed primary disease genes, dose and other heterogeneous effects of associated alleles, maximum likelihood estimates of relative penetrance values normalized to a reference genotype, and the effects of additional marker genes. Simmonds et al. (2005) using a stepwise logistic-regression analysis showed that the association of HLA DR-DQ with Graves disease could be explained with either DRB1 or DQA1 but not by DQB1. These data could also be analyzed using conditional haplotype and genotype methods (CHM, OCHM, CGM, and OCGM).

When a logistic regression, including CLR, is used to model the relationship between genetic factors and disease, as the distribution of data across numerous combinations of loci becomes sparse, the parameter estimates become unreasonably biased (for review see Thornton-Wells et al. 2004). In other words, the analysis then suffers from the curse of dimensionality. In analyses considering a combination of loci, one or more of which have low minor allele frequencies, the number of individuals with certain multilocus genotype combinations will be so small (or perhaps equal to zero) that a reasonable estimate for that combination of genotypes cannot be derived. For HLA data on the classical loci this is particularly critical as most multilocus haplotypes are <5%.

In addition, it is quite possible that the effect of a secondary locus is due to its interaction with a major susceptibility locus and not to a "main" effect. For example, a SNP allele or genotype may only have a role on certain predisposing HLA haplotypes but not on all others. In those situations some of the traditional implementations of logistic regression models (that is, forward stepwise regression) which require significant main effects to be modeled before including interaction effects between factors represent a major methodological limitation. In practice loci with relatively small main (non-interactive) effects but more substantial interactive effects would never be even included in the analyses.

This is not to deny the power and advantages of CLR analyses, for example to include covariates based on age of onset, sex, and population heterogeneity via principal components effects. The potential pitfalls are emphasized as there is more reliance on computer software

outputs with CLR than with previous analyses of HLA data using contingency table analyses, and careful inspection of the data by the analysts.

E. Combined association and IBD data

If a primary gene has been identified, and no additional secondary genes are involved in disease in the genetic region under study, then *all* features of the data must be explained by the primary gene. If not, then additional secondary genes remain to be identified.

The marker associated segregation chi-square (MASC) method of Clerget-Darpoux et al. (1988, 1991) fits the most parsimonious model explaining the overall linkage and association observations and tests for fit to the data to test the hypothesis that a primary disease gene has been identified. The MASC method was extended (Dizier et al. 1994) to take account of the role of two unlinked candidate genes in T1D, in this case HLA and IDDM2. With molecular HLA data, fitting DR-DQ as the sole HLA susceptibility locus to T1D was strongly rejected (Valdes et al. 2001); addition of HLA DPB1 gave a better fit to the data. T1D probands were stratified into two groups: those not carrying the alleles DPB1*0301 and *0202 (which are associated with disease after accounting for the DR-DQ primary association) and those with at least one copy of either of these alleles. Interestingly, both groups have almost identical frequencies of DR-DQ haplotypes but significantly different IBD distributions in the subset of families with probands who do not carry the highly predisposing DR3/DR4 genotype. We stress again the necessity to understand all aspects of disease heterogeneity at the primary and additional disease loci in a region.

Rheumatoid arthritis is associated with HLA DRB1 alleles, and in particular the group of alleles associated with disease has in common closely related amino acids in the third hypervariable region of the DR molecule at positions 70-74: the “shared epitope” (SE) hypothesis (Gregerson et al. 1987). Application of the MASC method to rheumatoid arthritis HLA “shared epitope” data shows lack of fit of the linkage and association data (Genin et al. 1998). However, if the heterogeneity in risk of the SE alleles due to variation at amino acid positions 70 and 71 is taken into account, then both the linkage and association data fit the model (du Montcel et al. 2005). This fit does not exclude the possibility that additional HLA region variation may modulate disease risk; a role of additional variation on HLA DR3 haplotypes has been shown with rheumatoid arthritis (Jawaheer et al. 2002, Nelson et al. 2007).

For PTPN22 data and rheumatoid arthritis, Bourgey et al. (2007) applied the MASC method and showed that the R620W variant (Begovich et al. 2004) alone could not explain the observed association and linkage data; the data were compatible with 3 SNPs studied or possibly via the role of two untyped SNPs. This does not exclude the role of additional variants in this region. Carlton et al. (2005) have demonstrated SNP associations additional to R620W.

With application of all methods described in this Section, including the MASC method, it is advantageous in terms of cell sizes to pool subsets of data when appropriate. However, we must always account for all known heterogeneity, and additionally be on the lookout for different aspects of the data which highlight heterogeneity, for example, the IBD distributions described above with DPB1 and T1D (Valdes et al. 2001).

V. Detecting Amino Acids at Classical HLA Genes Involved in Disease Risk

A. Introduction

Peptide motifs important for binding to HLA molecules, including critical residues, have been defined by sequence analysis of naturally processed peptides eluted from HLA molecules, analysis with synthetic peptides, phage display libraries, and predictive inference of binding preference based on similarity of peptide-binding environments, see e.g., Leisner et al. (2008), Nielsen et al. (2004), and Frahm et al. (2008). Peptide motifs important for binding to HLA molecules, including critical residues, e.g., position 57 of DQB1, have been defined by sequence analysis of naturally processed peptides, analysis with synthetic peptides, and predictive inference of binding preference based on similarity of peptide-binding environments.

The variation in ability of different HLA alleles to present specific peptides is believed to be the basis of their associations with infectious and autoimmune diseases. The peptide epitopes of Epstein-Barr virus, human immunodeficiency virus, and other infectious agents have been elucidated in model systems, as have specific MHC alleles involved in their binding/presenting (reviewed in Karp et al. 2010). For most autoimmune diseases, while the HLA allelic and genotypic associations are well identified, the hypothesized antigenic peptides contributing to these associations are not known. However, if the amino acids directly involved in disease risk can be identified, predictions could be made about peptide epitopes, and this could lead to the design of novel vaccines and a better understanding of autoimmunity (Karp et al. 2010).

Using a variety of methods summarized below, specific amino acids and combinations thereof have been identified that are potentially directly involved in differential disease risk in a number of HLA associated diseases.

B. Within serogroup and sequence alignment comparisons

Initial analyses usually rely on differential risk effects within serogroups of alleles, since alleles within the same serotype are more closely related at the AA level, hence significant differences in risk within serotypes, and between specific pairs of alleles within a serotype, may more easily identify specific amino acids, or a few amino acids, involved in disease. For example, study of T1D and DRB1*04:XX alleles on DQB1*03:02 haplotypes established heterogeneity in risk and a direct role of DRB1 in disease risk; also the initial observation of DQB1*03:02 versus *03:01 differential risk effects on DRB1*04:01 haplotypes implicating a direct role of DQB1 and amino acid position 57 in disease risk; and similarly other informative comparisons of haplotypes with risk heterogeneity (see Thomson et al. 2007a for references and more details). Within serogroup analyses involve a more restricted set of amino acids compared to overall allele level comparisons.

With the JIA-OP data of Table III.C.1, pairwise comparisons within serotypes of alleles with sufficient sample size—DRB1*01:XX, *04:XX, *11:XX, and *13:XX—were performed using a chi-square heterogeneity test of the respective patient and control counts; this was followed by manual inspection of their respective sequences for comparisons where significant risk heterogeneity was detected (Thomson et al. 2010). Significant results (ordered by p-value)

identified specific amino acids or a set of amino acids required to explain the differential disease risk (Table V.B.1).

Table V.B.1: Within serogroup differential risk comparisons of DRB1 alleles and JIA-OP

Alleles compared[a]	p-value[b]	Amino Acids[c,d]
*04:03 vs *04:01 + *04:04	0.002	**74**
*11:04 vs *11:03	0.003	**86**
*04:03 vs *04:01	0.004	71, **74**, or **86**
*01:01 vs *01:03	0.02	**67**, 70, or 71
*11:03 vs *11:01	0.04	71 or **86**
*13:01 vs *13:02	0.05	**86**

[a] Within serogroup allele comparisons of data in Table III.C.1
[b] Uncorrected p-value from the chi-square test of heterogeneity
[c] Amino acid residues that uniquely define these specific allele differences
[d] Amino acids indicated in **bold** are those identified by other analyses as potentially playing a major or important role in disease risk for JIA-OP, those underlined as potentially having an effect, albeit weaker

The strong evidence for the role of amino acid **86** in differential disease risk is of particular interest, since with SFVT analysis (see below) this amino acid shows no significant effect. There is also sufficient evidence for a direct role of amino acid **74**, which, in contrast to **AA 86**, individually shows a very significant effect with SFVT results (see Section V.D).

Identification of patterns from the sequence alignments of polymorphic sites at all alleles, again stratified by risk categories, has also been successfully applied, e.g., DRB1 and the so-called "shared epitope" set of amino acids 70-74 of DRB1 and rheumatoid arthritis. Recently autoimmunity to citrullinated protein antigens has been shown to define a clinically and genetically distinct subset of rheumatoid arthritis that is specifically associated with the "shared epitope" alleles (reviewed in Imboden 2005). However, note that there is disease risk heterogeneity within the "shared epitope" set of alleles.

C. The Unique Combinations Method

In the original Unique Combinations algorithm (Salamon et al. 1996), two categories of amino acid sequences are defined by the user: those in the "check" category are compared against those in the "group" category in order to identify combinations of sites that are unique between these two groups of sequences. However, when there are two or more sequences in the check category, sites that are polymorphic between these check sequences are excluded from consideration. In Thomson et al. (2010) we extended the algorithm to allow inclusion of all sites that are

polymorphic in the check category, thus expanding the utility of the method. Also, this means that the group and check categories are now interchangeable, whereas before there was an asymmetry.

This extension of the Unique Combinations algorithm provides an ordered list of a minimal number of polymorphic positions, which as a haplotype can differentiate between the alleles in the "check" and "group" categories. Deriving the sets of amino acids that correspond to the resulting minimal unique combination generates unique sequences that either belong to the check category or the group category. Thus, if we compare HLA alleles divided into two mutually exclusive sets based on risk heterogeneity between the sets of alleles, amino acid variation identified as uniquely distinguishing the two sets are likely candidates as directly involved in disease risk.

For the JIA-OP data of Table III.C.1, using the subdivisions of common DRB1 alleles (and their extensions) into the three categories defined as sets A (column 2) and B (column 3)—I and Ix (predisposing), II and IIx (intermediate), and III and IIIx (protective) —we performed various Unique Combinations comparisons (Thomson et al. 2010). In Table V.C.1 below, we report the results of the comparisons of each risk category versus the other 2 risk categories, e.g., I versus II + III. The single, pair, and most commonly seen triple combinations of amino acids differentiating the sequences in the group and check categories are listed. Note that if an amino acid appears in a single combination, it will not appear in any pair or triple combinations, etc.

The DRB1 amino acids targeted from the Unique Combinations algorithm as important in JIA-OP risk are **86** (as in the pairwise allele within serotype analyses above) combined with **13** and **37**, or **13** and **67** (Thomson et al. 2010).

Table V.C.1: JIA-OP HLA DRB1 Unique Combinations analyses

	A1	A2	B1	B2	C1	C2	
group[a]	**I**	**Ix**	**III**	**IIIx**	**II**	**IIx**	
check[a]	**II+III**	**IIx IIIx**	**I+II**	**Ix+IIx**	**I+III**	**Ix+IIIx**	
							Amino Acids[b]
			X				**13**
				X			**13, 67**
			X	X			**37, 67**
	X	X			X		**13, 37, 86**
	X	X			X	X	**13, 67, 86**

[a] The sets of predisposing (I and Ix), intermediate (II and IIx), and protective (III and IIIx) alleles are defined in Table III.C.1, the group and check categories define the 2 groups of alleles being compared (the results are symmetric for the group and check categories)

[b] Only amino acids in exon 2 are listed, other amino acids which appear in some comparisons are: 47, 57, 70, 71, 74

D. Sequence Feature Variant Type (SFVT) analysis

Sequence Feature Variant Type (SFVT) analysis was developed to systematically perform association tests focusing on variation at biologically relevant SFs, which are based on structural (e.g., beta-strand 1) and functional (e.g., peptide binding site) features of the protein (Karp et al. 2010). The SFs include classical HLA allele level and single amino acid polymorphisms. With systemic sclerosis, specific amino acids in pockets 4 and 7 of the peptide binding site are shown to explain much of the molecular determinant of disease risk (Karp et al. 2010).

Thomson et al. (2010) applied SFVT analysis (Tables V.D.1A and 1B) to the JIA-OP data of Table III.C.1; specific amino acids in pockets of the peptide binding site are shown to account for the major disease risk. After initial SFVT analysis (Table V.D.1A), so-called temporary SFs (tSFs) (Table V.D.1B) are defined by combinations of potentially informative amino acids.

Table V.D.1: SFVT results

A. JIA-OP HLA DRB1 SFVT analysis with SFs ranked by overall p-values

Rank	SF #	Description[a]	amino acid[b]	p-value	max OR	min OR
1	57	**position 13**	**13**	1.8E-28	4.91	0.33
2	136	**pocket 6**	*11*, **13**, 30	3.9E-28	7.07	0.31
3	134	**pocket 4**	**13**, 26, 28, 70, 71, **74**, 78	5.7E-28	6.84	0.28
4	151	*beta1_pep ant-TCR*	*11*, **13**	9.4E-28	4.89	0.33
5	1	**allele**[c]		1.1E-27	9.40	0.28
6	137	**pocket 7**	28, 30, 47, 61, **67**, 71	8.9E-27	9.40	0.28
7	142	pep ant position 12	*9*, 56, **57**, 60, 61, **67**	4.2E-26	7.14	0.31
8	55	*position 11*	*11*	8.8E-25	3.15	0.33
9	130	pep ant & TCR	60, **67**, 70, 71, 77, 78, 85	7.0E-24	9.40	0.32
10	56	*position 12*	*12*	1.1E-22	3.15	0.32
11	54	*position 10*	*10*	1.9E-22	3.14	0.32
12	162	alpha2_pocket 7	**67**, 71	2.9E-21	9.40	0.33
13	19	alpha-helix 1	52..62	3.5E-18	3.92	0.44
14	98	**position 67**	**67**	3.0E-17	3.39	0.54
15	138	**pocket 9**	*9*, **37, 57**	3.6E-16	3.92	0.33
16	104	**position 74**	**74**	3.8E-16	6.84	0.33
17	74	position 37	**37**	3.7E-13	1.80	0.34
18	59	*position 16*	*16*	5.4E-13	4.91	0.20
19	90	**position 57**	**57**	5.5E-13	3.92	0.44
20	53	*position 9*	*9*	2.3E-11	2.30	0.42
21	135	pocket 5	70, 71	1.4E-10	1.79	0.33
22	102	position 71	71	1.2E-09	1.48	0.33
23	101	position 70	70	4.5E-09	2.03	0.54
24	71	position 33	33	3.1E-07	2.62	0.38
25	58	*position 14*	*14*	3.6E-07	3.05	0.33
26	63	*position 25*	*25*	3.6E-07	3.05	0.33
27	91	position 58	58	1.0E-06	2.35	0.43

28	24	position 78	78	3.3E-06	2.56	0.39
29	93	position 60	60	3.5E-06	2.36	0.44
30	68	position 30	30	2.3E-05	1.55	0.33
31	152	beta2_pocket 7	28, 30	4.3E-05	1.55	0.33
32	155	beta2_pocket 4	26, 28	4.4E-05	1.31	0.34
33	141	pep ant position 4	77, 78, 81, 82, 85	7.4E-05	1.37	0.39
34	153	beta2_pep ant bind	26, 28, 30	0.0001	1.32	0.33
35	13	beta-strand 2	23..32	0.0002	1.29	0.33
36	66	position 28	28	0.004	1.54	0.63
37	22	position 73	73	0.007	1.47	0.68
38	81	position 47	47	0.03	1.28	0.78
39	178	beta1_CD4 bind	41..56	0.03	1.28	0.78
40	70	position 32	32	0.07	1.25	0.80
41	154	beta2_alpha chain	29, 31, 32	0.13	1.27	0.80
42	110	**position 86**	**86**	0.34	1.12	0.90
43	106	position 77	77	0.41	1.16	0.86
44	132	**pocket 1**	82, 85, **86**, (89, 90)	0.58	1.14	0.89
45	173	alpha4_pocket1	82, 85, **86**	0.63	1.13	0.90
46	64	position 26	26	0.67	1.16	0.93
47	109	position 85	85	0.74	1.13	0.88
48	69	position 31	31	0.86	1.03	0.97
49	75	position 38	38	0.91	1.04	0.96

[a] These are abbreviated descriptions, see Karp et al (2010) for the full definition of each SF

[b] Amino acids indicated in **bold** are those identified as having a major or important role in disease risk, those underlined as potentially having an effect, albeit weaker, and for those in *italics* their effect may be explained by LD with **AA 13**.

[b] SF127 (peptide antigen binding site - PBS) amino acids: 9, 11, 13, 26, 28, 30, 37, 47, 56, 57, 60, 61, 67, 70, 71, 74, 77, 78, 81, 82, 85, 86, 89, 90, has identical VT counts as SF1 (allele)

B. JIA-OP HLA DRB1 SFVT analysis of tSFs ranked by overall p-values

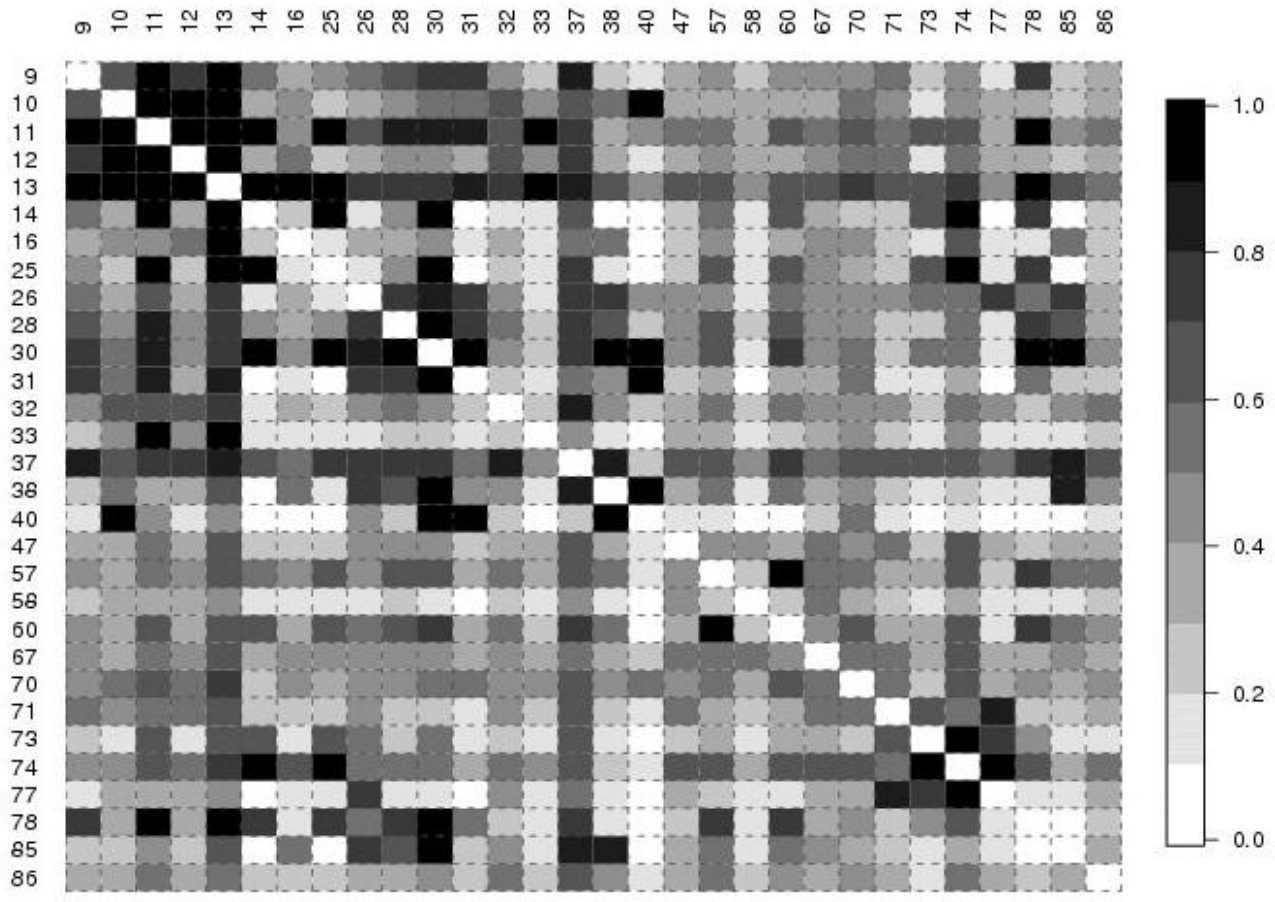
rank	SF #	Description	amino acid	p-value	max OR	min OR
1	t201		**13, 37**	5.6E-30	7.04	0.37
2	t205		**13, 37, 74, 86**	2.7E-29	6.69	0.37
3	t206		**13, 67, 74, 86**	1.2E-28	6.47	0.28
4	t224		**13, 37, 67**	1.3E-28	6.84	0.28
5	57	position 13	**13**	1.8E-28	4.91	0.33
6	t203		**13, 67**	2.3E-28	6.84	0.28
7	t225		**37, 67**	2.2E-28	3.90	0.31
8	t204		**13, 67, 86**	2.5E-28	6.69	0.28
9	t218		**13, 37, 57, 67, 74, 86**	3.6E-28	6.90	0.28
10	t207		**13, 37, 67, 74, 86**	3.3E-28	6.47	0.28
11	t215		**13, 57, 67,** 71**, 74, 86**	1.1E-27	9.40	0.28
12	1	allele		1.1E-27	9.40	0.28

E. Conditional haplotype analyses of HLA amino acid data

Many amino acids at DRB1 for JIA-OP, and DRB1 and DQB1 for T1D, show significant associations with disease, and continue to do so even with conditional analyses, as they are in LD with additional amino acids directly involved in differential disease risk. Additional work remains to take account of the complex LD pattern of amino acids at the classical HLA loci (see Figure V.E.1 below for DRB1 amino acid LD data, and Single et al. 2011 for DQB1 amino acid LD data). Note further, that when the LD is very high it may be impossible to distinguish between highly correlated amino acid sites; studies in other ethnic groups where the LD pattern may be different are then useful.

Figure V.E.1: JIA -OP HLA DRB1 amino acid LD[a]

.

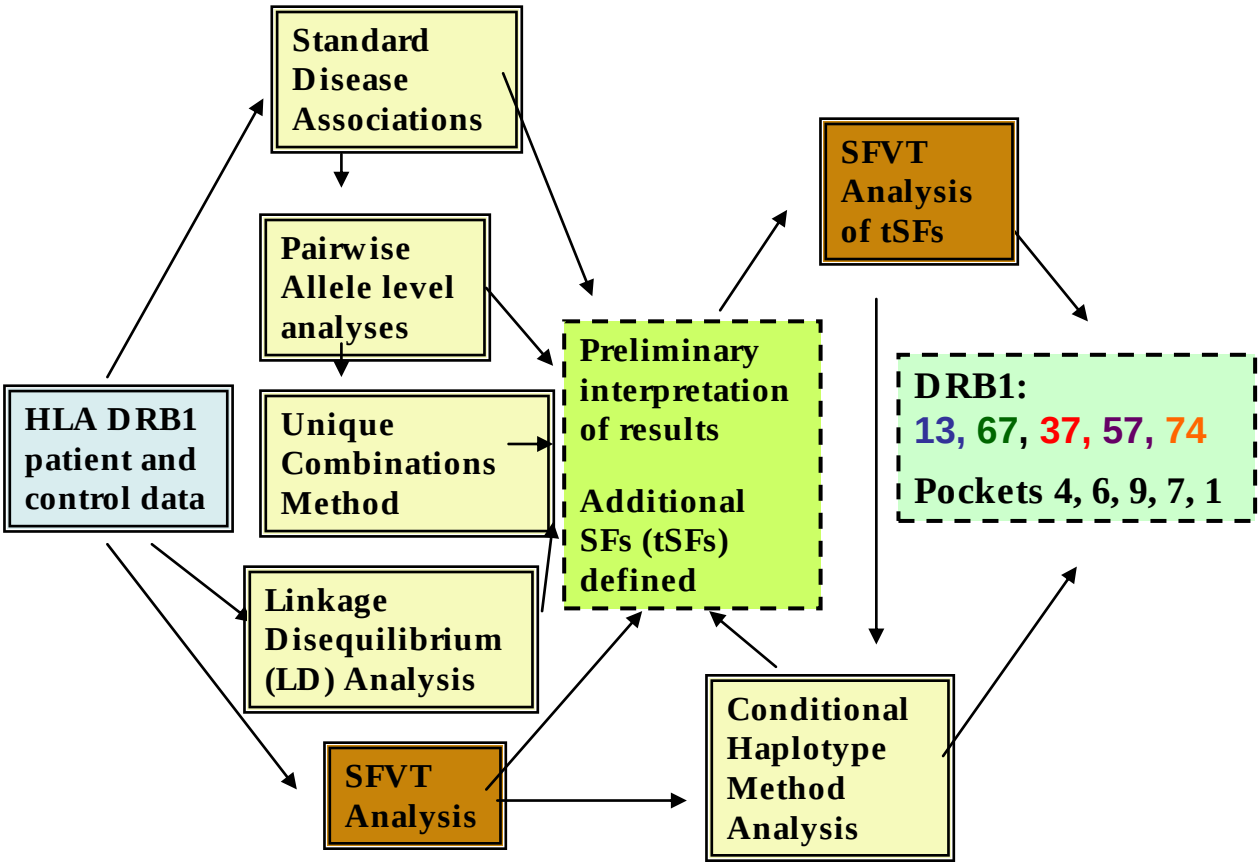


[a]Linkage Disequilibrium Measure Plot: Wn (All populations, DRB1) (Lancaster 2006)

As mentioned above, the application of a series of complementary methods (see Figure V.E.2 and Figure 1 of Thomson et al. 2010)—within serogroup comparisons (Section V.B), the Unique Combinations method (Section V.C), SFVT analyses (Section V.D), and Conditional Haplotype method (CHM) analyses —provides a powerful and systematic approach to the detection of the biologically relevant amino acids contributing to disease risk. For JIA-OP the

amino acids identified are (in numeric order): AAs 13 (pockets 4 and 6), 37 and 57 (both pocket 9), 67 (pocket 7), 74 (pocket 4), and 86 (pocket 1), and to a lesser extent 30 (pockets 6 and 7) and 71 (pockets 4, 5, and 7). In some cases we cannot exclude the involvement of other amino acids in high LD with these.

Figure V.E.2: Methods to apply to detect HLA amino acids directly involved in disease risk[a]



[a]See Thomson et al. (2010)

Application of different methods can help validate evidence of a direct role of specific amino acids and identify amino acids *additional* to those identified by for example SFVT analysis, particularly when there are interaction effects, e.g., DRB1 amino acid position 86 (Karp et al. 2010, Thomson et al. 2010, and unpublished data). However, the combinational magnitude of considering many combinations of amino acids, and CHM analyses thereof, is computationally intensive, and even when feasible, the interpretation of results is difficult. Also, for HLA data one quickly runs out of variation to test since HLA allelic variation can be uniquely defined by a limited number of amino acid sites (see Thomson et al. 2010, and Table V.E.1 below).

Table V.E.1: JIA -OP HLA DRB1 amino acid variation

		aa position [a]	13	67	74	86	37	57	30	71
JIA-OP	Controls	# alleles	6	3	4	2	5	5	5	4
12	1	DRB1*1103	S	F	A	V	Y	D	Y	E
102	13	DRB1*0801	G	F	L	G	Y	S	Y	R
57	11	DRB1*1104	S	F	A	V	Y	D	Y	R
9	3	DRB1*0403	H	L	E	V	Y	D	Y	R
90	38	DRB1*1301	S	I	A	V	N	D	Y	E
9	5	DRB1*0102	F	L	A	V	S	D	C	R
60	36	DRB1*1101	S	F	A	G	Y	D	Y	R
9	6	DRB1*0901	F	F	E	G	N	V	G	R
74	50	DRB1*0101	F	L	A	G	S	D	C	R
89	61	DRB1*0301	S	L	R	V	N	D	Y	K
10	8	DRB1*1201	G	I	A	V	L	V	H	R
28	23	DRB1*1302	S	I	A	G	N	D	Y	E
10	9	DRB1*1303	S	I	A	G	Y	S	Y	K
6	8	DRB1*1601	R	F	A	G	S	D	Y	R
11	18	DRB1*1401	S	L	E	V	F	A	Y	R
5	10	DRB1*1502	R	I	A	G	S	D	Y	A
7	16	DRB1*0404	H	L	A	V	Y	D	Y	R
38	80	DRB1*1501	R	I	A	V	S	D	Y	A
30	65	DRB1*0701	Y	I	Q	G	F	V	L	R
21	47	DRB1*0401	H	L	A	G	Y	D	Y	K
4	11	DRB1*0103	F	I	A	G	S	D	C	E

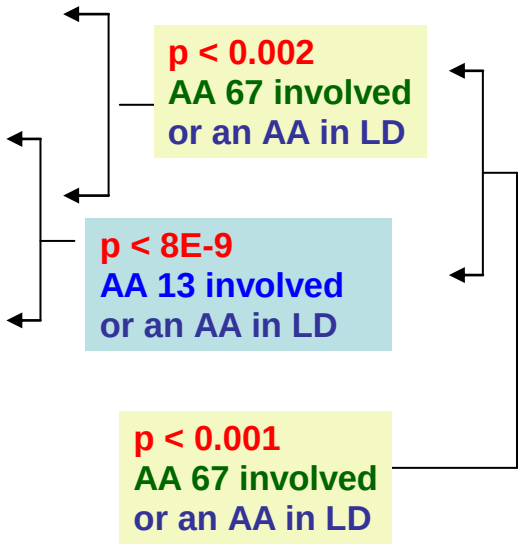
Preliminary CHM analyses have been applied to the JIA-OP HLA DRB1 amino acid data (Thomson et al. 2010) and to T1D data (unpublished). For JIA-OP pairwise conditional analyses of amino acids 13 and 67 show evidence of a direct role in disease risk (or high LD with other amino acids with a direct role in disease) (see Figure V.E.3). Note also the high blocks of LD of amino acids 10-13 in Figure V.E.1, and the strong associations of all these with JIA-OP. Amino acid 13 is more polymorphic than the others, so CHM analysis can be applied and indicates a direct role of amino acid 13 over the others. More work is required on how to systematically apply CHM analyses to the HLA amino acid data, and interpret the results.

Figure V.E.3: **Conditional haplotype method analysis of JIA -OP HLA DRB1 amino acids 13 and 67 variation**

DRB1 Amino Acids 13 and 67

13 - 67	patients	controls	OR
G - F	108	14	6.8
S - F	130	49	2.3
S - I	131	71	1.5
G - I	13	8	1.3
S - L	102	80	1.0
R - I	44	91	0.2
others	270	233	
Sum	798	546	

overall p < 2E-28



References

Ahmad T, Neville M, Marshall SE, et al. (2003) Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC**.** Hum. Mol. Genet. 12 (6): 647-656.

Amiel JL (1967) Study of the leucocyte phenotypes in Hodgkin's disease, in "Histocompatibility Testing 1967" (E. S. Curtoni, P. L. Mattiuz, and R. M. Tosi, Eds.), Munksgaard, Copenhagen.

Barcellos, L. F., Sawcer, S., Ramsay, P. P., Baranzini, S. E., Thomson, G., Briggs, F., Cree, B. C., Begovich, A. B., Villoslada, P., Montalban, X., et al. (2006). Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis, Hum. Mol. Genet. 15(18)**:** 2813-2824.

Barcellos LF, May SL, Ramsay PP, Quach HL, Lane JA, Nititham J, Noble JA, Taylor KE, Quach DL, Chung SA, Kelly JA, Moser KL, Behrens TW, Seldin MF, Thomson G, Harley JB, Gaffney PM, Criswell LA. 2009. High density SNP screening of the major histocompatibility complex in systemic lupus erythematosus demonstrates strong evidence for independent susceptibility regions. PlosGenetics 5(10): e1000696. PMID: 19851445

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

Begovich, A. B., Carlton, V. E., Honigberg, L. A., Schrodi, S. J., Chokkalingam, A. P., Alexander, H. C., Ardlie, K. G., Huang, Q., Smith, A. M., Spoerke, J. M., et al. (2004). A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. Am. J. Hum. Genet. 75(2)**,** 330-337.

Begovich, A. B., Caillier, S. J., Alexander, H. C., Penko, J. M., Hauser, S. L., Barcellos, L. F., and Oksenberg, J. R. (2005). The R620W polymorphism of the protein tyrosine phosphatase PTPN22 is not associated with multiple sclerosis. Am. J. Hum. Genet. 76(1)**,** 184-187.

Bennett, S. T., Wilson, A. J., Cucca, F., Nerup, J., Pociot, F., McKinney, P. A., Barnett, A. H., Bain, S. C., and Todd, J. A. (1996). IDDM2-VNTR-encoded susceptibility to type 1 diabetes: dominant protection and parental transmission of alleles of the insulin gene-linked minisatellite locus. J. Autoimmun. 9(3)**,** 415-421.

Blacker, D., Haines, J. L., Rodes, L., Terwedow, H., Go, R. C., Harrell, L. E., Perry, R. T., Bassett, S. S., Chase, G., Meyers, D., et al. (1997). ApoE-4 and age at onset of Alzheimer's disease: the NIMH genetics initiative. Neurology 48(1)**,** 139-147.

Blomhoff A, Olsson M, Johansson S, Akselsen HE, Pociot F, Nerup J, Kockum I, Cambon-Thomsen A, Thorsby E, Undlien DE, Lie BA. (2006) Linkage disequilibrium and haplotype blocks in the MHC vary in an HLA haplotype specific manner assessed mainly by DRB1*03 and DRB1*04 haplotypes. Genes Immun. 2006 Mar;7(2):130-40.

Bourgey, M., Pedry, H., and Clerget-Darpoux, F. (2007). Modeling the effect of PTPN22 in rheumatoid arthritis. In: Genetic Analysis Workshop 15: Gene Expression Analysis and

Approaches to Detecting Multiple Functional Loci, (H.J. Cordell, M. de Andrade, M-C. Babron, C.W. Bartlett, J. Beyene, H. Bickeböller, R. Culverhouse, L.A. Cupples, E.W. Daw, J. Dupuis, C.T. Falk, S. Ghosh, K.A. Goddard, E.L. Goode, E.R. Hauser, L.J. Martin, M. Martinez, K.E. North, N.L. Saccone, S. Schmidt, W. Tapper, D. Thomas, D. Tritchler, V.J. Vieland, E.M. Wijsman, M.A. Wilcox, J.S. Witte, Q. Yang, A. Ziegler, L. Almasy, J.W. MacCluer, eds.), . BMC Proceedings 2007, 1(Suppl 1): S37-43.

Bronson PG, Ramsay PP, Thomson G, Barcellos LF, and the T1DGC. 2009. Analysis of maternal-offspring HLA compatibility, parent-of-origin and non-inherited maternal effects for the classical HLA loci in type 1 diabetes. Diabetes, Obesity and Metabolism 11 (Supplement 1): 74-83. PMID: 19143818

Browning, S. R. (2006). Multilocus association mapping using variable-length Markov chains. Am. J. Hum. Genet. 78(6)**,** 903-913.

Browning, S. R., Briley, J. D., Briley, L. P., Chandra, G., Charnecki, J. H., Ehm, M. G., Johansson, K. A., Jones, B. J., Karter, A. J., Yarnall, D. P., et al. (2005). Case-control single-marker and haplotypic association analysis of pedigree data. Genet. Epidemiol. 28(2)**,**110-22.

Cano P, Klitz W, Mack SJ, et al. (2007) Common and wel-documented HLA alleles: report of the ad-hoc committee of the American Society of Histocompatibility and Immunologeneitcs. Hum Immunol 68: 392-417.

Carlton, V. E., Hu, X., Chokkalingam, A. P., Schrodi, S. J., Brandon, R., Alexander, H. C., Chang, M., Catanese, J. J., Leong, D. U., Ardlie, K. G., et al. (2005). PTPN22 genetic variation: evidence for multiple variants associated with rheumatoid arthritis. Am. J. Hum. Genet. 77(4), 567-581.

Clerget-Darpoux, F., Babron, M. C., Prum, B., Lathrop, G. M., Deschamps, I., and Hors, J. (1988). A new method to test genetic models in HLA associated diseases: the MASC method. Ann. Hum. Genet. 52(3)**,** 247-258.

Clerget-Darpoux, F., Babron, M. C., Deschamps, I., and Hors, J. (1991). Complementation and maternal effect in insulin-dependent diabetes. Am. J. Hum. Genet. **49**(1)**,** 42-48.

Clerget-Darpoux, F., Babron, M. C., and Bickeboller, H. (1995). Comparing the power of linkage detection by the transmission disequilibrium test and the identity-by-descent test. Genet. Epidemiol. 12(6)**:** 583-588.

Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, Allen JE, Downes K, Barrett JC, Healy BC, Mychaleckyj JC, Warram JH, Todd JA. 2008. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nat Genet. 40(12):1399-1401.

Cudworth AG, Woodrow JC. Evidence for HL-A-linked genes in ‘‘juvenile’’ diabetes mellitus. Br Med J 1975; 3:133–135.

Dausset J, Svejgaard A (eds). 1977. HLA and Disease. Munksgaard, Copenhagen

Dizier, M. H., Babron, M. C., and Clerget-Darpoux, F. (1994). Interactive effect of two candidate genes in a disease: extension of the marker-association-segregation chi^2 method. Am. J. Hum. Genet. 55(5)**,** 1042-1049.

Doherty PC, Zinkernagel RM. (1975) Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature 256, 50–52.

du Montcel, S. T., Michou, L., Petit-Teixeira, E., Osorio, J., Lemaire, I., Lasbleiz, S., Pierlot, C., Quillet, P., Bardin, T., Prum, B., et al. (2005). New classification of HLA-DRB1 alleles supports the shared epitope hypothesis of rheumatoid arthritis susceptibility. Arthritis Rheum. 52(4), 1063-1068.

Erlich H, Valdes AM, Noble J, et al. (2008) HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk. Diabetes 57: 1084-1092

Fajardy, I., Vambergue, A., Stuckens, C., Weill, J., Danze, P. M., and Fontaine, P. (2002). CTLA-4 49 A/G dimorphism and type 1 diabetes susceptibility: a French case-control study and segregation analysis. Evidence of a maternal effect. Eur. J. Immunogenet. 29(3), 251-257.

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227-233

Field LL (1989) Genes predisposing to IDDM in multiplex families. Genet Epidemiol 6:101-106

Field LL, Fothergill-Payne C, Bertrams J, Baur MP (1986) HLA-DR effects in a large German IDDM data set. Genet Epidemiol Suppl 1:323-328

Fingerlin, T. E., Boehnke, M., and Abecasis, G. R. (2004). Increasing the power and efficiency of disease-marker case-control association studies through use of allele-sharing information. Am. J. Hum. Genet. 74(3),432-43.

Fisher, R. (1970). "Statistical Methods for Research Workers." Oliver and Boyd, Edinburgh.

Frahm N, B. Baker, C. Brander (2008) Identification and optimal definition of HIV-derived cytotoxic T-lymphocyte (CTL) epitopes for the study of CTL escape, functional avidity and viral evolution. In: HIV Molecular Immunology 2008, Los Alamos, NM, 3-24

Fujisawa, T., Ikegami, H., Kawaguchi, Y., Yamato, E., Takekawa, K., Nakagawa, Y., Hamada, Y., Ueda, H., Shima, K., and Ogihara, T. (1995). Class I HLA is associated with age-at-onset of IDDM, while class II HLA confers susceptibility to IDDM. Diabetologia 38(12), 1493-1495.

Gaudieri S, Desantis D, McKinnon E, Moore C, Nolan D, Witt CS, Mallal SA, Christiansen FT (2005) Killer immunoglobulin-like receptors and HLA act both independently and synergistically to modify HIV disease progression. Genes Immun 6:683-690

Gedil MA, Steiner NK, Hurley CK (2007) KIR3DL2: diversity in a hematopoietic stem cell transplant population. Tissue Antigens 70:228-232

Genin, E., Babron, M. C., McDermott, M. F., Mulcahy, B., Waldron-Lynch, F., Adams, C., Clegg, D. O., Ward, R. H., Shanahan, F., Molloy, M. G., et al. (1998). Modeling the major histocompatibility complex susceptibility to RA using the MASC method. Genet. Epidemiol. 15(4), 419-430.

Gourraud PA, Gagne K, Bignon JD, Cambon-Thomsen A, Middleton D. (2007) Preliminary analysis of a KIR haplotype estimation algorithm: a simulation study. Tissue Antigens 69 Suppl 1: 96-100.

Gourraud P-A, Hollenbach JA, Barnetche T, Single RM, Mack SJ. 2011. Standard methods for the management of immunogenetic data. In: Methods in Molecular Genetics: Immunogenetics, ed. Walker J, Humana press USA, in press.

Greenberg, D. A. (1993). Linkage analysis of "necessary" disease loci versus "susceptibility" loci. Am. J. Hum. Genet. 52(1), 135-143.

Greenberg, D. A., and Doneshka, P. (1996). Partitioned association-linkage test: distinguishing "necessary" from "susceptibility" loci. Genet. Epidemiol. 13(3)**,** 243-252.

Gregersen, P. K., Silver, J., and Winchester, R. J. (1987). The shared epitope hypothesis: an approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. Arthritis Rheum. 30, 1205-1213.

Grote M, Klitz W, Thomson G. 1998. Constrained disequilibrium values and hitchhiking in a three-locus system. Genetics 150: 1295-1307. PMID: 9799280

Guillet, J. G., Lai, M. Z., Briner, T. J., Smith, J. A. & Gefter, M. L. (1986). Interaction of Peptide Antigensand Class II Major Histocompatibility Complex Antigens. Nature 324, 260–262.

Hamza TH, Zabetian CP, Tenesa A, et al. (2010) Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. Nat Genet 42(9): 781-786

Hanifi Moghaddam, P., de Knijf, P., Roep, B. O., Van der Auwera, B., Naipal, A., Gorus, F., Schuit, F., and Giphart, M. J. (1998). Genetic structure of IDDM1: two separate regions in the major histocompatibility complex contribute to susceptibility or protection. Belgian Diabetes Registry. Diabetes 47(2)**,** 263-269.

Harney, S., Newton, J., Milicic, A., Brown, M. A., and Wordsworth, B. P. (2003). Non-inherited maternal HLA alleles are associated with rheumatoid arthritis. Rheumatology (Oxford) 42(1)**,** 171-174.

Harris, E. E., and Meyer, D. (2006). The molecular signature of selection underlying human adaptations. Am. J. Phys. Anthropol. Suppl. 43, 89-130.

Hedrick, P. W. (1987). Gametic disequilibrium measures: proceed with caution. Genetics 117(2)**,** 331-341.

Hedrick, P. W. (2006). Genetic polymorphism in heterogeneous environments: the age of genomics. Annu. Rev. Ecol. Evol. Syst. 37, 67-93.

Hermann, R., Veijola, R., Sipila, I., Knip, M., Akerblom, H. K., Simell, O., and Ilonen, J. (2003) To: Pani MA, Van Autreve J, Van Der Auwera BJ, Gorus FK, Badenhoop K (2002) Non-transmitted maternal HLA DQ2 or DQ8 alleles and risk of Type I diabetes in offspring: the importance of foetal or post partum exposure to diabetogenic molecules. Diabetologia 45:1340-1343. Diabetologia 46(4)**,** 588-589; author reply 591-592.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. Genet. Med. 4(2)**,** 45-61.

Hodge, S. E. (1993). Linkage analysis versus association analysis: distinguishing between two models that explain disease-marker associations. Am. J. Hum. Genet. 53(2)**,** 367-384.

Hollenbach JA, Ladner LM, Saeteurn K, Taylor KD, Mei L, Haritunians T, McGovern DPB, Erlich HA, Rotter JI, Trachtenberg EA (2009) Susceptibility to Crohn's Disease is mediated by KIR2DL2/KIR2DL3 heterozygosity and the HLA-C ligand. Immunogenetics 61(10): 663-671

Hollenbach JA, Thompson SD, Bugawan TL, Ryan M, Sudman M, Marion M, Langefeld CD, Thomson G, Erlich HA, Glass DN. 2010. Juvenile idiopathic arthritis and HLA class I and class II interaction and age of onset effects. Arthritis and Rheumatism 62(6): 1781-1791. PMID: 20191588

Hollenbach JA, Mack SJ, Thomson G, Gourraud P-A. 2011a. Methods for disease association studies with immunogenetic data. In: Methods in Molecular Genetics: Immunogenetics, ed. Walker J, Humana press USA, in press.

Hollenbach JA, Mack SJ, Gourraud P-A, Single RM, Maiers M, Middleton D, Thomson G, Marsh SGE, Varney MD; for the Immunogenomics Data Analysis Working Group*. 2011b. A Community Standard for Immunogenomic Data-reporting and Analysis: Proposal for a STrengthening the REporting of Immunogenomic Studies (STREIS) statement. Tissue Antigens, under revision.

Horn, G. T., T. L. Bugawan, C. M. Long, and H. A. Erlich. 1988. Allelic sequence variation of the HLA-DQ loci: relationship to serology and to insulin-dependent diabetes susceptibility. Proc. Natl. Acad. Sci. USA 85: 6012-6016.

Hsu, L., Li, H., and Houwing-Duistermaat, J. J. (2002). A method for incorporating ages at onset in affected sibpair linkage studies. Hum. Hered. 54(1)**,** 1-12.

Imboden JB (2009) The immunopathogeneis of rheumatoid arthritis. Ann Rev Pathol Mech Dis 4: 417-434

International Multiple Sclerosis Genetic Consortium (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Eng J Med 357: 851-862.

Jawaheer, D., Li, W., Graham, R. R., Chen, W., Damle, A., Xiao, X., Monteiro, J., Lee, A., Lundsten, R., Begovich, A., et al. (2002). Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. Am. J. Hum. Genet. 71(3)**,** 585-594.

Johansson, S., Lie, B. A., Todd, J. A., Pociot, F., Nerup, J., Cambon-Thomsen, A., Kockum, I., Akselsen, H. E., Thorsby, E., and Undlien, D. E. (2003). Evidence of at least two type 1 diabetes susceptibility genes in the HLA complex distinct from HLA-DQB1, -DQA1 and -DRB1. Genes Immun. 4(1)**,** 46-53.

Karp DR, Marthandan N, Marsh SGE, Ahn C, Arnett FC, DeLuca DS, Diehl AD, Dunivin R, Eilbeck K, Feolo M, Guidry PA, Helmberg W, Lewis S, Mayes MD, Mungall C, Natale DA, Peters B, Petersdorf E, Reveille JD, Smith B, Thomson G, Waller MJ, Scheuermann RH. 2010. Novel sequence feature variant type analysis of the HLA genetic association in systemic sclerosis. Human Molecular Genetics 19(4): 707-719. PMID: 19933168

Khakoo SI, Carrington M (2006) KIR and disease: a model system or system of models? Immunol Rev 214:186-201

Khakoo SI, Chloe LT, Martin MP, Brooks CR, Gao X, Astemborski J, Cheng J, Goedert JJ, Vlahov D, Hilgartner M, Cox S, Little AM, Alexander GJ, Cramp ME, O'Brien S, Rosenberg WMC, Thomas DL, Carrington M (2004) HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. Science 305:872-874

Khoury MJ (1994) Case-parental control method in the search for disease-susceptibility genes. Am J Hum Genet 55: 414-415

Klein J (1975) "Biology of the mouse Histocompatibility-2 Complex." Springer-Verlag, Berlin

Klitz W, Thomson G. 1987. Disequilibrium pattern analysis. II. Application to Danish HLA-A and B locus data. Genetics 116: 633-643. PMID: 3476350

Knapp M, Seuchter SA, Baur MP (1993) The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. Am J Hum Genet 52:1085-1093

Koeleman, B. P., Dudbridge, F., Cordell, H. J., and Todd, J. A. (2000a). Adaptation of the extended transmission/disequilibrium test to distinguish disease associations of multiple loci: the Conditional Extended Transmission/Disequilibrium Test. Ann. Hum. Genet. 64(3)**,** 207-213.

Koeleman, B. P., Herr, M. H., Zavattari, P., Dudbridge, F., March, R., Campbell, D., Barnett, A. H., Bain, S. C., Mulargia, A. P., Loddo, M., et al. (2000b). Conditional ETDT analysis of the human leukocyte antigen region in type 1 diabetes. Ann. Hum. Genet. 64(3)**,** 215-221.

Kunert K, Seiler M, Mashreghi MF, Klippert K, Schonemann C, Neumann K, Pratschke J, Reinke P, Volk HD, Kotsch K (2007) KIR/HLA ligand incompatibility in kidney transplantation. Transplantation 84:1527-1533

Lambert, A. P., Gillespie, K. M., Bingley, P. J., and Gale, E. A. (2003). To: Pani MA, Van Autreve J, Van der Auwera BJ, Gorus FK, Badenhoop K (2002) Non-transmitted maternal HLA DQ2 or DQ8 alleles and risk of Type 1 diabetes in offspring: the importance of foetal or post partum exposure to diabetogenic molecules. Diabetologia 45:1340-1343. Diabetologia 46(4)**,** 590-591; author reply 591-592.

Lancaster A. 2006. Interplay of selection and molecular function in HLA genes. PhD thesis, University of California at Berkeley.

Lancaster, A., Nelson, M. P., Meyer, D., Thomson, G., and Single, R. M. (2003). PyPop: a software framework for population genomics: analyzing large-scale multi-locus genotype data. Pac. Symp. Biocomput.**,** 514-525.

Lancaster A, Nelson MP, Single RM, Meyer D, Thomson G. 2007a. Software framework for the Biostatistics Core of the International Histocompatibility Working Group. In: Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress, Volume I, ed. Hansen JA. IHWG Press, Seattle, WA, pp. 510-517.

Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. 2007b. PyPop update - a software pipeline for large-scale multi-locus population genomics. Tissue Antigens 69 (Suppl. 1): 192-197. PMID: 17445199

Lanier LL (1999) Natural killer cells fertile with receptors for HLA-G? Proc Natl Acad Sci U S A 96:5343-5345

Lechler R, ed. (1994) HLA and disease. Academic Press, London.

Leisner C, N. Loeth, K. Lamberth, et al., (2008) One-pot, mix-and-read peptide-MHC teramers. PLoS ONE 3(2), e1678

Li, C., Scott, L. J., and Boehnke, M. (2004). Assessing whether an allele can account in part for a linkage signal: the Genotype-IBD Sharing Test (GIST). Am. J. Hum. Genet. 74(3)**,** 418-431.

Li, H. (1999). The additive genetic gamma frailty model for linkage analysis of age-of-onset variation. Ann. Hum. Genet. 63(5)**,** 455-468.

Li, H., and Hsu, L. (2000). Effects of age at onset on the power of the affected sib pair and transmission/disequilibrium tests. Ann. Hum. Genet. 64(3)**,** 239-254.

Li, M., Boehnke, M., and Abecasis, G. R. (2005). Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. Am. J. Hum. Genet. 76(6)**,** 934-949.

Li Y, Zhang T, Ho C, Orange JS, Douglas SD, Ho WZ (2004) Natural killer cells inhibit hepatitis C virus expression. J Leukoc Biol 76:1171-1179

Liang, K. Y., Chiu, Y. F., and Beaty, T. H. (2001). A robust identity-by-descent procedure using affected sib pairs: multipoint mapping for complex diseases. Hum. Hered. 51(1-2)**,** 64-78.

Lie, B. A., Sollid, L. M., Ascher, H., Ek, J., Akselsen, H. E., Ronningen, K. S., Thorsby, E., and Undlien, D. E. (1999a). A gene telomeric of the HLA class I region is involved in predisposition to both type 1 diabetes and coeliac disease. Tissue Antigens 54(2)**,** 162-168.

Lie, B. A., Todd, J. A., Pociot, F., Nerup, J., Akselsen, H. E., Joner, G., Dahl-Jorgensen, K., Ronningen, K. S., Thorsby, E., and Undlien, D. E. (1999b). The predisposition to type 1 diabetes linked to the human leukocyte antigen complex includes at least one non-class II gene. Am. J. Hum. Genet. 64(3)**,** 793-800.

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., and Hirschhorn, J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat. Genet. 33(2)**,** 177-182.

Louis, E. J., Thomson, G., and Payami, H. (1983). The affected sib method. II. The intermediate model. Ann. Hum. Genet. 47, 225-243.

Louis EJ, Thomson G. 1986. The three allele synergistic mixed model for insulin dependent diabetes mellitus. Diabetes 35: 958-963. PMID: 3488932

Mack SJ, Gourraud P-A, Single RM, Thomson G, Hollenbach JA. 2011. Analytical methods for immunogenetic population data. In: Methods in Molecular Genetics: Immunogenetics, ed. Walker J, Humana Press USA, in press.

Malkki, M., Single, R., Carrington, M., Thomson, G., and Petersdorf, E. (2005). MHC microsatellite diversity and linkage disequilibrium among common HLA-A, HLA-B, DRB1 haplotypes: implications for unrelated donor hematopoietic transplantation and disease association studies. Tissue Antigens 66(2)**,** 114-124.

Martin, E. R., Monks, S. A., Warren, L. L., and Kaplan, N. L. (2000). A test for linkage and association in general pedigrees: The pedigree disequilibrium test. Am. J. Hum. Genet. 67(1)**,** 146-154.

Martin, E. R., Bass, M. P., Gilbert, J. R., Pericak-Vance, M. A., and Hauser, E. R. (2003). Genotype-based association test for general pedigrees: The genotype-PDT. Genet. Epidemiol. 25(3)**,** 203-213.

Meyer, D., and Thomson, G. (2001). How selection shapes variation of the human major histocompatibility complex: a review. Ann. Hum. Genet. 65(1)**,** 1-26.

Meyer, D., Single, R. M., Mack, S. J., Erlich, H. A., and Thomson, G. (2006). Signatures of demographic history and natural selection in the human major histocompatibility complex loci. Genetics 173(4)**,** 2121-2142.

Meyer, D., Single, R., Mack, S. J., Lancaster, A., Nelson, M. P., Fernandez-Vina, M., Erlich, H. A., and Thomson, G. (2007). Single locus polymorphism of classical HLA genes. In "Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress, Volume I" (J. A. Hansen, ed.), pp. 653-704. IHWG Press, Seattle.

Mignot, E., Lin, L., Rogers, W., Honda, Y., Qiu, X., Lin, X., Okun, M., Hohjoh, H., Miki, T., Hsu, S., et al. (2001). Complex HLA-DR and -DQ interactions confer risk of narcolepsy-cataplexy in three ethnic groups. Am. J. Hum. Genet. 68(3)**,** 686-699.

Mignot, E., Lin, L., Li, H., Thomson, G., Lathrop, M., Thorsby, E., Tokunaga, K., Honda, Y., Dauvilliers, Y., Tafti, M., et al. (2007). HLA allele and microsatellite studies in narcolepsy. In "Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress, Volume I" (J.A. Hansen, ed.), pp. 817-812. IHWG Press, Seattle.

Moffett A, Hiby SE (2007) How does the maternal immune system contribute to the development of pre-eclampsia? Placenta 28 Supp1: S51-S56

Morris, A. P. (2006). A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. Am. J. Hum. Genet. 79(4)**,** 679-694.

Motro, U., and Thomson, G. (1985). The affected sib method. I. Statistical features of the affected sib- pair method. Genetics 110(3)**,** 525-538.

Nakanishi, K., Kobayashi, T., Inoko, H., Tsuji, K., Murase, T., and Kosaka, K. (1995). Residual beta-cell function and HLA-A24 in IDDM. Markers of glycemic control and subsequent development of diabetic retinopathy. Diabetes 44(11)**,** 1334-1339.

Nelson, J. L., Lambert, N. C., Brautbar, C., El-Gabalaway, H., Fraser, P., Gorodezky, C., Li, H., Jonas, B., Konenkov, V., Lathrop, M., et al. (2007). The 13th International Histocompatibility Working Group for rheumatoid arthritis joint report. In "Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress, Volume I" (J. A. Hansen, ed.), pp. 797-804. IHWG Press, Seattle.

Nemes S JJ, Genell A, Steineck G (2009) Bias in odds ratios by logistic regression modelling and sample size. BMC Medical Research Methodology 9: 56

Nielsen M, C. Lundegaard, P. Worning, et al. (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. Bioinformatics **20**, 1388-1397

Noble, J. A., Valdes, A. M., Cook, M., Klitz, W., Thomson, G., and Erlich, H. A. (1996). The role of HLA class II genes in insulin-dependent diabetes mellitus: molecular analysis of 180 Caucasian, multiplex families. Am. J. Hum. Genet. 59(5)**,** 1134-1148.

Noble, J. A., Valdes, A. M., Thomson, G., and Erlich, H. A. (2000). The HLA class II locus DPB1 can influence susceptibility to type 1 diabetes. Diabetes 49(1)**,** 121-125.

Noble, J. A., Valdes, A. M., Bugawan, T. L., Apple, R. J., Thomson, G., and Erlich, H. A. (2002). The HLA class I A locus affects susceptibility to type 1 diabetes. Hum. Immunol. 63(8)**,** 657-664.

O'Connell JR, Weeks DE. 1998. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Gen 63: 259-266.

Oksenberg, J. R., Barcellos, L. F., Cree, B. A. C., Baranzini, S. E., Bugawan, T. L., Khan, O., Lincoln, R. R., Swerdlin, A., Mignot, E., Lin, L., et al. (2004). Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans. Am. J. Hum. Genet. 74(1)**,** 160-167.

Ott J (1989) Statistical properties of the haplotype relative risk. Genet Epidemiol 6:127-130

Pani, M. A., Van Autreve, J., Van der Auwera, B. J., Gorus, F. K., and Badenhoop, K. (2002). Non-transmitted maternal HLA DQ2 or DQ8 alleles and risk of Type I diabetes in offspring: the importance of foetal or post partum exposure to diabetogenic molecules. Diabetologia 45(9)**,** 1340-1343.

Payami, H., Thomson, G., Motro, U., Louis, E. J., and Hudes, E. (1985). The affected sib method. IV. Sib trios. Ann. Hum. Genet. 49(4)**,** 303-314.

Payami H, Khan MA, Grennan DM, Sanders PA, Dyer PA, Thomson G. 1987. Analysis of genetic interrelationship among HLA-associated diseases. American Journal of Human Genetics 41: 331-349. PMID: 3631074

Payami, H., Joe, S., Farid, N. R., Stenszky, V., Chan, S. H., Yeo, P. P., Cheah, J. S., and Thomson, G. (1989). Relative predispositional effects (RPEs) of marker alleles with disease: HLA-DR alleles and Graves disease. Am. J. Hum. Genet. 45(4)**,** 541-546.

Payami, H., Zhu, M., Montimurro, J., Keefe, R., McCulloch, C. C., and Moses, L. (2005). One step closer to fixing association studies: evidence for age- and gender-specific allele frequency variations and deviations from Hardy-Weinberg expectations in controls. Hum. Genet. 118(3-4)**,** 322-330.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. 69(1)**,** 124-137.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Pugliese, A., Dorman, J. S., Steenkiste, A., Li, H., Thorsby, E., Lathrop, M., Schoch, G., Thomson, G., Caillat-Zucman, S., Awdeh, Z., et al. (2007). Joint Report of the 13th IHWS type 1 diabetes (T1D) component. In "Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress, Volume I" (J. A. Hansen, ed.), pp 788-796. IHWG Press, Seattle.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575

Ramagopalan SV, Knight JC, Ebers GC. 2009. Multiple sclerosis and the major histocompatibility complex. Curr Opin Neurol. 22(3):219-225.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science 273(13 September)**,** 1516-1517.

Robinson WP, Asmussen M, Thomson G. 1991a. Three locus systems impose additional constraints on pairwise disequilibria. Genetics 129: 925-930. PMID: 1752428

Robinson WP, Cambon-Thomsen A, Borot N, Klitz W, Thomson G. 1991b. Selection, hitchhiking and disequilibrium analysis at three linked loci with application to HLA data. Genetics 129: 931-948. PMID: 1752429

Robinson, W. P., Barbosa, J., Rich, S. S., and Thomson, G. (1993). Homozygous parent affected sib pair method for detecting disease predisposing variants: application to insulin dependent diabetes mellitus. Genet. Epidemiol. 10(5)**,** 273-288.

Rotter JI, Anderson CE, Rubin R, et al. (1983) HLA genotype distribution of insulin-dependent diabetes. The excess of DR3/DR4 heterozygotes allows rejection of the recessive hypothesis. Diabetes 32: 169-174.

Rubinstein P, Walker M, Carpenter C, Carrier C, Krassner J, Falk C, Ginsberg F (1981) Genetics of HLA disease association: the use of the haplotype relative risk (HRR) and the "Haplo-Delta" (Dh) estimates in juvenile diabetes from three racial groups. Hum Immunol 3:384

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909)**,** 832-837.

Salamon H, Tarhio J, Rønningen K, Thomson G. 1996. On distinguishing unique combinations in biological sequences. Journal of Computational Biology 3: 407-423. PMID: 8891958

Sasaki, T., Nemoto, M., Yamasaki, K., and Tajima, N. (1999). Preferential transmission of maternal allele with DQA1*0301-DQB1*0302 haplotype to affected offspring in families with type 1 diabetes. J. Hum. Genet. 44(5)**,** 318-322.

Schaid DJ, Sommer SS (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet 53:1114-1126

Schaid DJ, Sommer SS (1994) Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 55:402-409

Sekigawa, I., Okada, M., Ogasawara, H., Kaneko, H., Hishikawa, T., and Hashimoto, H. (2003). DNA methylation in systemic lupus erythematosus. Lupus 12(2)**,** 79-85.

Sledin MF, Price AL (2008) Application of ancestry informative markers to association studies oin European Americans. PLos Genetics 4(1): e5

Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, Holmans PA, Whittemore AS, Mowry BJ, Olincy A, Amin F, Cloninger CR, Silverman JM, Buccola NG, Byerley WF, Black DW, Crowe RR, Oksenberg JR, Mirel DB, Kendler KS, Freedman R, Gejman PV. 2009. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460(7256):753-757.

Silman, A. J., Hay, E. M., Worthington, J., Thomson, W., Pepper, L., Davidson, J., Dyer, P. A., and Ollier, W. E. (1995). Lack of influence of non-inherited maternal HLA-DR alleles on susceptibility to rheumatoid arthritis. Ann. Rheum. Dis. 54(4)**,** 311-313.

Single RM, Martin MP, Gao X, Meyer D, Yeager M, Kidd JR, Kidd KK, Carrington M. 2007a. Global diversity and evidence for co-evolution of KIR and HLA genes. Nature Genetics 39(9): 1114-1119.

Single, R., Meyer, D., Mack, S., Lancaster, A., Nelson, M. P., Fernandez-Vina, M., Erlich, H. A., and Thomson, G. (2007b). Haplotype frequencies and linkage disequilibrium. In "Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress, Volume I" (J.A. Hansen, ed.), pp 705-746. IHWG Press, Seattle.

Single R, Meyer D, Thomson G. 2007c. Statistical methods for analysis of population genetic data. In: Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress, Volume I, ed. Hansen JA. IHWG Press, Seattle, WA, pp. 518-522.

Single RM, Martin MP, and Carrington M. 2008. Statistical and population genetic methods for KIR genes typed for presence/absence. Immunogenetics 60(12): 711-725. PMCID: PMC2663517

Single RM, Gourraud P-A, Maldonado-Torres H, Lancaster AK, Briggs F, Barcellos LF, Hollenbach JA, Mack SJ, Thomson G. 2011. Estimating Haplotype Frequencies and Linkage Disequilibrium Parameters in the HLA and KIR Regions (Haplotype Estimation and Linkage Disequilibrium Methods Manual, Version 0.1.0 – June 3, 2011) https://www.immport.org/docs/standards/MethodsManual_HaplotypeFreqs+LD_v8.pdf

Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, Thomson G. 2008. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: A meta-analytic review of 497 population studies. Human Immunology 69: 443-464. PMID: 18638659

Snyder JA, Demchuk E, McCanlies EC, et al. (2008) Risk of chronic beryllium disease by HLA-DPB1 E69 genotype and beryllium exposure. J R Soc Interface 5(24): 749-758.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. 52**,** 506-516.

Steenkiste, A., Valdes, A. M., Feolo, M., Hoffman, D., Concannon, P., Noble, J., Schoch, G., Hansen, J., Helmberg, W., Dorman, J. S., et al. (2007). The HLA component of type 1 diabetes: Report on the activities of the 13th International Histocompatibility Group and 14th International Histocompatibility Workshop. Tissue Antigens 69**,** 214-225.

Sun JY, Gaidulis L, Dagis A, Palmer J, Rodriguez R, Miller MM, Forman SJ, Senitzer D (2005) Killer Ig-like receptor (KIR) compatibility plays a role in the prevalence of acute GVHD in unrelated hematopoietic cell transplants for AML. Bone Marrow Transplant 36:525-530

Svejgaard A, Platz P, Ryder LP (1980) Insulin-dependent diabetes mellitus. In: Histocomaptibility Testing 1980, Terasaki P (ed), UCLA Tissue typing Lab, Los Angeles, pp 638-656.

Svejgaard A, Ryder LP (1981) HLA genotype distribution and genetic models of insulin-dependent diabetes mellitus. Ann Hum Genet 45: 293-298.

Tait, B. D., Harrison, L. C., Drummond, B. P., Stewart, V., Varney, M. D., and Honeyman, M. C. (1995). HLA antigens and age at diagnosis of insulin-dependent diabetes mellitus. Hum. Immunol. 42(2)**,** 116-122.

ten Wolde, S., Breedveld, F. C., de Vries, R. R., D'Amaro, J., Rubenstein, P., Schreuder, G. M., Claas, F. H., and van Rood, J. J. (1993). Influence of non-inherited maternal HLA antigens on occurrence of rheumatoid arthritis. Lancet 341(8839)**,** 200-202.

Thomson G. 1980. A two locus model for juvenile diabetes. Annals of Human Genetics 43(4): 383-398. PMID: 7396412

Thomson G. 1981. A review of theoretical aspects of HLA and disease associations. Theoretical Population Biology 20(2): 168-208. PMID: 7342351

Thomson, G. (1983). Investigation of the mode of inheritance of the HLA associated diseases by the method of antigen genotype frequencies among diseased individuals. Tissue Antigens 21(2)**,** 81-104.

Thomson, G. 1984). HLA DR antigens and susceptibility to insulin-dependent diabetes mellitus. Am. J. Hum. Genet. 36(6)**,** 1309-1317.

Thomson, G. (1993). AGFAP method: applicability under different ascertainment schemes and a parental contributions test. Genet. Epidemiol. 10(5)**,** 289-310.

Thomson, G. (1995a). Analysis of complex human genetic traits: an ordered-notation method and new tests for mode of inheritance. Am. J. Hum. Genet. 57(2)**,** 474-486.

Thomson, G. (1995b). Mapping disease genes: family-based association studies. Am. J. Hum. Genet. 57(2)**,** 487-498.

Thomson, G. (1995c). HLA disease associations: models for the study of complex human genetic disorders. Crit. Rev. Clin. Lab. Sci. 32(2)**,** 183-219.

Thomson G, Bodmer WF. 1977a. The genetics of HLA and disease associations. In: Measuring Selection in Natural Populations, eds. Christiansen FB, Fenchel T, Barndorff-Nielson O. Springer-Verlag, Heidleberg, pp. 545-564.

Thomson G, Bodmer WF. 1977b. The genetic analysis of HLA and disease associations. In: HLA and Disease, eds. Dausset J, Svejgaard A. Munksgaard, Copenhagen, pp. 84-93.

Thomson G, Klitz W. 1987. Disequilibrium pattern analysis. I. Theory. Genetics 116: 623-632. PMID: 3623083

Thomson, G., Robinson, W. P., Kuhner, M. K., Joe, S., MacDonald, M. J., Gottschall, J. L., Barbosa, J., Rich, S. S., Bertrams, J., Baur, M. P., et al. (1988). Genetic heterogeneity, modes of inheritance, and risk estimates for a joint study of Caucasians with insulin-dependent diabetes mellitus. Am. J. Hum. Genet. 43(6)**,** 799-816.

Thomson G, Robinson WP, Kuhner MK, Joe S, Klitz W. 1989. HLA and insulin gene associations with IDDM. Genetic Epidemiology 6: 155-160. PMID: 2567257

Thomson, G., and Valdes, A. M. (2005). Mapping of disease loci. In "Pharmacogenomics" (W. Kalow, U. A. Meyer, and R. F. Tyndale, eds.), pp. 557-587. Taylor and Francis Group, Boca Raton.

Thomson, G., and Valdes, A. M. (2007). Conditional genotype analysis: detecting secondary disease loci in linkage disequilibrium with a primary disease locus. In: Genetic Analysis Workshop 15: Gene Expression Analysis and Approaches to Detecting Multiple Functional Loci, (H.J. Cordell, M. de Andrade, M-C. Babron, et al, eds.) BMC Proceedings 2007, 1(Suppl 1): S163-169.

Thomson G, Valdes AM, Noble JA, et al. 2007a. Relative predispositional effects of HLA Class II DRB1-DQB1 haplotypes and genotypes on type 1 diabetes: a meta-analysis. Tissue Antigens 70: 110- 127. PMID: 17610416

Thomson G, Li H, Dorman SJ, Lie BA, Mignot E, Thorsby E, Steenkiste A, Akey JM, McWeeney S, Single R. 2007b. Statistical approaches for analyses of HLA-associated and other complex diseases. In: Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress, Volume I, ed. Hansen JA. IHWG Press, Seattle, WA, pp. 782-787.

Thomson G, Barcellos LF, Valdes AM. 2008. Searching for additional disease loci in a genomic region. In: Genetic Dissection of Complex Traits, 2nd ed., Advances in Genetics, Vol. 60, ed. Rao DC. Academic Press, pp. 253-292. PMID: 18358324

Thomson G, Marthandan N, Hollenbach JA, Mack SJ, Erlich HA, Single RM, Waller MJ, Marsh SGE, Guidry PA, Karp DR, Scheuermann RH, Thompson SD, Glass DN, Helmberg W. 2010. Sequence Feature Variant Type (SFVT) Analysis of the HLA Genetic Association in Juvenile Idiopathic Arthritis. Pacific Symposium of Biocomputing 2010, 359-370. PMID: 19908388

Thorsby, E. (1997). Invited anniversary review: HLA associated diseases. Hum. Immunol. 53(1)**,** 1-11.

Thorsby E, Caillat-Zucman S, Dorman J, Eliaou JF, Lathrop M, Li H, Lie BA, Mazzilli C, Mignot E, Nelson JL, Pugliese A, Reveille J, Thomson G, Toubert A. 2007. Additional disease predisposing genes in the HLA compex: Summary of the 13th IHWS Disease component studies. In: Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress, Volume I, ed. Hansen JA. IHWG Press, Seattle, WA, pp. 823-827.

Thornton-Wells, T. A., Moore, J. H., and Haines, J. L. (2004). Genetics, statistics and human disease: analytical retooling for complexity. Trends in Genetics 20(12), 640-647.

Tiwari JL, Terasaki PI. (1985) "HLA and Disease Associations." Mew York: Springer-Verlag

Todd, J. A., J. I. Bell, and H. 0. McDevitt. 1987. HLA-DQI gene contributes to susceptibility and resistance to insulin dependent diabetes mellitus. Nature 329: 599-604.

Valdes, A. M., and Thomson, G. (1997) Detecting disease-predisposing variants: the haplotype method. Am. J. Hum. Genet. 60(3)**,** 703-716.

Valdes, A. M., McWeeney, S., and Thomson, G. (1997) HLA class II DR-DQ amino acids and insulin-dependent diabetes mellitus: application of the haplotype method. Am. J. Hum. Genet. 60(3)**,** 717-728.

Valdes, A. M., Thomson, G., Erlich, H. A., and Noble, J. A. (1999) Association between type 1 diabetes age of onset and HLA among sibling pairs. Diabetes 48(8)**,** 1658-1661.

Valdes, A. M., Noble, J. A., Genin, E., Clerget-Darpoux, F., Erlich, H. A., and Thomson, G. (2001). Modeling of HLA class II susceptibility to Type I diabetes reveals an effect associated with DPB1 Genet. Epidemiol. 21(3)**,** 212-223.

Valdes, A. M., Erlich, H. A., and Noble, J. A. (2005a) Human leukocyte antigen class I B and C loci contribute to Type 1 Diabetes (T1D) susceptibility and age at T1D onset. Hum. Immunol. 66(3)**,** 301-313.

Valdes, A. M., Thomson, G., Graham, J., Zarghami, M., McNeney, B., Kockum, I., Smith, A., Lathrop, M., Steenkiste, A. R., Dorman, J. S., et al. (2005b) D6S265*15 marks a DRB1*15, DQB1*0602 haplotype associated with attenuated protection from type 1 diabetes mellitus. Diabetologia 48(12)**,** 2540-2543.

Valdes, A. M., Wapelhorst, B., Concannon, P., Erlich, H. A., Thomson, G., and Noble, J. A. (2005c) Extended DR3-D6S273-HLA-B haplotypes are associated with increased susceptibility to type 1 diabetes in US Caucasians. Tissue Antigens 65(1)**,** 115-119.

Valdes AM, Thomson G, and the T1DGC. 2009. Several loci in the HLA class III region are associated with T1D risk after adjusting for DRB1-DQB1. Diabetes, Obesity and Metabolism 11 (Supplement 1): 46-52. PMID: 19143814

Valdes AM, Thomson G, Barcellos LF, and the T1DGC. 2010. Genetic variation within the HLA Class III influences T1D susceptibility conferred by high-risk HLA haplotypes. Genes and Immunity 11(3): 209-218. PMID: 20054343

van der Horst-Bruinsma, I. E., Hazes, J. M., Schreuder, G. M., Radstake, T. R., Barrera, P., van de Putte, L. B., Mustamu, D., van Schaardenburg, D., Breedveld, F. C., and de Vries, R. (1998) Influence of non-inherited maternal HLA-DR antigens on susceptibility to rheumatoid arthritis. Ann. Rheum. Dis. 57(11)**,** 672-675.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006) A map of recent positive selection in the human genome. PLoS Biol 4(3)**,** e72.

Williams AP, Bateman AR, Khakoo SI (2005) Hanging in the balance. KIR and their role in disease. Molecular interventions 5:226-240

Yeo, T. W., De Jager, P. L., Gregory, S. G., Barcellos, L. F., Walton, A., Goris, A., Fenoglio, C., Ban, M., Taylor, C. J., Goodman, R. S., et al. (2007). A second major histocompatibility complex susceptibility locus for multiple sclerosis. Ann. Neurol. 61(3), 228-236

Zareparsi, S., James, D. M., Kaye, J. A., Bird, T. D., Schellenberg, G. D., and Payami, H. (2002). HLA-A2 homozygosity but not heterozygosity is associated with Alzheimer disease. Neurology 58(6)**,** 973-975.

Zavattari, P., Lampis, R., Motzo, C., Loddo, M., Mulargia, A., Whalen, M., Maioli, M., Angius, E., Todd, J. A., and Cucca, F. (2001). Conditional linkage disequilibrium analysis of a complex disease superlocus, IDDM1 in the HLA region, reveals the presence of independent modifying gene effects influencing the type 1 diabetes risk encoded by the major HLA-DQB1, -DRB1 disease loci. Hum. Mol. Genet. 10(8)**,** 881-889.

Zinkernagel RM, Doherty PC. (1974) Restriction of *in vitro* T-cell mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. Nature, 248, 701–702.

Appendix A: Relative Predispositional Effects (RPE) Method Applied to Juvenile Idiopathic Arthritis (JIA) Oligoarticular Persistent (OP) Data

The data are HLA DRB1 high resolution (4 digit amino acid level) data on 354 JIA patients with the clinical phenotype oligoarticular persistent (OP) and 273 controls (Hollenbach et al. 2010, Thomson et al. 2010). The alleles are ranked by their ORs (from most predisposing to most protective). The first analysis, detailed in this Appendix, is the RPE method of Payami et al. (1989) using a chi-square test of heterogeneity with uncorrected p-values. All alleles with an expected < 5 in patients or controls are placed in the 'binned' category (Table II.H.1). The binned category is then removed from analysis (keeping the binned category in the analyses does not alter the results, and whether or not to remove the binned category is at the discretion of the analyst). Alleles with the strongest effects (measured by their contribution to the overall chi-square value) are then sequentially removed from the analysis. Note that the *individual* allele p-values are biased (conservative with respect to finding significant effects), as the assumption of a chi-square with 1 degree of freedom (df) is incorrect; the p-values can be used however for a relative ranking of the allelic effects.

The overall analysis of the complete data set shows considerable heterogeneity in risk (Chi-sq = 182.1, df = 21, p < 1.1E-27) (Table A.1). Removal of the binned category gives a very similar result (Chi-sq = 181.5, df = 20, p < 5.0E-28) (Table A.2), and in both cases DRB1*0801 (predisposing) and DRB1*1501 and DRB1*0701 (protective) are the strongest effects (these alleles are denoted as belonging to Category 1 in the RPE analysis, and see Column 1 of Table II.H.1). Removal of these alleles gives a highly significant result for risk heterogeneity of the remaining alleles (Chi-sq = 79.9, df = 18, p < 4.1E-10) (Table A.3), with DRB1*1104 (predisposing) and DRB1*0401 (protective) as the strongest effects (Category 2). Note that a strong argument could be made for deleting these alleles also at the previous round, but it does not alter the outcome. With removal of these strong effects, there is only minimal evidence of remaining risk heterogeneity (Chi-sq = 25.2, df = 13, p < 0.02) (Table A.4), with the strongest effects due to DRB1*1103 (p < 0.01) (predisposing) and DRB1*0103 (p < 0.02) (protective) (Category 3). Note that these latter p-values would not be significant with corrections for multiple comparisons of either the overall tests, or those for individual alleles. Notwithstanding, our principal aim is to detect heterogeneity that may be relevant to detecting additional disease genes in a genetic region, hence one wants to err on the side of not missing potential heterogeneity, and these observations should be studied with replication data sets. The results from RPE analysis are all compatible with heterogeneity testing of all pairwise allele comparisons (Appendix B below), and consideration of ORs and P/C ratios (which are similar in these cases).

Table A.1: Heterogeneity test of HLA DRB1 alleles and JIA-OP including the binned class

DRB1	Patients	controls	Total	Exp P	Exp C	Chi sq	P-value	df
1103-	12	1	*13*	*7.3397*	*5.6603*	6.80	0.009136	21
0801-	102	13	*115*	*64.928*	*50.072*	48.61	3.12E-12	**Ov'l Chi Sq**
1104-	57	11	*68*	*38.392*	*29.608*	20.71	5.34E-06	**182.1**
0403-	9	3	*12*	*6.7751*	*5.2249*	1.68	0.195186	**P-value**
1301-	90	38	*128*	*72.268*	*55.732*	9.99	0.001572	**1.1E-27**

0102-	9	5	*14*	*7.9043*	*6.0957*	0.35	0.554774
1101-	60	36	*96*	*54.201*	*41.799*	1.42	0.232585
0901-	9	6	*15*	*8.4689*	*6.5311*	0.08	0.782105
0101-	74	50	*124*	*70.01*	*53.99*	0.52	0.469828
0301-	89	61	*150*	*84.689*	*65.311*	0.50	0.477746
1201-	10	8	*18*	*10.163*	*7.8373*	0.01	0.938356
1302-	28	23	*51*	*28.794*	*22.206*	0.05	0.822511
1303-	10	9	*19*	*10.727*	*8.2727*	0.11	0.736482
binned	27	27	*54*	*30.488*	*23.512*	0.92	0.338393
1601-	6	8	*14*	*7.9043*	*6.0957*	1.05	0.304658
1401-	11	18	*29*	*16.373*	*12.627*	4.05	0.044175
1502-	5	10	*15*	*8.4689*	*6.5311*	3.26	0.070845
1501-	38	80	*118*	*66.622*	*51.378*	28.24	1.07E-07
0701-	30	65	*95*	*53.636*	*41.364*	23.92	1E-06
0401-	21	47	*68*	*38.392*	*29.608*	18.10	2.1E-05
0404-	7	16	*23*	*12.986*	*10.014*	6.34	0.011826
0103-	4	11	*15*	*8.4689*	*6.5311*	5.42	0.019953
total	708	546	*1254*	*708*	*546*	0.00	1

Table A.2: Binned category removed

DRB1	Patients	controls	Total	Exp P	Exp C	Chi sq	P-value	df
1103-	12	1	*13*	*7.3775*	*5.6225*	6.70	0.009659	20
0801-	102	13	*115*	*65.263*	*49.738*	47.82	4.68E-12	Ov'l Chi Sq
1104-	57	11	*68*	*38.59*	*29.41*	20.31	6.6E-06	**181.5**
0403-	9	3	*12*	*6.81*	*5.19*	1.63	0.201928	P-value
1301-	90	38	*128*	*72.64*	*55.36*	9.59	0.001954	**5.0E-28**
0102-	9	5	*14*	*7.945*	*6.055*	0.32	0.569267	
1101-	60	36	*96*	*54.48*	*41.52*	1.29	0.255465	
0901-	9	6	*15*	*8.5125*	*6.4875*	0.06	0.799442	
0101-	74	50	*124*	*70.37*	*53.63*	0.43	0.510544	
0301-	89	61	*150*	*85.125*	*64.875*	0.41	0.523063	
1201-	10	8	*18*	*10.215*	*7.785*	0.01	0.918528	
1302-	28	23	*51*	*28.943*	*22.058*	0.07	0.789938	
1303-	10	9	*19*	*10.783*	*8.2175*	0.13	0.717088	
1601-	6	8	*14*	*7.945*	*6.055*	1.10	0.294063	
1401-	11	18	*29*	*16.458*	*12.543*	4.18	0.040797	
1502-	5	10	*15*	*8.5125*	*6.4875*	3.35	0.067159	
1501-	38	80	*118*	*66.965*	*51.035*	28.97	7.36E-08	
0701-	30	65	*95*	*53.913*	*41.088*	24.52	7.34E-07	
0401-	21	47	*68*	*38.59*	*29.41*	18.54	1.67E-05	
0404-	7	16	*23*	*13.053*	*9.9475*	6.49	0.010853	
0103-	4	11	*15*	*8.5125*	*6.4875*	5.53	0.018684	
total	681	519	*1200*	*681*	*519*	0.00	1	

Category 1: DRB1*0801 – the strongest predisposing effect is removed, and DRB1*1501 and DRB1*0701 - the two strongest protective effects, are removed for the next round of analysis

Table A.3: DRB1*0801, DRB1*1501, and DRB1*0701 (Category 1) removed

DRB1	Patients	Controls	Total	Exp P	Exp C	Chi sq	P-value	df
1103-	12	1	*13*	*7.6181*	*5.3819*	6.09	0.013609	18
1104-	57	11	*68*	*39.849*	*28.151*	17.83	2.41E-05	Ov'l Chi Sq
0403-	9	3	*12*	*7.0321*	*4.9679*	1.33	0.248765	**79.9**
1301-	90	38	*128*	*75.009*	*52.991*	7.24	0.007142	P-value
0102-	9	5	*14*	*8.2041*	*5.7959*	0.19	0.665852	**4.1E-10**
1101-	60	36	*96*	*56.257*	*39.743*	0.60	0.437972	
0901-	9	6	*15*	*8.7901*	*6.2099*	0.01	0.9124	
0101-	74	50	*124*	*72.665*	*51.335*	0.06	0.807714	
0301-	89	61	*150*	*87.901*	*62.099*	0.03	0.855489	
1201-	10	8	*18*	*10.548*	*7.4518*	0.07	0.793077	
1302-	28	23	*51*	*29.886*	*21.114*	0.29	0.591744	
1303-	10	9	*19*	*11.134*	*7.8658*	0.28	0.597312	
1601-	6	8	*14*	*8.2041*	*5.7959*	1.43	0.231703	
1401-	11	18	*29*	*16.994*	*12.006*	5.11	0.023827	
1502-	5	10	*15*	*8.7901*	*6.2099*	3.95	0.046941	
0401-	21	47	*68*	*39.849*	*28.151*	21.54	3.47E-06	
0404-	7	16	*23*	*13.478*	*9.5218*	7.52	0.006098	
0103-	4	11	*15*	*8.7901*	*6.2099*	6.31	0.012037	
total	511	361	*872*	*511*	*361*	0.00	1	

Category 2: DRB1*1104 – the next strongest predisposing effect is removed, and DRB1*0401 - the next strongest protective effect, are removed for the next round of analysis

Table A.4: Category 1 alleles and DRB1*1104 and DRB1*0401 (Category 2) removed

DRB1	Patients	controls	Total	Exp P	Exp C	Chi sq	P-value	df
1103-	12	1	*13*	*7.4667*	*5.5333*	6.47	0.010993	13
0403-	9	3	*12*	*6.8923*	*5.1077*	1.51	0.218487	Ov'l Chi Sq
0102-	9	5	*14*	*8.041*	*5.959*	0.27	0.604209	**25.2**
1101-	60	36	*96*	*55.138*	*40.862*	1.01	0.315612	P-value
0901-	9	6	*15*	*8.6154*	*6.3846*	0.04	0.840817	**0.02**
0101-	74	50	*124*	*71.221*	*52.779*	0.25	0.613682	
0301-	89	61	*150*	*86.154*	*63.846*	0.22	0.638354	
1201-	10	8	*18*	*10.338*	*7.6615*	0.03	0.87182	
1302-	28	23	*51*	*29.292*	*21.708*	0.13	0.714374	
1303-	10	9	*19*	*10.913*	*8.0872*	0.18	0.671901	
1601-	6	8	*14*	*8.041*	*5.959*	1.22	0.269922	

1401-	11	18	*29*	*16.656*	*12.344*	4.51	0.03364
1502-	5	10	*15*	*8.6154*	*6.3846*	3.56	0.05903
0103-	4	11	*15*	*8.6154*	*6.3846*	5.81	0.015945
total	336	249	*585*	*336*	*249*	0.00	1

Category 3: weak evidence for heterogeneity of effects for DRB1*1103 predisposing, and DRB1*0103 protective effect, no further rounds of analyses are performed, although note that these are very rare alleles and significant effects may be detected in a larger sample

Appendix B: Pairwise Risk Comparisons Applied to Juvenile Idiopathic Arthritis (JIA) Oligoarticular Persistent (OP) Data

The data are as in Appendix A. The alleles are ranked by their ORs (from most predisposing to most protective) (see Table II.H.1). The Tables given below list the uncorrected p-values for a chi-square test of heterogeneity of each pairwise combination listed. The diagonals are shaded grey, and the values are symmetric around the diagonal.

The p-values for the pairwise tests for alleles that are *relatively common* in patients or controls are given in Table B.1 below (this is identical to Table II.H.2 in the text, and is repeated for ease of comparison with the other Tables in this Appendix). These show a pattern of nearly mutually exclusive blocks of predisposing (shaded teal), neutral (intermediate) (shaded light grey), and protective (shaded pink) alleles (these are referred to as categories I, II, and III in column 2 of Table II.H.1 in the text).

Table B.1: Pairwise allele risk heterogeneity comparison p-values for common HLA DRB1 alleles and JIA-OP

Reduced data set A - categories I, II, and III

DRB1	*0801	*1104	*1301	*1101	*0101	*0301	*1302	*0404	*1501	*0701	*0401
*0801		0.3454	0.0004	7E-06	4E-07	1E-07	1E-06	4E-10	1E-18	2E-17	8E-16
*1104	0.3454		0.0376	0.0029	0.0006	0.0004	0.0005	1E-06	1E-11	4E-11	4E-10
*1301	0.0004	0.0376		0.2186	0.0766	0.0567	0.0496	0.0002	2E-09	1E-08	1E-07
*1101	7E-06	0.0029	0.2186		0.6705	0.6201	0.371	0.0054	1E-05	2E-05	7E-05
*0101	4E-07	0.0006	0.0766	0.6705		0.9539	0.5604	0.0096	2E-05	4E-05	0.0001
*0301	1E-07	0.0004	0.0567	0.6201	0.9539		0.5794	0.0094	1E-05	2E-05	1E-04
*1302	1E-06	0.0005	0.0496	0.371	0.5604	0.5794		0.051	0.0055	0.006	0.0084
*0404	4E-10	1E-06	0.0002	0.0054	0.0096	0.0094	0.051		0.8678	0.9155	0.9679
*1501	1E-18	1E-11	2E-09	1E-05	2E-05	1E-05	0.0055	0.8678		0.9226	0.8521
*0701	2E-17	4E-11	1E-08	2E-05	4E-05	2E-05	0.006	0.9155	0.9226		0.9246
*0401	8E-16	4E-10	1E-07	7E-05	0.0001	1E-04	0.0084	0.9679	0.8521	0.9246	

Table B.2 below extends the alleles considered, such that rarer alleles within the bounds of the three categories I, II and III above are now included. The pattern now deviates more from the block like patterns of significance expected based on their ORs, and this reflects in the main the fact that the additional alleles are much rarer. Nonetheless their categorization into the three risk categories is substantiated by much of the data (these are referred to as categories Ix, IIx, and IIIx, see Table II.H.1 of the text).

Table B.2: Pairwise allele risk heterogeneity comparison p-values for an extended set of DRB1 alleles

Reduced data set B - categories Ix, IIx, and IIIx

DRB1	*1103	*0801	*1104	*1301	*0102	*1101	*0901	*0101	*0301	*1201	*1302
*1103		0.6925	0.4301	0.0912	0.0801	0.0332	0.049	0.0206	0.0188	0.0261	0.0129
*0801	0.6925		0.3454	0.0004	0.0128	7E-06	0.0031	4E-07	1E-07	0.0003	1E-06
*1104	0.4301	0.3454		0.0376	0.093	0.0029	0.0385	0.0006	0.0004	0.0102	0.0005
*1301	0.0912	0.0004	0.0376		0.6412	0.2186	0.4129	0.0766	0.0567	0.207	0.0496
*0102	0.0801	0.0128	0.093	0.6412		0.8973	0.8121	0.7385	0.7178	0.6179	0.53
*1101	0.0332	7E-06	0.0029	0.2186	0.8973		0.8527	0.6705	0.6201	0.5786	0.371
*0901	0.049	0.0031	0.0385	0.4129	0.8121	0.8527		0.9808	0.96	0.797	0.7266
*0101	0.0206	4E-07	0.0006	0.0766	0.7385	0.6705	0.9808		0.9539	0.7395	0.5604
*0301	0.0188	1E-07	0.0004	0.0567	0.7178	0.6201	0.96	0.9539		0.7582	0.5794
*1201	0.0261	0.0003	0.0102	0.207	0.6179	0.5786	0.797	0.7395	0.7582		0.9618
*1302	0.0129	1E-06	0.0005	0.0496	0.53	0.371	0.7266	0.5604	0.5794	0.9618	
*0404	0.0004	4E-10	1E-06	0.0002	0.0438	0.0054	0.0712	0.0096	0.0094	0.1052	0.051
*1501	2E-05	1E-18	1E-11	2E-09	0.0178	1E-05	0.0339	2E-05	1E-05	0.0535	0.0055
*0701	3E-05	2E-17	4E-11	1E-08	0.0172	2E-05	0.0325	4E-05	2E-05	0.0511	0.006
*0401	4E-05	8E-16	4E-10	1E-07	0.0181	7E-05	0.0336	0.0001	1E-04	0.0525	0.0084

DRB1	*0404	*1501	*0701	*0401
*1103	0.0004	2E-05	3E-05	4E-05
*0801	4E-10	1E-18	2E-17	8E-16
*1104	1E-06	1E-11	4E-11	4E-10
*1301	0.0002	2E-09	1E-08	1E-07
*0102	0.0438	0.0178	0.0172	0.0181
*1101	0.0054	1E-05	2E-05	7E-05
*0901	0.0712	0.0339	0.0325	0.0336
*0101	0.0096	2E-05	4E-05	0.0001
*0301	0.0094	1E-05	2E-05	1E-04
*1201	0.1052	0.0535	0.0511	0.0525
*1302	0.051	0.0055	0.006	0.0084
*0404		0.8678	0.9155	0.9679
*1501	0.8678		0.9226	0.8521
*0701	0.9155	0.9226		0.9246
*0401	0.9679	0.8521	0.9246	

Table B.3 below considers all alleles, including the binned category. Rare alleles which cannot be placed into one of the three risk categories are shaded in green.

Table B.3: Pairwise allele risk heterogeneity comparison p-values for all DRB1 alleles

Full data set

DRB1	*1103	*0801	*1104	*0403	*1301	*0102	*1101	*0901	*0101	*0301	*1201
*1103		0.6925	0.4301	0.2383	0.0912	0.0801	0.0332	0.049	0.0206	0.0188	0.026
*0801	0.6925		0.3454	0.1737	0.0004	0.0128	7E-06	0.0031	4E-07	1E-07	0.000
*1104	0.4301	0.3454		0.4583	0.0376	0.093	0.0029	0.0385	0.0006	0.0004	0.010
*0403	0.2383	0.1737	0.4583		0.733	0.5551	0.3954	0.4113	0.2987	0.2854	0.278
*1301	0.0912	0.0004	0.0376	0.733		0.6412	0.2186	0.4129	0.0766	0.0567	0.207
*0102	0.0801	0.0128	0.093	0.5551	0.6412		0.8973	0.8121	0.7385	0.7178	0.617
*1101	0.0332	7E-06	0.0029	0.3954	0.2186	0.8973		0.8527	0.6705	0.6201	0.578
*0901	0.049	0.0031	0.0385	0.4113	0.4129	0.8121	0.8527		0.9808	0.96	0.797
*0101	0.0206	4E-07	0.0006	0.2987	0.0766	0.7385	0.6705	0.9808		0.9539	0.739
*0301	0.0188	1E-07	0.0004	0.2854	0.0567	0.7178	0.6201	0.96	0.9539		0.758
*1201	0.0261	0.0003	0.0102	0.2789	0.207	0.6179	0.5786	0.797	0.7395	0.7582	
*1302	0.0129	1E-06	0.0005	0.2032	0.0496	0.53	0.371	0.7266	0.5604	0.5794	0.961
*1303	0.0174	8E-05	0.0043	0.213	0.1231	0.5032	0.4207	0.6675	0.5613	0.5764	0.858
binned	0.0055	3E-08	6E-05	0.1157	0.009	0.3399	0.1365	0.4928	0.2309	0.235	0.683
*1601	0.0065	1E-05	0.0009	0.0982	0.0371	0.2556	0.1611	0.3559	0.2268	0.2324	0.476
*1401	0.0011	3E-09	6E-06	0.0307	0.001	0.1045	0.0192	0.1634	0.0339	0.0336	0.237
*1502	0.0014	1E-07	5E-05	0.0313	0.0041	0.0955	0.033	0.1432	0.0517	0.0525	0.201
*0404	0.0004	4E-10	1E-06	0.012	0.0002	0.0438	0.0054	0.0712	0.0096	0.0094	0.105
*1501	2E-05	1E-18	1E-11	0.0033	2E-09	0.0178	1E-05	0.0339	2E-05	1E-05	0.053
*0701	3E-05	2E-17	4E-11	0.0032	1E-08	0.0172	2E-05	0.0325	4E-05	2E-05	0.051
*0401	4E-05	8E-16	4E-10	0.0036	1E-07	0.0181	7E-05	0.0336	0.0001	1E-04	0.052
*0103	0.0005	6E-09	6E-06	0.0125	0.0008	0.0418	0.009	0.0654	0.015	0.015	0.094

DRB1	*1302	*1303	binned	*1601	*1401	*1502	*0404	*1501	*0701	*0401	*010
*1103	0.0129	0.0174	0.0055	0.0065	0.0011	0.0014	0.0004	2E-05	3E-05	4E-05	0.00
*0801	1E-06	8E-05	3E-08	1E-05	3E-09	1E-07	4E-10	1E-18	2E-17	8E-16	6E-0
*1104	0.0005	0.0043	6E-05	0.0009	6E-06	5E-05	1E-06	1E-11	4E-11	4E-10	6E-0
*0403	0.2032	0.213	0.1157	0.0982	0.0307	0.0313	0.012	0.0033	0.0032	0.0036	0.01
*1301	0.0496	0.1231	0.009	0.0371	0.001	0.0041	0.0002	2E-09	1E-08	1E-07	0.00
*0102	0.53	0.5032	0.3399	0.2556	0.1045	0.0955	0.0438	0.0178	0.0172	0.0181	0.04
*1101	0.371	0.4207	0.1365	0.1611	0.0192	0.033	0.0054	1E-05	2E-05	7E-05	0.00
*0901	0.7266	0.6675	0.4928	0.3559	0.1634	0.1432	0.0712	0.0339	0.0325	0.0336	0.06
*0101	0.5604	0.5613	0.2309	0.2268	0.0339	0.0517	0.0096	2E-05	4E-05	0.0001	0.01
*0301	0.5794	0.5764	0.235	0.2324	0.0336	0.0525	0.0094	1E-05	2E-05	1E-04	0.01
*1201	0.9618	0.8584	0.683	0.476	0.2374	0.2018	0.1052	0.0535	0.0511	0.0525	0.09
*1302		0.8654	0.6152	0.4241	0.1443	0.1419	0.051	0.0055	0.006	0.0084	0.05

***1303**	0.8654		0.8436	0.5787	0.3154	0.2605	0.1447	0.0832	0.0792	0.0801	0.126
binned	0.6152	0.8436		0.6337	0.2927	0.2522	0.1136	0.0255	0.0261	0.0318	0.108
***1601**	0.4241	0.5787	0.6337		0.7569	0.5974	0.4427	0.424	0.4022	0.3853	0.359
***1401**	0.1443	0.3154	0.2927	0.7569		0.7638	0.5725	0.5577	0.5245	0.4991	0.454
***1502**	0.1419	0.2605	0.2522	0.5974	0.7638		0.851	0.9298	0.8922	0.853	0.690
***0404**	0.051	0.1447	0.1136	0.4427	0.5725	0.851		0.8678	0.9155	0.9679	0.802
***1501**	0.0055	0.0832	0.0255	0.424	0.5577	0.9298	0.8678		0.9226	0.8521	0.663
***0701**	0.006	0.0792	0.0261	0.4022	0.5245	0.8922	0.9155	0.9226		0.9246	0.702
***0401**	0.0084	0.0801	0.0318	0.3853	0.4991	0.853	0.9679	0.8521	0.9246		0.747
***0103**	0.0544	0.1266	0.108	0.3593	0.4549	0.6903	0.8023	0.6639	0.702	0.7474	