Haplotype Estimation and Linkage Disequilibrium Methods Manual:
Version 0.1.8 (June 3, 2011)

Estimating Haplotype Frequencies and Linkage Disequilibrium Parameters in the HLA and KIR Regions

Available online at: www.ImmPort.org

Richard M. Single[1], Pierre-Antoine Gourraud[2], Alex K. Lancaster[3,4], Farren Briggs[5], Lisa F. Barcellos[5], Jill A. Hollenbach[6], Steven J. Mack[6], Glenys Thomson[3]

[1] Department of Mathematics and Statistics, 306 Mansfield House, University of Vermont, Burlington, VT 05405, USA, richard.single@uvm.edu

[2] Department of Neurology, 513 Parnassus Avenue, University of California, San Francisco, CA 94143-0436, USA, e-mail: pierreantoine.gourraud@ucsf.edu

[3] Department of Integrative Biology, 3060 Valley Life Sciences Building MC #3140, University of California, Berkeley, CA 94720-3140, USA, e-mails: alexl@cal.berkeley.edu, glenys@berkeley.edu

[4] Mailing address: Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA

[5] Division of Epidemiology, School of Public Health, University of California, Berkeley, CA 94720-, USA, e-mails: fbsbriggs@gmail.com, barcello@genepi.berkeley.edu

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

Acknowledgments

This publication was supported in part by NIH/NIAID contract number HHSN266200400076C, ABD N01-AI-40076 (RMS and GT). This research utilizes resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD), and Juvenile Diabetes Research Foundation International (JDRF) and supported by U01 DK062418. The contents of this Methods Manual are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Outline

Abbreviations

BMDRs—Bone Marrow Donor Registries;

CDV—constrained disequilibrium values;

D_{ij}—pairwise LD statistic between alleles Ai and Bj at two loci, also written as D_{AiBj};

D'_{ij} ($=D_{ij}$ /Dmax)—the standard normalized linkage disequilibrium between alleles A_i and B_j at two loci;

DPA—disequilibrium pattern analysis;

EM—expectation-maximization;

F—the expected proportion of homozygotes under HWP: $F_A = \Sigma_i\ p_{Ai}^2$;

FND—the normalized deviate of the homozygosity F statistic;

H (= 1—F)— the expected proportion of heterozygotes under HWP, also referred to as the gene diversity index;

HF—haplotype frequency;

HLA—human leukocyte antigen;

HSF—haplotype specific homozygosity;

HSH—haplotype specific heterozygosity;

HWP—Hardy-Weinberg proportions;

IHW—International Histocompatibility Workshop;

k_A—the observed number of genetic variants, e.g., alleles or amino acids at locus A;

KIR—killer inhibitory receptors;

LD—linkage disequilibrium;

LE—linkage equilibrium;

n—# of individuals in a sample;

NK—Natural Killer

PyPop— www.pypop.org, www.ImmPort.org (Python for Population Genomics – PyPop, current release version 0.7.0) (Lancaster et al. 2003, 2007a, 2007b, 2008; Lancaster 2006);

r—the correlation coefficient between the allele frequency distributions at two bi-allelic loci denoted A and B, with $r^2 = D^2/[p_{A1}p_{A2}p_{B1}p_{B2}]$;

SIRE—self identified race/ethnicity

W_n— also denoted W_{AB}, the multi-allelic extension of the bi-allelic correlation coefficient r of LD between two loci denoted A and B: W_n = r for bi-allelic loci

I. Overview

A. Introduction

Note that references in Section I are kept to a minimum. Detailed references are listed in later sections.

There are many challenges in trying to identify appropriate techniques for analysis of the extensive data generated by whole genome association and/or linkage studies as well as detailed study of specific genetic regions such as HLA and KIR. Even when computational power is sufficient to consider all possible combinations of genes/alleles/haplotypes, the combinatorial magnitude makes interpretation of the results problematic. Some single nucleotide polymorphisms (SNPs) or microsatellites (MSATs) that do not show a single locus association with disease (marginal effect), may nevertheless be directly involved in disease predisposition, *and* this may only be detected via appropriate haplotype or stratification analyses that take account of the linkage disequilibrium (LD) structure of the data. Conditional analyses of haplotype data are important for identifying primary disease predisposing genes, as well as additional (secondary) genes that are also involved in disease, but whose effects are weaker and may be restricted to specific allele/haplotype/genotype subsets at the primary gene.

With HLA disease associations, the high level of LD between many of the classical HLA genes means that multiple disease associations may be observed, some of which may indicate a causative genetic factor and others may be due to LD with this causative factor. This is illustrated by the original associations of the serologically defined A3, B7, and DR2 alleles with multiple sclerosis in samples of European origin, and using molecular results 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) as the primary association in Europeans. LD is in many cases nearly complete between the tightly linked class II DRB1, DQA1, and DQB1 loci. For example, almost all DR3 haplotypes in many ethnic groups are DRB1*03:01-DQA1*05:01-DQB1*02:01. The class I B and C loci also exhibit very high LD, as do DPA1-DPB1, and the DRB3/4/5 loci and DRB1, while less closely linked HLA loci show more moderate, but still quite strong, levels of LD (e.g., class I A and B). Structural variation and recombination hotspots within the HLA region have been identified, particularly between the DP and DR-DQ regions, and explain the lower LD generally seen between the DP loci and the other classical HLA loci.

Reports of disease associations for MSATs and SNPs in the HLA region have appeared in the literature for a number of diseases for which classical HLA genes have been identified as a primary disease risk factor. In many of these studies it has been difficult to determine if an additional HLA region gene is involved in disease, versus the associations reported reflecting LD with the antigen presenting HLA molecules directly involved in disease. 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 (reviewed in Thomson et al. 2008, also see Thomson et al. 2007a, and references therein).

The frequency of haplotypes and the strength of LD among loci are also informative with respect to evolutionary forces acting on genes, and the history of human populations. An understanding of LD patterns is also useful for detecting evidence of selection. Distinguishing among demographic and selective explanations for patterns of variation observed with HLA genes is a challenge (Meyer and Thomson 2001, Meyer et al. 2006, 2007, Single et al. 2007a). Allele frequency distributions for the HLA loci, with the exception of DPB1, deviate from neutral expectations in the direction of balancing selection (for a meta-analysis and review of previous studies see Solberg et al. 2008), and this is unlikely to be explained by demographic factors. The fact that DPB1 allele distributions do not deviate from neutral expectations does not mean selection is not acting; it is detected at the amino acid level for DPB1 as well as for the other classical HLA genes (Salamon et al. 1999, Valdes et al. 1999, Cano and Fernández-Viña 2009). Other features of HLA variation are explained in part by demographic history, including *decreased* heterozygosity and *increased* LD for populations at greater distances from Africa. Examining all locus pairs of the classical HLA genes, LD is seen to be lowest in sub-Saharan African populations, and highest in the native populations of North and South America (Single et al. 2007a,). HLA allele and haplotype frequencies (HFs) vary even within Caucasians, for example from Northern to Southern Europe (Single et al. 2007a, Meyer et al. 2007), and this must be taken into account in disease studies, and also with analyses of SNP and MSAT data. Combinations of these clines in frequency and LD create a complex setting for analyses involving individuals of mixed ancestry (e.g., African-, European-, Hispanic-, and Asian- Americans, see Maiers et al 2007). The percentage of ancestry from heterogeneous regions must be taken into account in disease studies in order to avoid spurious association signals.

Understanding and incorporating the LD structure of a genetic region into analyses is crucial for detecting disease predisposing variants, as well as for understanding the evolutionary history of a genetic region. Our knowledge in this area, both within and between populations, is continuing to evolve with greater availability of molecular-level HLA data as well as typing of many additional marker loci (MSATS and SNPs).

B. Linkage disequilibrium (LD)

As stated by Slatkin (2008) "Linkage disequilibrium is one of those unfortunate terms that does not reveal its meaning." Non-random associations found between alleles at different loci are generally referred to as linkage disequilibrium (LD) although they may not be due to linkage. Under some selection models, a population at *equilibrium* can maintain LD between loci, and although rare, *unlinked* markers may show significant LD; also very closely linked loci may be in linkage equilibrium. Gametic disequilibrium is a more descriptive term, but is not as commonly used.

The LD parameter D_{ij} (also written as D_{AiBj}) between a pair of alleles $\mathrm{A_i}$ and $\mathrm{B_j}$ at two loci is defined as the difference between the observed haplotype (gametic) frequency at the population level and that expected under random association of the two alleles, i.e., $D_{\mathrm{ij}} = \mathrm{f}(\mathrm{A_iB_j}) - p_{\mathrm{Ai}}\, p_{\mathrm{Bj}}$. However, the maximum value D_{ij} can take is a function of the observed allele frequencies and defining the *strength* of any observed non-random association is complicated by this fact. A number of normalized measures to reflect the strength of LD have been proposed; both for bi- and multi-allelic data (see Hedrick 1987 for review). No single summary statistic measuring LD strength captures all aspects of LD, which is multi-dimensional in nature. Thus, each measure has different properties and hence different strengths and weaknesses with respect to the question being addressed (Lewontin 1988). Existing measures are not always well suited for direct use with all data.

LD can be created by various evolutionary factors: selection either directly on the two loci or indirectly via a hitchhiking event, migration and admixture, inbreeding and genetic drift. The most likely cause though is historical—when a new mutation arises there is a non-random association created with respect to variation at other polymorphic loci—this association is broken down by recombination, but remains for a very long time between closely linked loci.

The observed levels of overall LD for HLA data have been shown to be incompatible with neutrality expectations (Hedrick and Thomson 1986, Klitz and Thomson 1987, Klitz et al. 1992). Two methods to specifically detect selection via LD patterns of HLA genes have been developed and have identified specific HLA haplotypes that show signs of past selection in specific populations: disequilibrium pattern analysis (DPA) (Thomson and Klitz 1987, Klitz and Thomson 1987, Williams et al. 2004) and constrained disequilibrium values (CDV) (Robinson et al. 1991a, b, Grote et al. 1998). There is agreement between the results of application of these two methods, however, there are many instances where selection will not be detected via these methods.

The two most common measures of the strength of LD for *bi-allelic* data are: (1) the normalized measure of the LD, namely $D' = D/D_{max}$ (for bi-allelic loci there is only one LD parameter, denoted D, since $D_{A1B1} = -D_{A1B2} = -D_{A2B1} = D_{A2B2}$, and hence $D_{\mathrm{AiBj}} = D$, and D_{max} is the maximum value D can take given the allele frequencies and the sign of D); and (2) the correlation coefficient r, which is most often reported as $r^2 = D^2/[p_{A1}p_{A2}p_{B1}p_{B2}]$. $D' = 1$ whenever one (or two) of the possible four haplotypes is not observed, irrespective of its expected frequency. In contrast, $r = 1$ only when the two loci show 100% correlation, i.e., when both loci have (a) equal allele frequencies, and (b) only two complementary haplotypes are observed: either A_1B_1 and A_2B_2, or A_1B_2 and A_2B_1. This correlation property is of particular interest to many research issues, e.g., if $r = 1$, or is very close to 1, then there is no, or probably insufficient, variation that can be analyzed by any stratification method to distinguish between two potentially disease predisposing genetic variants. Similarly, in population genetic or evolutionary studies, the two loci (genes or amino acids thereof) would show very similar allele frequency distributions and it would be difficult to disentangle evolutionary forces acting on them. For multi-allelic markers appropriate extensions of these two bi-allelic measures are used. These are defined in section II.B.

To distinguish between the effects of two markers which are highly correlated there must be some breakdown in the LD pattern so that stratification analyses can be applied. If the markers are truly 100% correlated, or so strongly correlated that their effects cannot be disentangled, then of course their individual effects cannot be distinguished, e.g., DRB1*15:01 and DQB1*06:02 and multiple sclerosis and narcolepsy in Caucasians. However, note that markers with D' = 1, can nevertheless often allow a heterogeneity test. For example, with two locus HFs of A_1B_1 = 0.3, A_1B_2 = 0.2, A_2B_2 = 0.5, although D' = 1, r = 0.65 and one can possibly distinguish between the effects of A_1 and A_2 on B_2 haplotypes, although not on A_1 haplotypes, and similarly for the vice versa situation. For this reason we prefer the use of the correlation coefficient r for bi-allelic loci, and its extension for multi-allelic loci. With complete correlation of sites, e.g., when A_1B_1 and A_2B_2 are the only two haplotypes seen, or haplotypes which break up this association are too rare for use in stratification analyses, then the individual effects of the two loci cannot be disentangled (as mentioned in the Introduction above). Also to remember is that markers in different haplotype block structures can still show high LD with each other. Stratification analyses should *not be* restricted to within block analyses.

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 more details will be given in a later version of this Methods Manual.

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.

C. Estimating HLA and KIR haplotype frequencies (HFs)

Many software programs can handle large numbers of bi-allelic markers e.g., haploview (http://www.broad.mit.edu/mpg/haploview/index.php). However, most of these do not work with the high degree of polymorphism found in for example the classical HLA genes or MSAT loci. For example, with HLA data in Caucasians it is not uncommon to see 40 to 50 alleles at with DRB1, 15 or more alleles at DQB1 , and 50-80 DRB1-DQB1 haplotypes.. For more details on numbers of alleles seen in specific populations and regions world wide see Solberg et al. (2008). Other programs only allow a limited number of loci to be evaluated, not the hundreds to thousands that are the focus of current projects. These limitations apply to estimates of HFs for both family based and case/control data. Also, issues of missing data vary between programs, as well as how haplotypes are estimated with family data (see Niu 2004 and Salem et al. 2005 for reviews). Various mathematical frameworks have been proposed and implemented to address haplotype estimation: Parsimonious methods such as Clark's algorithm, Maximum likelihood methods using the expectation maximization (EM) algorithm, Pseudo-Bayesian and Bayesian methods (Niu et al. 2002, Niu 2004).

Maximum likelihood based algorithms to specifically handle the high level of polymorphism of the HLA loci (or MSATs) have been developed (these are applicable to less polymorphic loci as well): the haplotype and LD estimation module in PyPop (www.pypop.org, www.ImmPort.org) was developed specifically for analysis of the 13th International Histocompatibility Workshop (IHW) anthropology/ human diversity and disease association data (Lancaster et al. 2003, 2007a, b, 2008, Lancaster 2006, Single et al. 2007a, b, Meyer et al. 2007). It has also been applied to data analyses for the 14th and 15th IHWs (Single et al. 2007d) as well as other population studies. Precompiled versions using the default settings are limited to a total of 8 loci at a time and 5,000 individuals; however, these values can be increased by modifying the source code, which is freely available.

The haplotype frequency and LD estimates of Bone Marrow Donor Registries (BMDRs) (e.g., NMDP in the US, Anthony Nolan Research Institute in the UK, EUROMADO in Europe, Registre France Greffe de Moelle in France, ZKRD in Germany, and JMDP in Japan) include multi-locus genotypes for millions of individuals. To accomplish these analyses, registries have customized the PyPop software to run on their high-memory servers. Algorithms to handle more loci in the HLA region, as well as large sample sizes, have been developed e.g., the Estihaplo algorithm of Gourraud et al. (2007) (http://birl.supbiotech.fr/hla-estihaplo.html) (also see Salem et al. 2005). These methods, which are unconstrained by sample size and number of alleles, are only limited by the number of haplotype parameters they will exhaustively enumerate (hundreds of thousands of haplotypes can be formed by the classical HLA genes).

HLA data present additional issues since typing techniques may not completely identify all possible HLA alleles. Consequently haplotype estimation models and software have to adapt to an additional level of missing information (in addition to missing phase information) which is the ambiguous nature of the typing. This issue has been part of working with HLA data for a long time, as exemplified by the coexistence of broad and split HLA serological nomenclature (e.g., HLA-DR2 which was split into DR15 and DR16). The complexity of this issue is even greater with molecular typing techniques, as reflected by the NMDP code system (http://bioinformatics.nmdp.org/HLA/Allele_Codes/Allele_Code_Lists/index.html). Bone Marrow Donor registries have developed software that can account for missing typing and HLA typing ambiguities, but aspects of BMD registry nomenclature (e.g., NMDP codes) can make these less user friendly options for some users.

Most analysis tools require two unambiguous allele assignments per locus, per individual as input. While "high resolution" typing systems may generate data with fewer ambiguities, both *allele* and *genotype ambiguities* still occur. *Allele ambiguity* results when polymorphisms that distinguish alleles fall outside of assessed regions, while *genotype ambiguity* results from an inability to establish phase between identified polymorphisms. Although both types of ambiguity can occur in all typing systems (including sequence based typing (SBT) methods), "high resolution" typing systems generate less ambiguous HLA data, and it is anticipated that new generation deep sequencing may give unambiguous data. There are currently no universal standards for making these allelic assignments; they are based on the individual investigators accumulated and empirical knowledge of the data under study, LD patterns, etc. Thus, allele assignments for the same specimen may vary between laboratories. We have developed ambiguity reduction rules based on "common and well documented" (CWD) alleles (Cano et al. 2007) (see HLA Guidelines ("Standard Operating Procedure for HLA Quality Control (QC) Pipeline" https://www.immport.org). In collaboration with the NCBI, we have developed an algorithm (currently in beta testing) for ambiguity reduction which extends the use of CWD information to also include population/regional allele frequency data (Single RM, Mack SJ, Dunivin R, Feolo M, in progress).

Modules for data validation and binning of alleles that are indistinguishable under certain circumstances (typing methods and/or time frames) are available in PyPop; the latter is usually necessary for performing meta-analyses of data. Additional modules in PyPop can be used for data quality control (QC) including: validation of allele names, testing of Hardy Weinberg proportions in controls (significant deviations may indicate errors in allele calls), and similarly with multilocus haplotype patterns (previously unobserved HLA B-C, DR-DQ and DPA-DPB haplotypes *may* flag errors). The Immunogenomic Data-Analysis Working Group (IDAWG) (www.igdawg.org) is an International Consortium of HLA and KIR researchers co-chaired by Steven J Mack and Jill A Hollenbach, with the aims to develop solutions and recommend community standards for the management and analysis of immunogenomic data. The IDAWG has partnered with the NIAID Bioinformatics Integration Support Contract (BISC) to develop the "Silver Standard" data recording and data validation pipeline for HLA genotype data collection (www.immport.org). This includes, in addition to data validation and binning, specific rules for the reporting and recording of HLA data in such a way that they will have optimal utility through time, and across studies, and for other researchers, in compliance with standards emerging from the HLA community regarding the reporting of HLA data (e.g., Helmberg 2000). More recently the IDAWG has developed nomenclature standardization modules and an alpha version of an HLA ambiguity reduction module.

KIR data present additional analytical issues that need to be dealt with, including that some loci are not found in all individuals. That is, there is presence/absence polymorphism in addition to allele level polymorphism. Although allele level KIR data has been less frequently reported, it is now being typed for more often. Customized algorithms, described in section IV.B, have been developed (e.g., Yoo et al 2007 - HAPLO-IHP, Gourraud et al 2007 - Estihaplo, Nowak et al 2008 – NullHap, Erdem et al 2009 - HAPLO-ASP). Both for HLA and KIR a central issue is whether prior restriction of the possible haplotype space should be used. For maximum likelihood models, any constraints will affect estimates derived from maximum likelihood based procedures. One approach for avoiding the potential removal of haplotypes of interest based on these constraints is to perform a two-stage estimation as follows. The first step is to determine a preliminary set of haplotypes and frequencies based on unconstrained estimation. This is followed by a constrained estimation where the haplotype space is based on a biologically meaningful reduction of the set of possible haplotypes. parameter space.

Haplotype estimation with the *cases* (patients) in case/control data is problematic for diseases that do not show a mode of inheritance close to recessive (allowing for incomplete penetrance) as associated marker genes may then deviate significantly from Hardy Weinberg proportions (HWP) (Thomson 1993, 1995a, b). HWP are assumed in application of the EM algorithm that is the basis for many haplotype estimation algorithms. The haplotype estimates may be most problematic for diseases with a dominant or additive mode of inheritance (allowing for incomplete penetrance) and a strong disease association.

Further, it is important to keep in mind that most analyses that use haplotype frequency data as input only take into consideration sampling variability and do not account for potential biases in the estimation of the haplotypes. Analyses are conducted assuming that HFs are known and not estimated. While resampling techniques can address aspects of this issue, they are not standard and not fully adapted to this problem.

II. Linkage disequilibrium (LD)

A. Causes of linkage disequilibrium

The original models of population genetics dealt chiefly with single loci, so that the genome was regarded as a collection of individual independent loci each undergoing its separate evolution. Until recently, most observed polymorphic genetic data were for loci sufficiently far apart on the genetic map that this assumption of independence was reasonable. Theoretical and simulation studies of selection and two and three locus systems demonstrated that LD of alleles could be maintained in a stable polymorphism (see Thomson 1977 for references). Simulations of multi-locus systems involving more loci and simple selection schemes were made (Lewontin 1964a, b, Franklin and Lewontin 1970) and it was thought that LD might be quite ubiquitous across the genome.

The hitchhiking effects produced on the allele frequencies and LD of linked neutral loci as a selected locus evolves towards its equilibrium value were also studied at this time (Smith and Haigh 1974, Thomson 1977, as well as later studies (Stephan et al. 2006). As an approximate generality, significant effects on the neutral locus were seen if the recombination fraction between the neutral and selected loci is smaller than the order of magnitude of the selected differences at the selected locus (Thomson 1977). Further, significant LD could be generated between neutral loci that initially showed no LD.

In a simulation study of LD in humans, Kruglyak (1999) (also see Slatkin 2008) predicted that little LD would be seen beyond 3kb. Apart from the HLA community, the subsequent description of LD using a block like structure of human variation (Daly et al. 2001, Jeffreys et al. 2001, Gabriel et al. 2002, Wall and Pritchard 2003) was somewhat of a surprise. In the 1970's non-random association of alleles at the HLA system were well established at the population level, and in fact this LD led to the recognition that multiple loci were involved for class I. The HLA-A and –B loci were shown to display a high level of LD given that they are about 1cm apart. Several authors have found significant LD across large distances in the HLA region (Begovich et al. 1992, Gordon et al. 2000, Sanchez-Mazas et al. 2000).

Theoretical studies previously established that LD can be created by various evolutionary factors: as well as selection (including disease) either directly on the two loci or indirectly via a hitchhiking event as detailed above, migration and admixture, inbreeding and genetic drift can create significant LD. The most likely cause, though, is historical — when a new mutation arises there is a non-random association created with respect to variation at other polymorphic loci — for neutral loci, this association is broken down by recombination at a rate of (1 – θ) per generation where θ is the recombination fraction between the two loci. Thus, the linkage disequilibrium in this case converges to zero (random association of alleles) with time as $(1 - \theta)^t$, where *t* is the number of generations. The more loosely linked two loci are, the faster the decay of LD. However, for very tightly linked loci LD may exist for a very long time.

For monogenic traits, most disease genes mapped to date show LD with markers sufficiently close to the disease gene, 0.5cM or even more distant in some cases, e.g., cystic fibrosis, Huntington disease, myotonic dystrophy, hemochromatosis, adult onset polycystic kidney disease, and many others. The familial breast cancer gene BRCA1 is an exception to this rule; LD is not seen with closely linked markers since, except in the Ashkenazi Jewish population, each family showing linkage to this gene often has a unique mutation. For disease association studies in general, it is important to use a combination of methods, including stratification analyses to take account of the LD structure of the data. This is particularly true for diseases with an immunogenetic component where the loci to account for have often been the antigen presenting classical HLA genes. Additional detail is given in section III.

B. Measures of the strength and significance of linkage disequilibrium

(i) Two locus theory

Deviation from random association of alleles A_i and B_j at two loci is most often measured using the coefficient of LD, $D_{\mathrm{A_iB_j}} = D_{\mathrm{ij}} = \mathrm{f}(\mathrm{A_iB_j}) - p_{\mathrm{Ai}}\, p_{\mathrm{Bj}}$ (as indicated in Section I), and we now simplify the notation to be $D_{ij} = f_{ij} - p_i q_j$. Lewontin (1964a) developed a normalized measure, $D'_{ij} = D_{ij}/D_{max}$, to address the fact that the range for D_{ij} is not independent of allele frequencies. Values for D'_{ij} range from –1 to 1.

There are several statistics that can be used to measure overall LD between the set of alleles at two different loci. Hedrick's (1987) D' statistic sums the contributions of all individual haplotypes (D'_{ij}) in a multi-allelic two-locus system, using the products of allele frequencies at the loci, p_i and q_j, as weights, with I and J representing the number of alleles at the A and B loci, respectively.

$$D' = \sum_{i=1}^{I} \sum_{j=1}^{J} p_i q_j \left|D'_{ij}\right|$$

W_n, also known as Cramer's V Statistic (Cramer 1946, Cohen 1988), is a second overall measure of LD between two loci. This statistic is based on a weighted average of individual D_{ij} values. The W_n statistic can be written as a re-expression of the overall Chi-square statistic, X^2_{LD}, normalized to be between zero and one.

$$W_n = \left(a.k.a.\ W_{AB}\right) = \left[\frac{\sum_{i=1}^{I}\sum_{j=1}^{J} D_{ij}^2/p_i q_j}{\min(I-1, J-1)}\right]^{\frac{1}{2}} = \left[\frac{X_{LD}^2/2N}{\min(I-1, J-1)}\right]^{\frac{1}{2}}$$

When there are only two alleles per locus, W_n is equivalent to the correlation coefficient between the two loci (see Section I), defined as $r = \sqrt{D_{11} / p_1 p_2 q_1 q_2}$

Both D' and r have been used in definitions of haplotype blocks for bi-allelic markers. Studies by Single et al. (2007a, b) using highly polymorphic HLA data have shown that the multi-allelic *D'* statistic is more strongly influenced by the number and total frequency of singleton haplotypes. For this reason W_n is considered a better measure of overall LD for highly polymorphic loci, although as stated above this and all other measures have considerable limitations.

A third measure of overall LD is a standardized version of the likelihood-ratio statistic used in a permutation test of the significance of overall LD between two loci (Zhao et al. 1999). It is defined in the section below on significance testing.

A fourth measure of LD is the number of haplotypes observed relative to the number expected under linkage equilibrium (LE), given the sample size and the number of alleles at the constituent loci for the haplotype. This statistic, defined below, is a modification (Mateu et al. 2001) of one minus the fraction of extra haplotypes statistic described by Slatkin (2000).

$$FNF = 1 - \frac{k_{obs} - k_{\min}}{k_{LE} - k_{\min}} = \frac{k_{LE} - k_{obs}}{k_{LE} - k_{\min}},$$

where k_{obs} is the number of different haplotypes seen in the sample, k_{min} is the minimum number of possible haplotypes (k_{min} is the larger value of the number of alleles, k, for each of the individual loci in the haplotype), and k_{LE} is the number of different haplotypes expected under LE: k_{LE} can be estimated by simulation conditioning on the sample size and allele frequencies at each locus as follows. For each observed sample of size N, $2N$ haplotypes are generated by sampling independently, with probabilities based on the allele frequencies for that population and LE, from the distribution of alleles at each locus. The number of distinct haplotypes is then counted in this sample simulated under LE. k_{LE} is estimated as the average value over 10,000 replications. *FNF* takes a value of zero if the number of haplotypes observed is equal to the number expected under LE and one if the minimum number of haplotypes is observed in the sample.

(ii) Significance testing

The significance of overall LD can be tested using the X^2_{LD} statistic defined above in the equation for W_n. It has better statistical properties for less highly polymorphic loci. Instead, the significance of overall LD between two loci can be tested using the permutation distribution of the likelihood-ratio statistic (Slatkin and Excoffier 1996) and is implemented in the PyPop program for population level data. The statistical significance of individual LD coefficients can be tested using $X^2_{ij} = (2N)D^2_{ij}/p_i(1\text{-}p_i)q_j(1\text{-}q_j)$ (Weir 1996).

Likelihood ratio based tests relate the likelihood of the observed data, with no constraints, to the likelihood of the data under a null hypothesis of linkage equilibrium (LE). If L_0 represents the likelihood of the data assuming LE (L_0 is computed with HFs given as the product of allele frequencies), and L_1 represents the likelihood of the data based on the estimated HFs without the assumption of LE, then the likelihood ratio statistic, $S = 2log(L_1/L_0)$, has an asymptotic Chi-square distribution with $(I\text{-}1)*(J\text{-}1)\ df$. The Chi-square approximation for S can be poor for highly polymorphic loci. A better approximation for the distribution of S under the null hypothesis that there is no LD can be approximated using the following resampling procedure. First, phenotypes at each locus are permuted between individuals. Second, the likelihood of the data, L_1, and a corresponding new value of S is computed for the "permuted sample" (L_0 is not changed by the permutations). These two steps are repeated a large number of times (e.g., 1000 times) to give the permutation distribution of S. The p-value for the test of no LD (i.e., LE) is the proportion of "permuted samples" yielding a value of S at least as large as the value computed from the original non-permuted data.

The third measure of overall LD mentioned in the above subsection is a standardized version of the likelihood-ratio statistic, *S*, based upon the permutation distribution (Zhao et al. 1999). The standardized statistic is defined as $\xi = \left(\sqrt{2df}/N\right)\left([S - \mu_S]/\sigma_S\right)$, where μ_s and σ_s are the mean and standard deviation of the permutation distribution for *S*, respectively.

(iii) Haplotype specific heterozygosity (HSH)

Haplotype specific heterozygosity (HSH) was developed (Malkki et al. 2005 and Single et al 2007c) to distinguish markers that act as proxies for classical HLA genes (low HSH) and markers that subdivide haplotypes at primary disease genes (high HSH) to detect additional genetic effects. HSH in these studies is the heterozygosity of a particular microsatellite (MSAT) marker given a specific HLA allele or haplotype. It is computed separately for each HLA haplotype by normalizing the MSAT allele frequencies found on the specific HLA haplotype and then calculating the heterozygosity using these normalized frequencies.

The normalized frequencies for the haplotype specific MSAT alleles are $p_i = h_i \Big/ \sum_{j=1}^{k} h_j$ and then $HSH = 1 - \sum_{i=1}^{k} p_i^2$, where *k* is the number of MSAT alleles observed on the specific HLA-A-B-DRB1 haplotype and $h_1, \ldots, h_k$ are the frequencies of the four-locus haplotypes (i.e., MSAT and HLA-A-B-DRB1). Markers with low HSH values can be used to predict specific HLA haplotypes or multi-locus genotypes to supplement the screening of HLA matched donors for transplantation. Markers with high HSH values will be most informative in studies investigating MHC-region disease susceptibility genes where HLA haplotypic effects are known to exist.

Looking at five common Caucasian HLA-A-B-DRB1 haplotypes, each had at least one MSAT marker with an HSH value of zero, indicating that only one MSAT allele was observed for that particular HLA haplotype. In terms of the ability of MSATs to predict HLA-A-B-DRB1 haplotypes, over 90% prediction probability was found for two of the common haplotypes using three MSATs. These preliminary data show the utility of this approach.

(iv) Asymmetric measures of linkage disequilibrium

Standard *multi-allelic* measures of LD are not fully informative for allele, haplotype, or amino acid level analyses of HLA data, where there is high polymorphism and often quite *different numbers of "alleles"* at the two loci being considered. The standard methods ignore possible differences in correlation patterns due to conditioning on one locus versus the other. W_n (= W_{AB}), the multi-allelic extension of the bi-allelic *r* LD correlation measure (i.e., square root of the r^2 measure), is *always* symmetric with respect to two loci; however, the pattern of variation can be quite different between two loci, especially when they have different numbers of alleles. Asymmetric measures of LD are needed to better explore the relationship between highly correlated amino acid sites and measures of functional and selective importance.

We have recently developed a set of new *asymmetric* conditional LD (CLD) measures (denoted $W_{A/B}$ and $W_{B/A}$) in order to dissect effects of correlated loci (e.g., amino acid sites within a gene, or combinations of genes in the *KIR* or *HLA* gene clusters), on measures of association and selection. These measures complement the HSH measure (Malkki et al. 2005, Single et al. 2007c) described above that was developed to assess how informative genetic markers are on different HLA haplotype backgrounds (e.g., MSATS, SNPs, and SNP blocks). With SNPs the asymmetric measures are recommended for analysis of haplotype block data, both for block-block comparisons of LD patterns, and for block, or single SNP, to HLA (or other primary disease locus) data. For bi-allelic loci, the two asymmetric measures coincide and equal the correlation measure *r*.

The new asymmetric LD measures are particularly relevant for sequence feature variant type (SFVT) analyses (Karp et al. 2010, Thomson et al. 2010) due to the large number of highly correlated SFs. We hypothesize that for polymorphic genes these asymmetric measures will be more powerful for identifying cases where conditional analyses can be applied to identify specific amino acids, and combinations of amino acids, directly involved in disease risk and for identifying additional disease genes in studies of allele and haplotype level variation.

(v) Three locus theory

We consider first the simplest case of three bi-allelic loci (loci A, B, and C, with alleles denoted A_1, A_2, B_1, B_2, and C_1, C_2). The eight haplotypes can be completely specified by seven parameters: three allele frequencies p_{A1}, p_{B1}, and p_{C1} ($p_{A2} = 1 - p_{A1}$, and similarly for p_{B2} and p_{C2}), three pairwise LD parameters denoted D_{AB}, D_{AC}, and D_{BC} (with $D_{AB} = D_{A1B1} = -D_{A1B2} = -D_{A2B1} = D_{A2B2}$, as above, and similarly for D_{AC} and D_{BC}) and one three locus LD parameter denoted D_{ABC} (with $D_{ABC} = D_{A1B1C1} = -D_{A1B1C2} = -D_{A1B2C1} = D_{A1B2C2} = -D_{A2B1C1} = D_{A2B1C2} = D_{A2B2C1} = -D_{A2B2C2}$).

$$f(A_1B_1C_1) = p_{A1}\, p_{B1}\, p_{C1} + p_{A1}\, D_{BC} + p_{B1}\, D_{AC} + p_{C1}\, D_{AB} + D_{ABC}$$

$$f(A_1B_1C_2) = p_{A1}\, p_{B1}\, p_{C2} - p_{A1}\, D_{BC} - p_{B1}\, D_{AC} + p_{C2}\, D_{AB} - D_{ABC}$$

$$f(A_1B_2C_1) = p_{A1}\, p_{B2}\, p_{C1} - p_{A1}\, D_{BC} + p_{B2}\, D_{AC} - p_{C1}\, D_{AB} - D_{ABC}$$

$$f(A_1B_2C_2) = p_{A1}\, p_{B2}\, p_{C2} + p_{A1}\, D_{BC} - p_{B2}\, D_{AC} - p_{C2}\, D_{AB} + D_{ABC}$$

$$f(A_2B_1C_1) = p_{A2}\, p_{B1}\, p_{C1} + p_{A2}\, D_{BC} - p_{B1}\, D_{AC} - p_{C1}\, D_{AB} - D_{ABC}$$

$$f(A_2B_1C_2) = p_{A2}\, p_{B1}\, p_{C2} - p_{A2}\, D_{BC} + p_{B1}\, D_{AC} - p_{C2}\, D_{AB} + D_{ABC}$$

$$f(A_2B_2C_1) = p_{A2}\, p_{B2}\, p_{C1} - p_{A2}\, D_{BC} - p_{B2}\, D_{AC} + p_{C1}\, D_{AB} + D_{ABC}$$

$$f(A_2B_2C_2) = p_{A2}\, p_{B2}\, p_{C2} + p_{A2}\, D_{BC} + p_{B2}\, D_{AC} + p_{C2}\, D_{AB} - D_{ABC}$$

The general formulation is

$$f(A_iB_jC_k) = p_{Ai}\, p_{Bj}\, p_{Ck} + p_{Ai}\, D_{BjCk} + p_{Bj}\, D_{AiCk} + p_{Ck}\, D_{AiBj} + D_{AiBjCk}$$

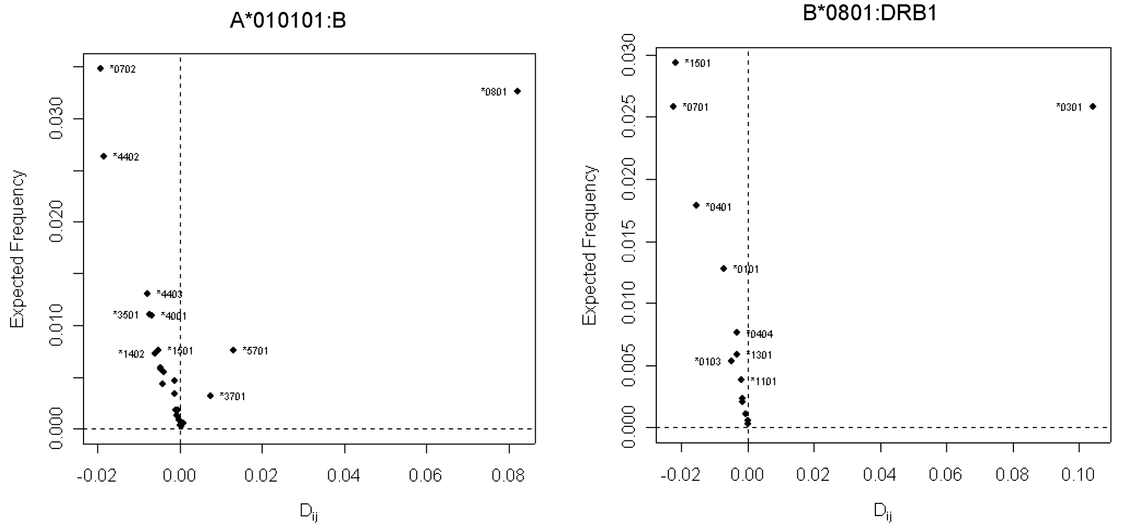
also see Bennett (1954), Geiringer (1994), Feldman et al. (1974), and for applications, see e.g., Thomson (1977) and Robinson et al. (1991, b).

C. Linkage Disequilibrium and Natural Selection

The disequilibrium pattern analysis (DPA) (Thomson and Klitz 1987, Klitz and Thomson 1987, Williams et al. 2004) and constrained disequilibrium values (CDV) (Robinson et al. 1991a, b, Grote et al. 1998) methods are two complementary approaches that have been used to detect selection acting on sets of HLA loci. Results from the two methods have identified specific HLA haplotypes that show signs of past selection in specific populations, e.g., in serological notation, the HLA haplotype A1 B8 DR3.

The DPA method identifies patterns of LD that are consistent with past selective events (Thomson and Klitz 1987, Klitz and Thomson 1987). For example, under selection in the recent past on the HLA class I haplotype A1 B8 (either directly on one or both of these alleles or via a hitchhiking event), then related haplotypes, i.e., all the A1 non-B8's, and B8 non-A1's are predicted to have an expected value of LD proportional to the frequency of the unshared allele. Figure II.C.1, from Williams et al (2004), show examples from a Caucasian population where the pattern is indicative of selection. The allele in the figure title is the one that is conditioned on (e.g., for the figure on the left, A*01:01:01 (A*010101 in the Figure) is conditioned on with frequencies and LD values plotted for all the B alleles observed, and similarly in the second figure the conditioning is on B*08:01 with all DRB1 alleles.

Figure II.C.1: Disequilibrium pattern analysis (DPA) for two of the pairwise combinations of the A:B:DRB1 loci.



The constrained disequilibrium values (CDV) method (Robinson et al. 1991a, 1991b, Grote et al. 1998) has been used to detect selection events in the HLA region by examining the pattern of pairwise LD values imposed by a three-locus system ($D_{ij}^{''}$) compared to those in the respective two locus system ($D_{ij}^{'}$). The difference between these two measures, $\Delta = \left| D_{ij}^{'} - D_{ij}^{''} \right|$, has a distribution which can be indicative of selection (as with DPA, not all cases of selection will be detected by this method).

The following criteria are used to infer selection based on Δ values:

1. If one of the three Δ values is positive (in practice greater than 0.1) and the remaining two are zero or negative, this is an indication of selection.
2. If more than one of the Δ values is positive, but one is much larger than the rest (in practice more than double the next largest), this is an indication of selection.
3. If all three Δ values are less than or equal to zero, or two are positive but close in value, no conclusion about selection can be drawn.

It was previously thought that in cases 1 and 2, the constraining allele that gives the high positive Δ value is the one experiencing selection (Robinson et al. 1991a). Further studies have shown that this can be misleading, especially when the center locus of three is the one leading to a high positive Δ value (Grote et al. 1998).

There are many instances where selection will not be detected via either of these methods. However, there is agreement between the results of application of these two methods.

III. Estimating HLA haplotype frequencies (HFs)

A. Population data

Estimated haplotypes and HFs play a central role in most genetic studies. Haplotype level analyses are important to studies of the etiology of human disease, selective forces acting on populations, and optimal sizes for BMDRs. Associations between markers and disease loci that are not evident with a single-marker locus may be identified in multi-locus marker analyses using estimated HFs. The term “haplotype” now includes any set of genetic polymorphisms (i.e., all DNA sequence variation including deletions) at contiguous loci. Except when recombination occurs, these neighboring genetic polymorphisms are co-transmitted by a single parental chromosome.

Haplotypes may be represented as blocks of DNA sequence variants like SNPs or groups of sequence variants can be abstracted into an allelic nomenclature at the level of a functional locus such as in the HLA system. For most studies, the type of experimental design and analysis is based on whether observed haplotypes will be determined by segregation analysis in families or estimation from phase-unknown unrelated individuals (Barnetche et al. 2005). Haplotypes are used for disease association mapping, QTL mapping, and imputing underlying genetic markers (Guan and Stephens 2008).

Early work on the estimation of HFs from unrelated genotype data were based on the EM algorithm with the assumption of HWP at the locus level (Dempster et al. 1977, Morton et al. 1983, Ott 1977, Piazza 1975, Yasuda 1978). Later work refined, explored, and extended aspects of the algorithm (Fallin and Schork 2000, Hawley and Kidd 1995, Kirk and Cardon 2002, Long et al. 1995, Single et al. 2002, Tishkoff et al. 2000). Application to haplotypes of SNPs (Niu et al. 2002, Qin et al. 2002, Stephens et al. 2001) and Bayesian methods (Niu 2004, Stephens and Donnelly 2003) are commonly used. The algorithm used in Estihaplo (Gourraud et al. 2004) allows for multiple alleles at each locus and missing data. There are at least two implementations that extend Clayton’s SNHAP algorithm to work with multi-allelic data. SNPHAP uses a progressive insertion algorithm to trim improbable haplotype assignments and allows for missing data. The EM Zipper algorithm in Arlequin 3.5 (Excoffier and Lischer 2010) and the haplo.em routine in Haplo Stats (Schaid et al. 2002) iteratively insert increasingly larger sets of loci into the EM estimation and remove poorly supported haplotypes from the space of possible haplotypes based on posterior probabilities. When all loci are included at the first stage and the fraction of trimmed haplotypes is set to zero (both user-defined parameters) these reduce to the ordinary EM algorithm, enumerating all possible haplotypes for the observed set of unphased genotypes.

The performance of haplotype frequency estimation algorithms is sensitive to various aspects of the population under study (Niu 2004). The accuracy of haplotype estimates is critical for association and candidate gene studies, and fine-mapping of disease genes. The presence or absence of specific low frequency alleles, and the corresponding estimated haplotypes that contain them, can influence the robustness of associations. In addition, available HF and LD estimation software are generally limited in their capacity to a few thousands of individuals; there is a need within the immunogenomic community for applications capable of handling very large (millions of individuals) data sets as well as very large numbers of alleles and haplotypes.

The diversity and complexity of Immunogenomic data poses additional challenges for haplotype estimation. Over the past thirty years, the Immunogenomic community has seen an exponential increase of the number of HLA alleles leading to regular nomenclature revisions. This phenomenon now also extends to the KIR genes (Robinson et al. 2006). In the HLA and KIR regions we have: heterogeneity of typing resolution, heterogeneity of typing techniques, heterogeneity of allele nomenclatures, continual discovery of new alleles; large numbers of alleles per locus observed; and high haplotype diversity. In addition, KIR and HLA data are very sensitive to ethnic background diversity. The potential for population sub-structure is particularly relevant for immunogenomic data due to the fact that HLA and KIR genes carry both selective and demographic histories of populations under study (see e.g., Meyer et al. 2006, 2007, Solberg et al. 2008 and references therein). These issues are exacerbated in BMDRs where sample sizes for specific research questions are often very large (> 100,000), as well as heterogeneity of typing levels and potential heterogeneity of ethnicities in a sample. Taken together, these features support the idea that the "HLA continues to provide new insights and remains in the vanguard of contemporary research in human genomics" (Vandiedonck and Knight 2009, p. 379).

For KIR and HLA, there is still much work to be done in order to understand haplotype estimation and potential biases. The frequency of the alleles, the sample size of the dataset, the fit or not of the single locus data to HWP, missing genotype information, and the variable levels of LD, all influence the accuracy of estimation.

B. Disease studies

While ABO blood group associations with disease were well known and replicated, the odds ratios (ORs) were all relatively small. Further, the biological mechanism of these disease associations was not known. In contrast, many striking HLA disease associations have been consistently found (some representative examples are given below in Table III.B.1 for serological data). The existence of LD has been a very powerful tool in mapping over 100 diseases (and possibly many more, the exact number is unknown) to the HLA region. An increased frequency of an HLA antigen (allele) in patients over that in an ethnically matched control population is inferred to be due either to the direct effect of the HLA antigen itself on disease, or to LD (association) of the HLA allele with the actual disease causing allele at a separate locus. In many cases the causal gene has been shown to be an antigen presenting classical HLA gene.

Table III.B.1: HLA associated diseases*

Disease	**HLA**	**Patients**	**Controls**	**Odds Ratio**
Ankylosing spondylitis				
	B27	90%	9%	81
Type 1 diabetes				
	DR3	52%	23%	3.6
	DR4	74%	24%	9
	DR3 or DR4	93%	43%	17.6
	DR2	4%	29%	0.1

Rheumatoid arthritis				
	DR4	81%	33%	8.7
Narcolepsy				
	DR2	95%	33%	38.6

* 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. For ease of reading, the data shown are older serological level HLA typing, rather than more recent molecular typing.

The difficulty in disease studies lies in identifying the actual predisposing loci and alleles in the context of this strong LD, and the fact that relatively common alleles in the population are involved. Historically, for HLA associated diseases, the detection of primary disease genes has taken a combination of typing both class I and II loci, study of their LD patterns within and between the two regions, comparisons of association strength among loci, stratification analyses, inter-population comparisons and inter-ethnic group comparisons (reviewed in Thomson et al. 2008, also see Thomson et al. 2007b). A breakdown in LD may be seen in some ethnic groups, particularly African and African American populations which show greater haplotype diversity than other populations. For example, in a study of African Americans (Just et al. 1997) over twice as many distinct DRB1-DQB1 haplotypes were present as found in a sample of Northern Europeans and consequently lower LD in African Americans than in African (McElroy et al. 2010).

A breakdown in the strength of LD may allow stratification analyses within the DR-DQ genes to identify the primary disease gene (e.g., see Mignot et al. 2001, 2007 for application to narcolepsy and Oksenberg et al. 2004 for application to multiple sclerosis). The direct involvement of HLA class II, and in some cases class I, genes in the disease process has been well documented for a number of diseases, for example, class II associations: HLA DR-DQ for type 1 diabetes (T1D) (reviewed in Thomson et al. 2007b), HLA-DR for rheumatoid arthritis (RA) and multiple sclerosis (MS), and HLA-DQ for narcolepsy, and class I associations: HLA-B for ankylosing spondilitis.

C. Rare Alleles and Haplotypes

Rare and low-frequency alleles contribute to single-copy haplotypes (“singletons”). The number and total frequency of singletons is a function of the polymorphism at the constituent loci, but haplotype frequency estimation accentuates the observed effect of polymorphism on singleton frequency as numerous haplotypes with estimated frequencies below one copy may occur. Estimated frequencies for rare haplotypes, which incorporate low-frequency alleles, are often incorrect, even when the EM algorithm finds the global maximum likelihood (Slatkin and Excoffier 1996, Fallin and Schork 2000, Tishkoff and Kidd 2004). In our analyses of the 13th and 14th IHWS Anthropology and Human Diversity (AHGD) data we evaluated several measures of overall LD in the context of the highly polymorphic HLA data (Single et al. 2007a, b, d). The overall D' and Wn statistics responded differently to the number and total frequency of “singleton” haplotypes, with D' more sensitive than Wn. The degree of this sensitivity is greater for more distant locus pairs, which usually have less overall LD. These relationships should be considered in HLA and KIR haplotype studies, where large numbers of low frequency estimates are common. Correction of HFs for estimates below a threshold of 1/2N, as initially implemented in Arelquin (Excoffier and Slatkin 1995), is not recommended since the frequencies will no longer be maximum likelihood estimates. Nevertheless, for small sample sizes in highly diverse datasets, such a correction can temporarily improve the convergence of EM iterations.

Low frequency or unique haplotypes can be of specific interest. For example, BMDRs need to characterize the degree of haplotype sharing among self identified race/ethnicity (SIRE) groups associated with each donor. Information about haplotypes that are rare or unique to certain SIRE groups is needed to make realistic registry size and donor recruitment predictions.

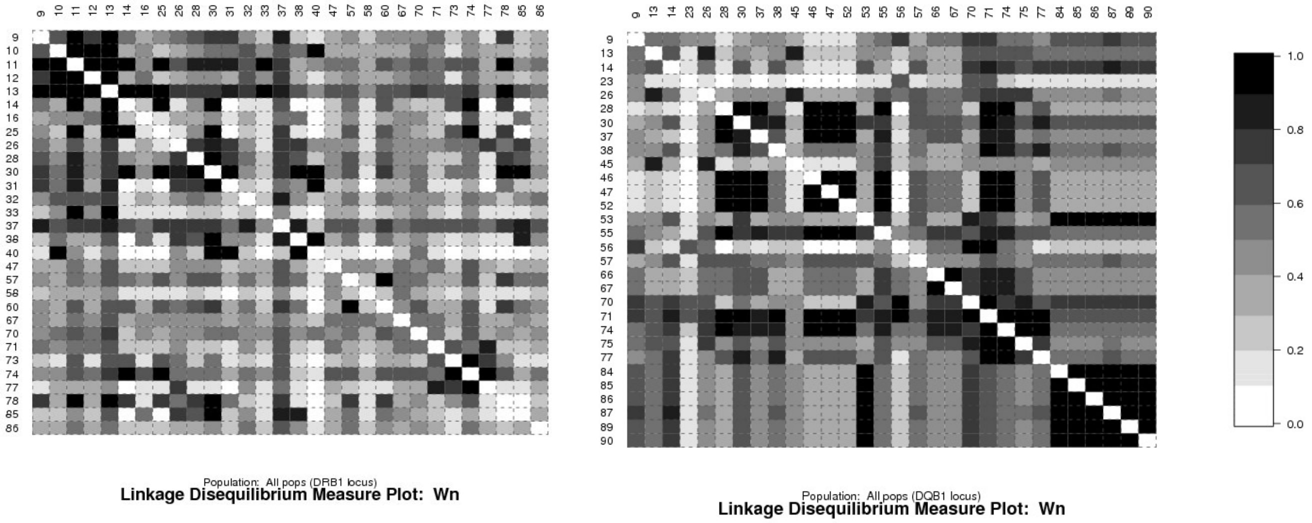
As mentioned in Section I.C, the haplotype frequency and LD estimates of BMDRs include genotype data for millions of individuals, with varying numbers of loci typed and levels of typing resolution. Registries have customized various software algorithms to run haplotype estimation algorithms on these large datasets. These also require high memory servers. It is remarkable that new bone marrow donors not only bring new phenotypes to the registry but also new haplotypes, suggesting that haplotype diversity is still underestimated with samples sizes in the millions Algorithms have been developed to deal specifically with HLA and KIR data and registry size samples

D. Amino acid level analyses

When a classical HLA gene (or genes) is identified as a primary disease risk factor, it is of interest to see if one can identify the combinations of biologically relevant amino acid residues directly involved in disease. This is difficult due to the pattern of amino acid variability, including the varying degrees of LD between amino acids both within, and between, the classical genes. Nonetheless, specific amino acid residues, as well as combinations of amino acids, have been implicated as potentially causal in a number of HLA associated diseases, e.g., HLA DRB1-DQB1 and type 1 diabetes risk (see Valdes and Thomson 1997, Valdes et al. 1997, Thomson et al. 2007b, including reviews of the literature), DRB1 and the “shared epitope” set of amino acids 70-74 and rheumatoid arthritis (reviewed in Imboden 2009). The recent development of a novel approach to genetic association analyses with genes/proteins sub-divided into biologically relevant smaller sequence features (SFs), and their variant types (VTs) (Karp et al. 2010), allows a systematic search focusing on the most likely actual causative genetic variants in HLA associated diseases. We have extended these analyses to include additional complementary methods of analysis, including the calculation of LD patterns of single amino acid and other SF variation, to guide our understanding of effects that may be due to high correlation of amino acid variation (Thomson et al. 2010).

Given the varying numbers of residues at polymorphic amino acid sites, e.g., in some cases up to 6, our new asymmetric CLD measures (see Section II.B above) are useful in identifying cases where stratification analyses can be carried out to aid identification of causal amino acids. The extent of LD seen within the classical HLA genes at the amino acid level varies considerably between genes. Data for the class II DRB1 and DQB1 loci are shown below (from Lancaster 2006).

Figure III.D.1:Linkage Disequilibrium plots for HLA-DRB1 and HLA-DQB1.



E. Accuracy of HLA haplotype frequency estimates

We must keep in mind a number of limiting factors regarding haplotype estimation. First, as mentioned in the Introduction (Section I.A), haplotype estimation with case/control data may be problematic for diseases that do not show a mode of inheritance close to recessive (allowing for incomplete penetrance) for the disease gene(s) in the genetic region under consideration. In this case, genotype frequencies at markers in high LD with the primary disease gene, or the primary disease gene itself, with an additive, dominant, or intermediate mode of inheritance (or other modes of inheritance that are not strictly recessive), are not expected to be in HWP (Thomson 1993, 1995a, b) which is a base assumption used in the EM algorithm to estimate HFs. This will lead to inaccuracies (the full extent of which is unknown at this time) in haplotype estimation and hence errors in association testing. Single et al. (2002) found (the study is described below) that the error in HLA haplotype estimation was more pronounced when the loci involved deviated from HWP, especially if there was excess heterozygosity. A similar result was found for bi-allelic markers by Fallin and Schork (2000); they also highlighted the fact that the EM algorithm is sensitive to sampling fluctuations. An algorithm to deal minimally with haplotype estimation for an additive disease model (the easiest algebraically and with expectations close to the dominant model) with case/control data is needed. Of course the mode of inheritance is usually unknown, but possibly with haplotype estimates from a recessive and an additive model, one could at least get a feel for the range of haplotype associations with disease.

Single et al. (2002) studied the accuracy of haplotype estimates with *known* HLA HFs for six HLA loci (A, B, DRB1, DQA1, DQB1, and DPB1) from family based data that were analyzed *as if unknown* using the expectation-maximization (EM) algorithm in ARLEQUIN. In general, the overall accuracy of the EM algorithm was shown to be very good. However, there were examples with large over- and underestimates of particular HFs, even for common haplotypes, especially, as mentioned above, when the loci involved deviated significantly from HWP. Estimating HFs for three or more loci and then collapsing over loci to generate two locus haplotypes was shown to often improve accuracy. This collapsing procedure was most beneficial when one of the loci in the two locus haplotype of interest deviated significantly from HWP and the locus collapsed over was in LD with the other loci.

The limit of accuracy of haplotype estimation for a large number of loci is of interest. It is likely that estimation accuracy will decline as the number of loci is greatly increased and many haplotypes then have very low estimated frequencies. We have investigated this issue using data from the type 1 diabetes genetic consortium (T1DGC), with dense SNP typing spanning 4.5Mb of the HLA region, including HLA class I, II and III regions in a total of 7,523 individuals in 1,640 pedigrees classified as of Caucasian ancestry (unpublished results of oral presentation of Briggs et al. 2007). A total of 2,050 SNPs with an average missing data of 1.2%, minor allele frequencies greater than 5%, and Hardy Weinberg Equilibrium p>0.001, were included. Pedigrees were excluded if there was missing genotype information for a founder; all remaining individuals within the nuclear family with genetic information were phased. We explored haplotype reconstruction with fastPHASE despite the assumptions of unrelatedness, due to its ability to phase large data sets fairly efficiently, and the comparable results obtained from other phasing algorithms (see Appendix A).

Haplotype blocks were assigned using Gabriel et al (2002). Haplotype estimation was then carried out using fastPHASE v1.2 (Scheet and Stephens 2006), with three levels of the number of SNPs considered at a time (Figure III.C.1): (i) 20 SNPs in one haplotype block (block 4), (ii) 159 SNPs spanning 5 blocks (including block 4), and (iii) all 2,050 SNPs. The same 20 SNPs of block 4 were extracted from each analysis for comparison, i.e., the data for the 5 blocks and all 2,050 SNPs at a time were collapsed to the 20 SNPs in Block 4 (Figure III.C.2). There was *consistency* (97% concordance), with and without collapsing, in estimation of the most common haplotypes of the 20 SNPs in Block 4 (note that we cannot assign accuracy in these analyses only concordance of results). These common 20 SNP haplotypes account for >95% of all haplotypes estimated in all three cases. Of interest, the total number of estimated haplotypes at the 20 SNPs in Block 4 decreased with increasing number of SNPs phased at a time (81, 60, and 56 for 20, 159, and 2,050 SNPs respectively), however, these all involve very rare haplotypes. Thus, although further study of this phenomenon is required, including the *accuracy* of the haplotype estimates (for which one would need to use simulated or other *known phase* data), collapsing or not over large numbers of SNPs does not appear to have major effects that would drastically alter the analyses of common haplotypes. These specific results also apply when haplotypes were assigned using MERLIN v1.0.1 (Abecasis et al. 2002, 2005).

Figure III.E.1: T1DGC HLA region SNP data

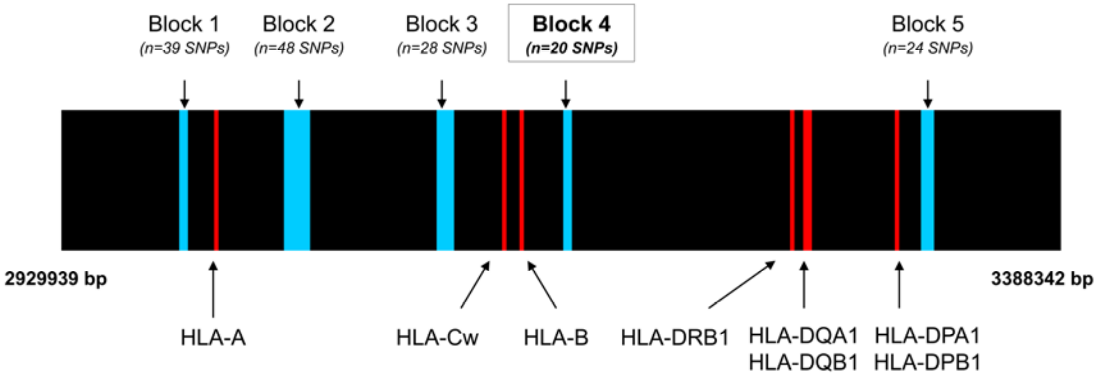
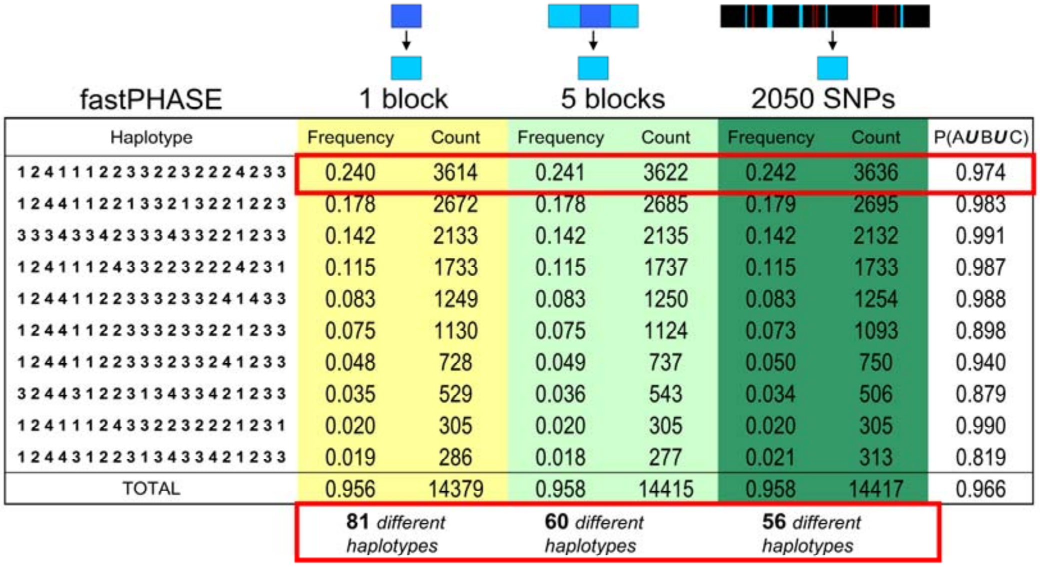


Figure III.E.2: Comparison of haplotype estimates from fastPHASE for block 4 with increasing amounts of SNP data



fastPHASE	1 block		5 blocks		2050 SNPs		
Haplotype	Frequency	Count	Frequency	Count	Frequency	Count	P(A*U*B*U*C)
1 2 4 1 1 1 2 2 3 3 2 2 3 2 2 2 4 2 3 3	0.240	3614	0.241	3622	0.242	3636	0.974
1 2 4 4 1 1 2 2 1 3 3 2 1 3 2 2 1 2 2 3	0.178	2672	0.178	2685	0.179	2695	0.983
3 3 3 4 3 3 4 2 3 3 3 4 3 3 2 2 1 2 3 3	0.142	2133	0.142	2135	0.142	2132	0.991
1 2 4 1 1 1 2 4 3 3 2 2 3 2 2 2 4 2 3 1	0.115	1733	0.115	1737	0.115	1733	0.987
1 2 4 4 1 1 2 2 3 3 3 2 3 3 2 4 1 4 3 3	0.083	1249	0.083	1250	0.083	1254	0.988
1 2 4 4 1 1 2 2 3 3 3 2 3 3 2 2 1 2 3 3	0.075	1130	0.075	1124	0.073	1093	0.898
1 2 4 4 1 1 2 2 3 3 3 2 3 3 2 4 1 2 3 3	0.048	728	0.049	737	0.050	750	0.940
3 2 4 4 3 1 2 2 3 1 3 4 3 3 4 2 1 2 3 3	0.035	529	0.036	543	0.034	506	0.879
1 2 4 1 1 1 2 4 3 3 2 2 3 2 2 2 1 2 3 1	0.020	305	0.020	305	0.020	305	0.990
1 2 4 4 3 1 2 2 3 1 3 4 3 3 4 2 1 2 3 3	0.019	286	0.018	277	0.021	313	0.819
TOTAL	0.956	14379	0.958	14415	0.958	14417	0.966
	81 *different haplotypes*		**60** *different haplotypes*		**56** *different haplotypes*		

Note: "1 block" refers to the 20 SNPs in the left column. "5 Blocks" refers to results for the same 20 SNPs after collapsing over results for 159 SNPs in 5 blocks. "2050 SNPs" refers to results for the 20 SNPs after collapsing over results for all 2050 SNPs.

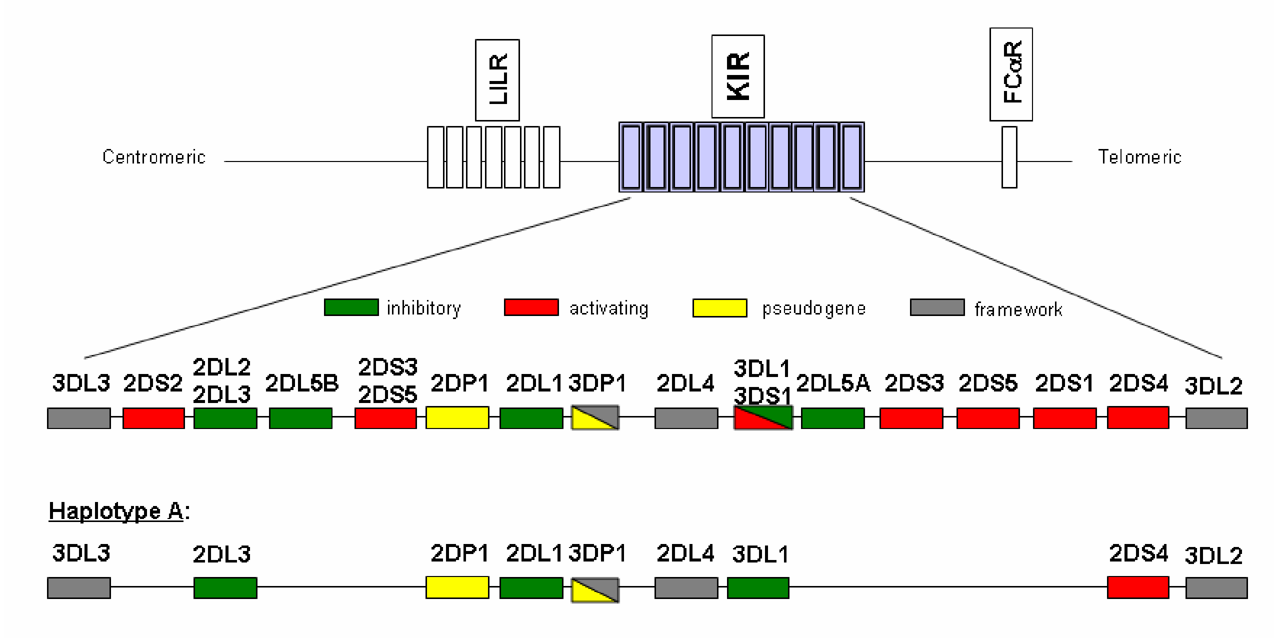
Note that even with family based data, haplotype assignment is subject to error. With family based data haplotype assignment is quite accurate for the highly polymorphic classical HLA loci. However, with SNP data there is always inherent ambiguity in haplotype assignment except for loci homozygous within each family. A number of haplotype phasing algorithms are available for SNP data, both for unrelated individuals and for pedigree data, with differences in terms of how ambiguous genotypes, missing data, and LD are taken into account (see Nui 2004, Salem et al. 2005). A brief overview of the details of the results from fastPHASE versus MERLIN is given in Appendix A.

IV. Estimating KIR haplotype frequencies

A. Background to the KIR gene complex

HLA class I proteins also serve as ligands for killer cell immunoglobulin-like receptors (KIR)(Vales-Gomez 1998), a family of inhibitory and activating receptors expressed on natural killer (NK) cells and a small percentage of cytotoxic T-cells to regulate cell killing and cytokine response (Biron 1997, Bashirova et al. 2006). The KIR gene complex is located on human chromosome 19q13.4 and includes 7-12 polymorphic genes per chromosome (a KIR haplotype). KIR haplotypes can be divided into two types, called A and B, depending on the activating genes present. KIR A haplotypes include only one activating gene (KIR2DS4), while KIR B haplotypes generally include more than one activating gene. Figure IV.A.1 illustrates the minimal KIR A haplotype, containing 7 expressed KIR genes and two pseudogenes.

Figure IV.A.1: KIR Gene Complex on Chromosome 19q13.4 and the minimal A haplotype



The extensive sequence homology among the KIR loci has resulted in a variety of KIR typing methods that generate KIR data with a wide range of resolutions. The level of resolution currently reported and analyzed in the literature is predominantly gene content variation (i.e., presence or absence of KIR loci). The extensive KIR locus- and haplotype-level polymorphisms include locus-level copy number variants. For example, KIR2DS3, can be found in the centromeric or telomeric regions of KIR haplotypes or in *both* portions. In addition to gene content variation, KIR genes are individually highly polymorphic (Table IV.A.1). Compounding these issues, nomenclature inconsistencies (e.g., distinct ‘loci’ that segregate as alleles, or duplicate loci on the centromeric and telomeric ends of the cluster) confound consistent interpretation of results by the KIR community, and can introduce analytical biases if not appropriately addressed.

Table IV.A.1: KIR Polymorphism (2010, http://www.ebi.ac.uk/ipd/kir/stats.html**)**

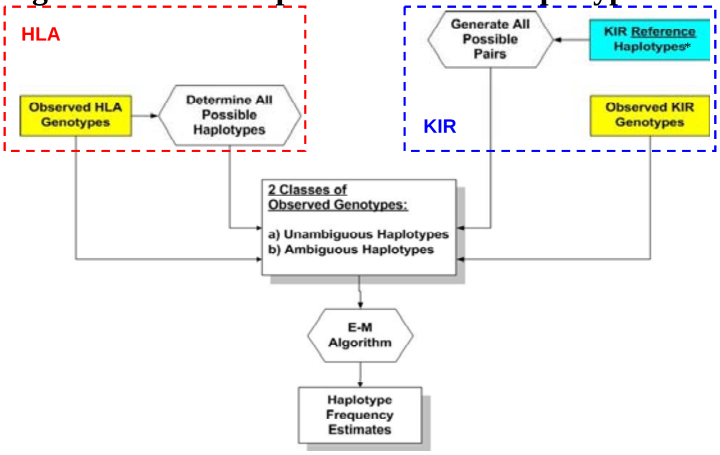
Gene	**2DL1**	**2DL2**	**2DL3**	**2DL4**	**2DL5**	**2DS1**	**2DS2**	**2DS3**	**2DS4**	**2DS5**
Alleles	43	29	33	47	41	15	22	14	30	15
Proteins	24	12	17	22	18	7	8	5	13	10
Nulls	1	0	1	0	0	0	0	1	0	0
Gene	**3DL1**	**3DS1**	**3DL2**	**3DL3**	**2DP1**	**3DP1**				
Alleles	74	16	84	107	22	23				
Proteins	58	12	62	56	0	0				
Nulls	1	1	1	0	0	0				

B. Haplotype frequency estimation for KIR

Because some KIR genes are present *only* on certain haplotypes, the space of possible KIR haplotypes excludes some locus combinations that *could* be generated from the observed genotypic data. The EM algorithm for estimating KIR HFs can be modified to account for this reduced combinatorial space (Figure IV.B.1) using a set of reference haplotypes using an a priori list of known/possible haplotypes (reference haplotypes) to constrain the EM algorithm (Gourraud et al

2007, Yoo et al 2007). The user-designated a priori haplotype list is said to "*span*" a set of observed genotypes if each observed genotype can be generated from at least one pair of haplotypes in the list. If the list does not span the observed genotypes the resulting estimates must be carefully interpreted.

Figure IV.B.1: Comparison of HLA Haplotype Estimation and KIR Haplotype Estimation


HLA
Observed HLA Genotypes
Determine All Possible Haplotypes
Generate All Possible Pairs
KIR Reference Haplotypes*
KIR
Observed KIR Genotypes
2 Classes of Observed Genotypes:
a) Unambiguous Haplotypes
b) Ambiguous Haplotypes
E-M Algorithm
Haplotype Frequency Estimates

Several KIR haplotype frequency estimation studies have noted shortcomings in the use of such constraints, imposed by the need to specify predefined haplotype patterns. Gourraud et al (2007) found that accuracy measures related to haplotype identification for KIR were particularly low for fewer than 200 individuals and suggested that more than 500 individuals would provide an acceptable estimation accuracy. Further simulations studies also suggest that the 1/2N correction may improve the convergence of the estimations and that estimation works better when achieved separately in the centromeric and telomeric region of the KIR gene cluster (delimited by KIR2DL4).

Yoo et al. (2007) developed an algorithm in the HAPLO-IHP software that incorporates information about specific haplotype patterns and an a priori list of haplotypes. In this approach, the algorithm first constructs a minimal set of haplotypes to resolve observed genotypes and then uses the EM algorithm to estimate HFs. A haplotype pattern file allows the user to require the presence of anchor genes, or specify an allelic relationship between specific KIR loci. Yoo et al. noted that rare or unusual haplotypes that are incompatible with constraints may be incorrectly rejected. When the a priori list of user-defined haplotypes does not span the observed genotypes, new haplotypes are "constructed" in an attempt to satisfy any user defined haplotype patterns.

Single et al. (2008) assessed the accuracy of KIR HF estimates from the HAPLO-IHP program using measures that compare the true HFs with those estimated by the EM algorithm in a sample of 90 unrelated individuals from the CEPH families. After adding three new haplotypes identified by segregation analysis to the a priori list in order to span the set of observed genotypes, no spurious haplotypes were created by the program. While the more common haplotypes were estimated relatively well, a large number of low frequency haplotypes were either missed, or were estimated but not actually present in the sample. Application of these methods to 23 global populations from Single et al. (2007e) revealed that most populations had individuals with presence/absence genotype profiles that could not be constructed using pairs of haplotypes from Khakoo and Carrington's (2006) list of well documented KIR haplotypes (i.e., these individuals had "new" haplotypes) (Single et al. 2008). The percentage of these individuals was greater than five percent in several populations and was more than 10% in four of six African populations. This higher percentage in the African populations is expected as these populations have lower LD, more haplotypes, and are less studied than Europeans for KIR. When using a set of a priori haplotypes it is important to first check to see if they span the set of observed genotypes and to take these findings into account when using estimated HFs in downstream analyses.

NullHap (Nowak and Ploski 2008) uses a modified EM algorithm to handle multi-allelic loci and loci with null alleles. Loci with a potential null allele are identified prior to estimation. The maximization step is then modified to account for the fact that a heterozygote with a null allele would have the same observed genotype as a homozygote without a null allele. Comparisons to results from Haplo-IHP on bi-allelic presence/absence data gave a difference of roughly 3% in estimated frequencies.

Haplo-ASP (Erdem et al 2009) uses an answer set programming (ASP) algorithms to accommodate restrictions on observable haplotype patterns and can work with multi-allelic data. ASP algorithms find a minimal set (answer set) that satisfies a prespecified list of constraints. As with Haplo-IHP, multi-locus genotypes that do not conform to pre-specified haplotype patterns can be problematic. Haplo-ASP had slightly higher accuracy than Haplo-IHP, based on the data in Yoo et al. (2007).

Appendix A: SNP haplotype estimation algorithms

There are several algorithms currently available for reconstructing extended haplotypes from SNP data, including fastPHASE (Scheet and Stephens 2006) and MERLIN (Abecasis et al. 2002, 2005), which we discuss in more detail below, and BEAGLE (Browning and Browning 2007), FAMHAP (Becker and Knapp 2004), HAPLORE (Zhang et al. 2005), and PHASE (Stephens and Donnelly 2003). A summary of the main features of MERLIN and fastPHASE are given in Table A.1, and of the other algorithms in Table A.2.

Appendix TableA.1: Comparing the main features of MERLIN and fastPHASE

	MERLIN	**fastPHASE**
Type of Data	Pedigrees	Unrelated individuals
Genetic Data	SNPs, multi-allelic	SNPs
Haplotype Algorithm	Sparse binary trees to summarize gene flow (Lander-Green algorithm) within a PEDIGREE only	Hidden Markov Model & EM algorithm across ALL individuals
Missing genotypes	Imputes SOME based on gene flow in pedigree	Imputes ALL using EM
Ambiguous Haplotypes	For heterozygous loci within a family	None
LD	Default: no LD; can factor in however results in more missing data if recombination is suspected	Factored in to analysis. Flexible: allows block-patterns to gradual decline with distance
Limitation	Map file: centimorgrams (assumed 1,000,000 bp =1 cM)	Assumes loci are equally spaced

Appendix TableA.2: Comparison of the main features of PHASE, HAPLORE, BEAGLE, and FAMHAP

PHASE v2.1 (Stephens and Donnelly 2003)

- Used for HapMap, though not recommended for large sample size and no more than 250 loci

HAPLORE (Zhang et al. 2005)

- Simple pedigrees, no recombination

BEAGLE (Browning and Browning 2007)

- Unrelated individuals

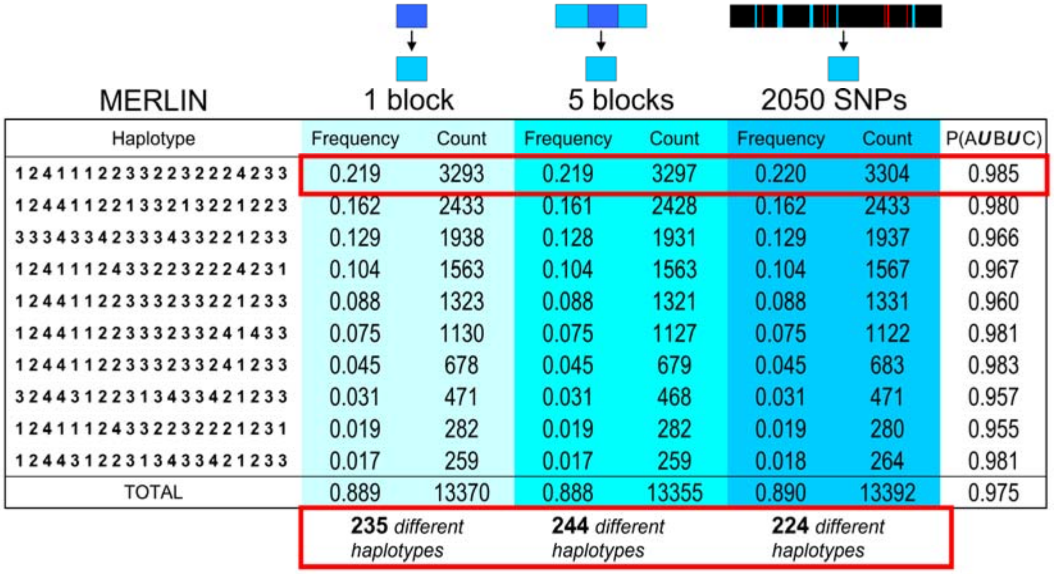
FAMHAP (Becker and Knapp 2004)

- Case-control data or family data (no recombination)
- Limited loci
- only 263 haplotypes are allowed, i.e. 63 SNPs or 21 microsatellite markers with 8 alleles

The haplotype estimation of fastPHASE and MERLIN were compared using the T1DGC HLA + SNP data described and analyzed in Section III.E above (Briggs et al. 2007). In summary, the fastPHASE algorithm assigns haplotypes in unrelated individuals under the assumption that similar haplotypes tend to cluster over short regions and cluster membership follows a hidden Markov model along the chromosome; while the MERLIN algorithm reconstructs haplotypes using sparse binary trees to identify the most likely gene flow within a pedigree. Furthermore, both algorithms implement methods for estimating missing genotypes, however not all genotypes were readily imputed with MERLIN v1.0.1, and thus affected the evaluation (*--infer* option) (for more details see Table A.1 above).

There was high concordance (87.3% overlap) between the algorithms with respect to phased chromosomes among the ten most frequent haplotypes based on the 20 SNPs within a single block (Block 4). A total of 81 and 235 (35% with at least one missing locus) haplotypes were reconstructed using fastPHASE and MERLIN, respectively (see Figure III.E.1 (above in text) and Figure A.1 below). The discrepancy in haplotype counts is due to both the persistence of missingness and ambiguity in haplotype assignment in MERLIN as inheritance patterns are limited to within a pedigree; ambiguity arises when all individuals within a pedigree are heterozygous at a locus. This limitation further reduced the phasing concordance with increasing genetic data. For example, using 159 SNPs from 5 haplotype blocks, a total of 6,899 (4,878 with an observed count of one (N=1)) and 5,702 (2,200 with N=1; 37.2 % with at least one missing locus) haplotypes were assigned using fastPHASE and MERLIN, respectively; with 33.5% concordance in chromosome reconstruction among the ten most frequent haplotypes (60 and 244 respectively were seen for Block 4 from the collapsed data). When using the full data set of 2,050 SNPs, a total of 14,433 (13,942 with N=1) and 9,071 (5,292 with N=1; 90% with at least one missing locus) haplotypes were assigned using fastPHASE and MERLIN, respectively; with less than 1% concordance in chromosome reconstruction among the ten most frequent haplotypes (56 and 224 respectively were seen for Block 4 from the collapsed data). However, the most frequent haplotypes for Block 4 derived from both phasing approaches appear similar, when using the original 20 SNPs or the collapsed data from 159 SNPs or 2,050 SNPs. A difference between the two approaches is that the most frequent haplotypes explain more chromosomes in fastPHASE than MERLIN (as above we cannot assign which is more accurate). The key differences are that with increasing chromosomal regions to phase, fastPHASE generated many more unique haplotypes despite imputing all missing loci; while MERLIN was limited by missingness, it was still able to construct haplotypes within families, marginally reducing the overall haplotype diversity.

Appendix FigureA.1 : Comparing haplotype estimates for block 4 (see Figure III.E.1) using increasing genetic data in MERLIN



MERLIN	1 block		5 blocks		2050 SNPs		
Haplotype	Frequency	Count	Frequency	Count	Frequency	Count	P(AUBUC)
1 2 4 1 1 1 2 2 3 3 2 2 3 2 2 2 4 2 3 3	0.219	3293	0.219	3297	0.220	3304	0.985
1 2 4 4 1 1 2 2 1 3 3 2 1 3 2 2 1 2 2 3	0.162	2433	0.161	2428	0.162	2433	0.980
3 3 3 4 3 3 4 2 3 3 3 4 3 3 2 2 1 2 3 3	0.129	1938	0.128	1931	0.129	1937	0.966
1 2 4 1 1 1 2 4 3 3 2 2 3 2 2 2 4 2 3 1	0.104	1563	0.104	1563	0.104	1567	0.967
1 2 4 4 1 1 2 2 3 3 3 2 3 3 2 2 1 2 3 3	0.088	1323	0.088	1321	0.088	1331	0.960
1 2 4 4 1 1 2 2 3 3 3 2 3 3 2 4 1 4 3 3	0.075	1130	0.075	1127	0.075	1122	0.981
1 2 4 4 1 1 2 2 3 3 3 2 3 3 2 4 1 2 3 3	0.045	678	0.045	679	0.045	683	0.983
3 2 4 4 3 1 2 2 3 1 3 4 3 3 4 2 1 2 3 3	0.031	471	0.031	468	0.031	471	0.957
1 2 4 1 1 1 2 4 3 3 2 2 3 2 2 2 1 2 3 1	0.019	282	0.019	282	0.019	280	0.955
1 2 4 4 3 1 2 2 3 1 3 4 3 3 4 2 1 2 3 3	0.017	259	0.017	259	0.018	264	0.981
TOTAL	0.889	13370	0.888	13355	0.890	13392	0.975
	235 *different haplotypes*		**244** *different haplotypes*		**224** *different haplotypes*		

Note: "1 block" refers to the 20 SNPs in the left column. "5 Blocks" refers to results for the same 20 SNPs after collapsing over results for 159 SNPs in 5 blocks. "2050 SNPs" refers to results for the 20 SNPs after collapsing over results for all 2050 SNPs.

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30(1): 97-101.

Abecasis GR, Wigginton JE. (2005) Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 77(5):754-767.

Barnetche T, Gourraud PA, Cambon-Thomsen A. (2005) Strategies in analysis of the genetic component of multifactorial diseases; biostatistical aspects. Transplant Immunology 14(3-4): 255-266.

Becker T, Knapp M. (2004) Maximum-likelihood estimation of haplotype frequencies in nuclear families. Genet Epidemiol 27(1): 21-32.

Begovich AB, McClure GR, Suraj VC, Helmuth RC, Fildes N, Bugawan TL, Erlich HA, Klitz W. (1992) Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. J Immunol 148(1): 249-258.

Bennett JH. (1954) On the theory of random mating. Annals of Eugenics 18: 311-317.

Biron, C.A (1997) Activation and function of natural killer cell responses during viral infections. Curr Opin Immunol. 9(1): p. 24-34.

Bashirova, A.A., et al.(2006) The Killer Immunoglobulin-like Receptor Gene Cluster: Tuning the Genome for Defense. Annu Rev Genomics Hum Genet.

Briggs FB, Shao X, Thomson G, Barcellos BF. (2007) A comparison of haplotype estimation methods within the MHC using genetic data from Type 1 Diabetes families. Oral presentation at the T1DGC MHC Results Workshop, August 27-28, 2007, Washington DC.

Browning SR, Browning BL. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81(5): 1084-1097.

Cano P, Fernández-Viña M. (2009) Two sequence dimorphisms of DPB1 define the immunodominant serologic epitopes of HLA-DP. Hum Immunol 70(10): 836-843.

Cano P, Klitz W, Mack SJ, Maiers M, Marsh SG, Noreen H, Reed EF, Senitzer D, Setterholm M, Smith A, Fernandez-Vina M. (2007) Common and Well-Documented HLA Alleles Report of the Ad-Hoc Committee of the American Society for Histocompatiblity and Immunogenetics. Hum Immunol. 68(5): 392-417.

Clayton D. SNPHAP: A program for estimating frequencies of large haplotypes of SNPs. (Version 1.3) [Software and documentation]. Unpublished instrument. Available from http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt

Cohen J. (1988) Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cramer H. (1946) Mathematical Methods of Statistics. Princeton, NJ: Princeton University Press.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. (2001) High-resolution haplotype structure in the human genome. Nature 29: 229-252.

Dempster AP, Laird NM, Rubin DB. (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39: 1-38.

Erdem E, Erdem O, Türe F. (2009) HAPLO-ASP: Haplotype Inference Using Answer Set Programming. Lecture Notes in Computer Science 5753/2009: 573-578.

Excoffier L and Lischer H (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 10: 564-567.

Excoffier L and Slatkin M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12(5): 921-927.

Fallin D, Schork NJ. (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67: 947-959.

Feldman MW, Franklin I, Thomson G. (1974) Selection in complex genetic systems. I. The symmetric equilibria of the three-locus symmetric viability model. Genetics 76: 145-162.

Franklin I, Lewontin RC. (1970) Is the gene the unit of selection? Genetics 65(4): 707-34.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. (2002) The structure of haplotype blocks in the human genome. Science 296(5576): 2225-2229.

Gordon D, Simonic I, Ott J. (2000) Significant evidence for linage disequilibrium over a 5-cM region among Afrikaners. Genomics 66: 87-92.

Gourraud PA, Genin E, Cambon-Thomsen A (2004) Handling missing values in population data: consequences for maximum likelihood estimation of haplotype frequencies. European Journal of Human Genetics (2004) 12, 805–812.

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.

Geiringer H. (1994) On the probability theory of linkage in Mendelian heredity. Annals of Mathematical Statistics 15: 25.

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

Guan Y, Stephens M. (2008) Practical issues in imputation-based association mapping. PLoS Genet 4(12): e1000279.

Hawley ME, Kidd KK. (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multisite haplotypes. J Hered 86: 409-411.

Hedrick PW. (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117: 331-41 (http://onlinelibrary.wiley.com/doi/10.1002/gepi.20024/pdf)

Hedrick PW, Thomson G. (1986) A two-locus neutrality test: applications to humans, E. coli and lodgepole pine. Genetics 112(1): 135-156.

Helmberg W. (2000) Storage and utilization of HLA genomic data--new approaches to HLA typing. Rev Immunogenet. 2(4): 468-476.

Imboden JB. (2009) The immunopathogenesis of rheumatoid arthritis. Annu Rev Pathol 4: 417-434.

Jeffreys AJ, Kauppi L, Neumann R. (2001) Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics 29: 217-222.

Just JJ, King MC, Thomson G, Klitz W. (1997) African-American HLA class II allele and haplotype diversity. Tissue Antigens 49(5): 547-55.

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.

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

Kirk KM, Cardon LR. (2002) The impact of genotyping error on haplotype reconstruction and frequency estimation. Eur J Hum Genet 10: 616-622.

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

Klitz W, Thomson G, Borot N, Cambon-Thomsen A. (1992) Evolutionary and population perspectives of the human HLA complex. Evolutionary Biology 26: 35-72.

Kruglyak L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics 22: 139-144.

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

Lancaster A, Nelson MP, Single RM, Meyer D, Thomson G. (2003) PyPop: a software framework for population genomics: analyzing large-scale multi-locus genotype data. Pacific Symposium of Biocomputing 2003: 514-525.

Lancaster A, Nelson MP, Single RM, Meyer D, Thomson G. (2007a) Software framework for the Biostatistics Core. In: J.A. Hansen, ed: Immunobiology of the Human MHC. Proceedings of the 13th International Histocompatibility Workshop and Conference. Vol I, IHWG Press, Seattle, 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.

Lancaster A, Nelson MP, Meyer D, Single RM. (2008) PyPop User Guide: User Guide for Python for Population Genetics, Version 0.7.0. Available at http://www.uvm.edu/~pypop/docs and http://www.pypop.org/docs, Copyright © 2003-2008. Regents of the University of California.

Lewontin RC. (1964a) The interaction of selection and linkage I. General considerations; heterotic models. Genetics 49: 49-67.

Lewontin RC. (1964b) The interaction of selection and linkage II. Optimum models. Genetics 50: 757-782.

Lewontin RC. (1988) On measures of gametic disequilibrium. Genetics 120: 849-852.

Long JC, Williams RC, Urbanek M. (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56: 799-810.

Maiers M, Gragert L, Klitz W. (2007) High-resolution HLA alleles and haplotypes in the United States population, Human Immunology, 68(9), 779-788.

Malkki M, Single R, Carrington M, Thomson G, 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 MP, Single RM, Wilson MJ, Trowsdale J, Carrington M. (2008) KIR haplotypes defined by segregation analysis in 59 Centre d'Etude Polymorphisme Humain (CEPH) families. Immunogenetics, 60: 767-774.

Mateu E, Calafell F, Lao O, Bonne-Tamir B, Kidd JR, Pakstis A, Kidd KK, Bertranpetit J. (2001) Worldwide genetic analysis of the CFTR region. Am J Hum Genet 68(1): 103-117.

McElroy JP, Cree BA, Caillier SJ, Gregersen PK, Herbert J, Khan OA, Freudenberg J, Lee A, Bridges SL Jr, Hauser SL, Oksenberg JR, Gourraud PA. (2010) Refining the association of MHC with multiple sclerosis in African Americans. Hum Mol Genet 19(15): 3080-3088.

Meyer D, Single RM, Mack SJ, Erlich HA, 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 RM, Mack SJ, Lancaster A, Nelson MP, Fernández-Viña M, Erlich H, Thomson G. (2007) Single locus polymorphism of classical HLA genes. In: J.A. Hansen, ed: Immunobiology of the Human MHC. Proceedings of the 13th International Histocompatibility Workshop and Conference. Vol I, IHWG Press, Seattle, 653-704.

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

Mignot E, Lin L, Rogers W, Honda Y, Qiu X, Lin X, Okun M, Hohjoh H, Miki T, Hsu S, Leffell M, Grumet F, Fernandez-Vina M, Honda M, Risch N. (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" (JA Hansen, ed.), pp. 817-812. IHWG Press, Seattle.

Morton NE, Simpson SP, Lew R, Yee S. (1983) Estimation of haplotype frequencies. Tissue Antigens 22(4): 257-62.

Niu T. (2004) Algorithms for inferring haplotypes. Genet Epidemiol 27(4): 334-347.

Niu T, Qin ZS, Xu X, Liu JS. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70: 157-169.

Nowak RM, Ploski R. (2008) NullHap--a versatile application to estimate haplotype frequencies from unphased genotypes in the presence of null alleles. BMC Bioinformatics 5(9): 330-338.

Oksenberg JR, Barcellos LF, Cree BA, Baranzini SE, Bugawan TL, Khan O, Lincoln RR, Swerdlin A, Mignot E, Lin L, Goodin D, Erlich HA, Schmidt S, Thomson G, Reich DE, Pericak-Vance MA, Haines JL, Hauser SL. (2004) Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans. Am J Hum Genet 74(1): 160-107.

Ott J. (1977) Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis. Ann Hum Genet. 40: 443-454.

Piazza A. (1975) Haplotypes and linkage disequilibria from three-locus phenotypes. In Histocompatibility testing (ed. Kissmeyer-Nielsen, F.), pp. 923-927. Copenhagen: Munskgaard.

Qin ZS, Niu T, Liu JS. (2002) Partition-ligation-expectationmaximization algorithm for haplotype inference with singlenucleotide polymorphisms. Am J Hum Genet 71: 1242-1247.

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

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.

Robinson J, Waller MJ, Fail SC, Marsh SG. (2006) The IMGT/HLA and IPD databases. Hum Mutat 27:1192-1199.

Salamon H, Klitz W, Easteal S, Gao X, Erlich HA, Fernandez-Vina M, Trachtenberg EA, McWeeney SK, Nelson MP, Thomson G. (1999). Evolution of HLA Class II molecules: Allelic and amino acid site variability across populations. Genetics 152(1): 393-400.

Salem RM, Wessel J, Schork NJ. (2005) A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. Hum Genomics 2(1): 39-66.

Sanchez-Mazas A, Djoulah S, Busson M, Le Monnier de Gouville I, Poirier J-C, Dehay C, Charron D, Excoffier L, Schneider S, Langaney A, Dausset J, Hors J. (2000) A linkage disequilibrium map of the MHC region based on the analysis of 14 loci haplotypes in 50 French families. Eur J Hum Genet 8: 33-41.

Scheet P, Stephens M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78(4): 629-644.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet. 2002;70:425–34.

Single RM, Meyer D, Hollenbeck J, Nelson M, Noble JA, Erlich HA, Thomson G. (2002) Haplotype frequency estimation in patient populations: The effect of departures from Hardy-Weinberg proportions and collapsing over a locus in the HLA region. Gen Epidemiol 22: 186-195.

Single RM, Meyer D, Mack SJ, Lancaster A, Nelson MP, Fernández-Viña M, Erlich H, Thomson G. (2007a) Haplotype Frequencies and Linkage Disequilibrium among classical HLA genes. In: J.A. Hansen, ed: Immunobiology of the Human MHC. Proceedings of the 13th International Histocompatibility Workshop and Conference. Vol I, IHWG Press, Seattle, 705-746.

Single RM, Meyer D, Thomson G. (2007b) Statistical methods for analysis of population genetic data. In: J.A. Hansen, ed: Immunobiology of the Human MHC. Proceedings of the 13th International Histocompatibility Workshop and Conference. Vol I, IHWG Press, Seattle, 518-522.

Single RM, Malkki M, Thomson G, Mather KA, Carrington M, Petersdorf E. (2007c) Linkage disequilibrium and HLA-A: B: DRB1 haplotype probabilities for Class I, II, III microsatellite markers in unrelated donor hematopoietic stem cell transplantation. In: J.A. Hansen, ed: Immunobiology of the Human MHC. Proceedings of the 13th International Histocompatibility Workshop and Conference. Vol I, IHWG Press, Seattle, 1372-1377.

Single RM, Meyer D, Mack SJ, Lancaster A, Erlich HA, Thomson G. (2007d) 14th International HLA and Immunogenetics Workshop: Report of progress in methodology, data collection, and analyses. Tissue Antigens 69 (Suppl. 1): 185-187.

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

Single RM, Martin MP, Meyer D, Gao X, Carrington M. (2008) Methods for assessing gene content diversity of KIR with examples from a global set of populations. Immunogenetics 60(12): 711-25.

Slatkin M. (2000) Balancing selection at closely linked, overdominant loci in a finite population. Genetics 154: 1367–1378.

Slatkin M. (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9: 477-485.

Slatkin M, Excoffier L. (1996) Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. Heredity 76: 377-383.

Smith JM, Haigh J. (1974) The hitch-hiking effect of a favourable gene. Genet Res 23(1): 23-35.

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.

Stephan W, Song YS, Langley CH. (2006) The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172(4): 2647-2663.

Stephens M, Donnelly P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73: 1162-1169.

Stephens M, Smith NJ, Donnelly P. (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68: 978-989.

Thomson G. (1977) The effect of a selected locus on linked neutral loci. Genetics 85: 753-788.

Thomson G. (1993) The AGFAP method: applicability under different ascertainment schemes and a parental contributions test. Genetic Epidemiology 10: 289-310.

Thomson G. (1995a) Analysis of complex human genetic traits: an ordered-notation method and new tests for mode of inheritance. American Journal of Human Genetics 57: 474-486.

Thomson G. (1995b) Mapping disease genes: family based association studies. American Journal of Human Genetics 57: 487-498.

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

Thomson G, Li H, Dorman SJ, Lie BA, Mignot E, Thorsby E, Steenkiste A, Akey JM, McWeeney S, Single R. (2007a) 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, Valdes AM, Noble JA, Kockum I, Grote MN, Najman J, Erlich HA, Cucca F, Pugliese A, Steenkiste A, Dorman JS, Caillat-Zucman S, Hermann R, Ilonen J, Lambert AP, Bingley PJ, Gillespie KM, Lernmark A, Sanjeevi CB, Ronningen KS, Undlien DE, Thorsby E, Petrone A, Buzzetti R, Koeleman BP, Roep BO, Saruhan-Direskeneli G, Uyar FA, Gunoz H, Gorodezky C, Alaez C, Boehm BO, Mlynarski W, Ikegami H, Berrino M, Fasano ME, Dametto E, Israel S, Brautbar C, Santiago-Cortes A, Frazer de Llado T, She JX, Bugawan TL, Rotter JI, Raffel L, Zeidler A, Leyva-Cobian F, Hawkins BR, Chan SH, Castano L, Pociot F, Nerup J. (2007b) Relative predispositional effects of HLA class II DRB1-DQB1 haplotypes and genotypes on type 1 diabetes: a meta-analysis. Tissue Antigens 70(2): 110-27.

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. 255-294.

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. Pac Symp Biocomput 2010: 359-370.

Tishkoff SA, Kidd KK. (2004) Implications of biogeography of human populations for 'race' and medicine. Nat Genet 36(11 Suppl): S21-27.

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK. (2000) The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67: 518-22.

Valdes AM, McWeeney S, 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): 717-28.

Valdes AM, Thomson G. (1997) Detecting disease-predisposing variants: the haplotype method. Am J Hum Genet 60(3): 703-16.

Valdes AM, McWeeney SK, Meyer D, Nelson MP, Thomson G. 1999. Locus and population specific evolution in HLA class II genes. Annals of Human Genetics 63: 27-43. PMID: 10738519

Vales-Gomez, M., et al. (1998) Kinetics of interaction of HLA-C ligands with natural killer cell inhibitory receptors. Immunity, 9(3): p. 337-44.

Vandiedonck C, Knight JC. (2009) The human Major Histocompatibility Complex as a paradigm in genomics research. Brief Funct Genomic Proteomic 8(5): 379-94.

Wall JD, Pritchard JK. (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nature Review Genetics 4: 587-597.

Weir BS. (1996) Genetic data analysis II. Sinauer Associates, Sunderland, MA.

Williams F, Meenagh A, Single R, McNally M, Kelly P, Nelson MP, Meyer D, Lancaster A, Thomson G, Middleton D. (2004) High resolution HLA-DRB1 identification of a Caucasian population. Human Immunology 65(1): 66-77.

Yasuda, N. (1978) Estimation of haplotype frequency and linkage disequilibrium parameter in the HLA system. Tissue Antigens 12: 315-322.

Yoo YJ, Tang J, Kaslow RA, Zhang K. (2007) Haplotype inference for present-absent genotype data using previously identified haplotypes and haplotype patterns. Bioinformatics 23(18): 2399-406.

Zhang K, Sun F, Zhao H. (2005) HAPLORE: a program for haplotype reconstruction in general pedigrees without recombination. Bioinformatics 1: 90-103.

Zhao H, Pakstis AJ, Kidd JR, Kidd KK. (1999) Assessing linkage disequilibrium in a complex genetic system I. Overall deviation from random association. Annals of Human Genetics 63: 167-179.