Published in final edited form as:

Methods Mol Biol. 2012 ; 882: 245–266. doi:10.1007/978-1-61779-842-9_14.

Analytical Methods for Disease Association Studies with Immunogenetic Data

Jill A. Hollenbach, Steven J. Mack, Glenys Thomson, and Pierre-Antoine Gourraud

Abstract

Disease association studies involving highly polymorphic immunogenetic data may involve analyses at one or many units of analysis, including amino acid, allele, genotype and haplotype levels, as well as consideration of gene–gene or gene–environment interactions. The selection of the appropriate statistical tests is critical and will be dependent on the nature of the dataset (e.g., case-control vs. family data) as well as the specific research hypotheses being tested. This paper describes the various study and analysis categories used for such analyses, including the advantages and limitations of such techniques.

Keywords

HLA; KIR; Immunogenetic; Data analysis; Disease association; Case-control; Family

1. Introduction

Statistics may be defined as “a body of methods for making wise decisions in the face of uncertainty.”

W.A. Wallis

There is no shortage of statistical methods and tools available for the analysis of genetic data; in disease association studies these provide the results upon which we draw inferences about human health, and often influence the direction of future research, clinical protocols, and public health policy. However, just as Wallis suggested statistics assist us in making wise decisions in an uncertain world, it is also imperative that we first *make wise decisions about how we apply statistical methods*. Unless utilized properly, statistics can provide misleading results in disease association studies, consuming precious resources and often leading investigators down blind alleys.

The human leukocyte antigen (HLA) loci serve as a model for the study of genetic variation in human health and disease (1). Their critical role in disease predisposition in a variety of autoimmune, infectious diseases and cancers (2–25), as well as transplant outcomes, has long been recognized. Likewise, a role for the killer cell immunoglobulin-like receptors (KIR) loci and their HLA ligands in autoimmune diseases (26–29) and infectious diseases (HIV and Hepatitis C (30, 31)) as well as solid organ and hematopoietic stem cell transplantation (32–34) and pregnancy (35, 36) is now established. However, the high levels of polymorphism associated with these loci mean that particular care must be taken when analyzing data for disease association studies, in order to avoid confounding results. In this

chapter, we discuss the appropriate application of common statistical methods in the analysis of highly polymorphic immunogenetic data. While most of these analyses can be undertaken with either customized or commercial computer software, no *existing* programs, either individually, or in aggregate, can handle the breadth and complexity of the analyses needed to detect primary and secondary disease genes in polymorphic immunogenetic regions. The complex linkage disequilibrium (LD) patterns common to many immunogenetic data are a major complicating factor in deciphering specific *causal* genetic variants. Further, many analysis programs do not work with the high degree of polymorphism found in some regions, for example HLA, while others only allow a limited number of highly polymorphic loci to be evaluated.

The effects of environmental factors, age of onset and gender-specific effects, undetected heterogeneity of disease, gene–gene and gene–environment interactions further complicate our attempts to detect predisposing (PRE) genes for complex diseases. We thus strongly recommend a multistrategy approach using a variety of complementary methods. The data should be studied carefully to understand differences in significance obtained from different methods. There is no “best practice” for the types of analyses needed for these projects. The power of different methods will vary depending on the specific genetic and environmental features of the disease under consideration; for most complex diseases this involves many unknown factors. Although we must rely heavily on computers, in the final analysis of multiple effects in a genetic region and/or interaction or independent effects between unlinked genes, manipulation of the data by the individual investigator must play a crucial role.

2. Study and Analysis Categories

2.1. Case-Control Studies

Case-control studies are a critical, population-based tool for epidemiological studies, including genetic association studies. These studies generally involve samples of unrelated individuals with the (disease) phenotype of interest and a corresponding control sample of unrelated, unaffected (or randomly ascertained) individuals drawn from the same (ethnic and preferably geographic) population. While a discussion of study design, including power calculations and sample size considerations is out of the scope of this chapter, a general recommendation in immunogenetic studies is to have a minimum of 200 chromosomes (100 individuals) in each of the case and control cohorts, although much larger sample sizes are preferable. Among the advantages of case-control studies is elimination of the necessity to collect data from family members, which is often logistically difficult and costly. However, case-control studies can be very sensitive to population stratification within the sample cohorts, a particularly important issue in immunogenetic data, where allele frequency distributions can vary considerably between human ethnic groups. If the samples are not collected with scrupulous attention to homogeneity of ancestry background, investigators run the risk of misinterpreting genetic difference between cases and controls. In these cases, heterogeneity between cases and controls due to allele frequency differences related to population stratification may be mistaken for association with a particular locus.

Population analyses, including calculations of gene (allele) frequencies, tests for fit to expectations under Hardy–Weinberg Equilibrium (HWE), and estimations of haplotype frequencies, are the first step in any study involving immunogenetic data, and will provide additional means to assess any underlying population substructure, as well as additional units of analysis on which association analyses will be performed. These calculations are detailed in Chap. 13. Many computer software programs are available to perform such calculations on genetic data, but as mentioned in Chap. 13, care must be taken to utilize tools either specifically designed for immunogenetic data (e.g., the PYPOP analytical package (<http://www.pypop.org>)) or those that are able to handle the high levels of polymorphism characteristic of these loci.

2.1.1. Contingency Tables

1. Basic construction and analysis of contingency tables, with particular attention to sparse cells: A contingency table, or cross-tabulation, is used to test the difference, or independence, of frequency distributions for categorical variables. Tests for heterogeneity between specific groups (e.g., cases and controls) for immunogenetic data can be performed using contingency table testing and a standard Chi-square (χ^2) measure. Software implementations of contingency table analyses are available through a variety of commercial packages (e.g., SAS (37), SPSS (38), GraphPad Prism (<http://www.graphpad.com>)); through open source statistical software packages (e.g., the “chisq.test” function in the R language for statistical computing (39)); and various online calculators (e.g., <http://www.physics.csbsju.edu/stats/contingency.html>).

A contingency table is always constructed utilizing raw *counts*, rather than frequency data. The preferred approach in this type of analysis for highly polymorphic immunogenetic data is to first examine heterogeneity in overall allele frequency distributions in cases vs. controls for a particular locus, using a $2 \times k$ row by column ($r \times c$) contingency table. Analyses of these data historically relied on series of 2×2 tests for each allele observed due to the presence of untyped (the so-called blank) alleles; due to the often high frequencies of blanks in initial studies, allele counts and frequencies could not easily be calculated. As described in Chap. 13, the widespread use of DNA typing in most immunogenetic studies now allows for direct ascertainment of allele frequencies.

Analysis based on a series of 2×2 tests requires corrections for multiple comparisons (usually utilizing the Bonferroni method), with a correction factor minimally equivalent to the number of alleles tested. However, when significance of the association of all individual alleles is assessed with the a priori knowledge of *overall* heterogeneity at the locus, it is not necessary to correct for multiple comparisons in subsequent 2×2 tests intended to identify the allele(s) with significant contributions to the overall deviation at the locus. An example of a $2 \times k$ contingency table is given in Table 1 and a related 2×2 table for one allele is given in Table 2.

Rare alleles pose a particular challenge in traditional association tests in disease studies that use the χ^2 test for goodness-of-fit. Although the χ^2 test has been the standard approach for testing fit at the overall locus-level, this test can lead to false acceptance or rejection of the null hypothesis when the expected genotype counts in a contingency table are small (aka,

“sparse cells”), as discussed in Chap. 13 with regard to Hardy–Weinberg (HW) testing; the χ^2 test is *inappropriate* if any expected count is less than one or if the expected count is less than five in more than 20% of all cells in a contingency table. It is not unusual for 30 or more alleles to be observed at the HLA loci for example, with a wide range of frequencies, resulting in many sparse cells. Although Yates’ correction for continuity is sometimes applied to account for numerous sparse cells, this may be overly conservative, and the preferred method is to combine low frequency alleles into a single category, often referred to as the “binned” category (Table 1). A conservative (and recommended) approach is to combine alleles with an expected value of less than five in cases or controls into the binned category prior to calculation of the χ^2 statistic. The χ^2 statistic for a contingency table analysis of case-control data for a genetic association is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},$$

where

O_i = the observed count of allele i

E_i = the expected count of allele i

And the derived values are summed over all cells in the tables. The expected count for each cell in the $r \times c$ table is calculated as

$$\frac{(\text{row total allele } i) \times (\text{column total})}{2n},$$

where

Column total = sum of the counts in the column

Row total = number observations of allele i in all subjects

n = number individuals (cases + controls)

$2n$ = number chromosomes (cases + controls)

The degrees of freedom (df) for the goodness of fit χ^2 analysis are calculated from the number of alleles with expected values in cases and controls of five or greater, plus the combined category, -1 (i.e., $k - 1$). A p -value is obtained by comparing the test statistic to the χ^2 distribution for the appropriate degrees of freedom.

In the example given in Table 1, we obtain a value of $\chi^2 = 39.71$, with $df = 10$. This indicated significant differences between cases and controls at the overall locus level, with $p < 0.001$. In the example in Table 2 for a 2×2 table for an individual allele, $\chi^2 = 11.88$, $df = 1$, $p < 0.001$.

These analyses can be performed at the allele level, as shown in the example, as well as for genotypes and haplotypes. In a haplotype-level analysis, population-level haplotype frequencies are first estimated (see Chap. 13) and the frequency estimates are converted to

counts based on the sample size. Particular care must be taken when working with estimated haplotypes to combine the many rare types that are typically estimated, as well as to not over-interpret the data.

Summary: Basic steps for contingency table analysis in case-control studies with immunogenetic data

- a. Construct a $2 \times k$ table of allele (or genotype or haplotype) counts for cases and controls (see Table 1 example).
- b. Combine all alleles with expected values of four or less in cases or controls into a common “binned” category.
- c. Calculate the χ^2 test statistic for the table and assess significance.
- d. If results are significant at the *overall locus level*, perform additional testing using 2×2 contingency tables for each allele (expected counts of five or more only) against all other alleles.

2. Relative Predispositional Effects (RPE): While contingency table testing will reveal whether the overall allele frequency distributions at a locus differ significantly between cases and controls, identifying all allele(s) that contribute significantly to these differences can be difficult. The RPE method can be used to identify all heterogeneity in disease risk at the primary disease gene; alleles, haplotypes, or genotypes with the strongest predisposing (PRE) or protective (PRO) effects are sequentially removed from the analysis until no further heterogeneity in risk effects is seen. Determining the order in which haplotypes, genotypes or alleles are sequentially removed requires interplay between the contribution to the χ^2 heterogeneity test, the ORs (see below) or Patient/ Control (*P/C*) ratio (40) and the control frequencies of the allele, haplotype, or genotype. In the simplest application of this method, the allele with the greatest contribution to the overall χ^2 values is removed, and the test statistic is recomputed, with degrees of freedom correspondingly reduced by one. If the new χ^2 is found to be statistically significant at the test level, the process is repeated, with each subsequent analysis involving the removal of the greatest contributor to the overall χ^2 from the previous analysis, until the test statistic is no longer found to be significant. In some cases more than one allele can be removed at a time, for example, where there are two PRE alleles with strong effects or a PRE and PRO allele when prior evidence has validated both as significant and strong effects.

The application of the RPE method to HLA data for juvenile idiopathic arthritis (JIA) (41, 42) is summarized in Table 3. The effect of individual alleles is indicated as either predisposing, neutral, or intermediate (INT) or protective. Rare alleles which cannot be placed into one of these categories either by the RPE method or all pairwise comparisons below have no effect given. Significant *p*-values for these comparisons are shaded gray.

The overall analysis of the complete dataset shows considerable heterogeneity in risk ($p < 1.1\text{E} - 27$). DRB1*0801 (PRE), DRB1*1501, and DRB1*0701 (PRO) are the strongest effects (the set of alleles labeled 1 in column 1). Repeat analysis after the removal of these three alleles still gives a highly significant result ($p < 4.1\text{E} - 10$), with DRB1*1104 (PRE)

and DRB1*0401 (PRO) as the strongest effects (set labeled 2 in column 1). Note that a strong argument could be made for deleting these alleles also at the previous round, but this does not alter the outcome. With the removal of these strong effects, there is only minimal evidence of remaining risk heterogeneity ($p < 0.02$), with DRB1*1103 ($p < 0.01$) (PRE) and DRB1*0103 ($p < 0.02$) (PRO) the strongest effects (set labeled 3 in column 1). Note that these latter p -values would not be significant with corrections for multiple comparisons of either the overall tests, or those for individual alleles. Notwithstanding, we mention these results, since our principal aim is to detect heterogeneity that may be relevant to detecting additional disease genes in a genetic region or identifying the causal amino acids in the *primary* disease gene.

3. Conditional Haplotype Method (CHM): With immunogenetic disease associations, the high level of LD between many of the genes means that multiple disease associations are often observed; some of these may indicate a causative genetic factor, and others may be due to LD with this causative factor. In order to account for LD within a genetic region, stratification by the disease-associated alleles using the CHM can be employed (43). We stress that it is important to identify *all* heterogeneity in disease risk at the *primary* disease PRE gene, *including* relatively weak effects, before proceeding with conditional analyses to detect *secondary* genetic effects. Otherwise, spurious results may be seen resulting from heterogeneity in relative risk effects at the primary disease gene at the allele, haplotype, or genotype levels. If all genes directly involved in disease susceptibility have been identified, then the relative frequencies of alleles at the other loci on high-risk haplotypes should be the same in cases and controls; similarly for INT and PRO haplotypes. While fit to these expectations does not exclude the possibility that other genes in the genetic complex are involved in disease, lack of fit unequivocally shows that all disease-PRE genes in the region have *not* been identified.

For the CHM, we denote the primary disease locus by A (or the primary plus secondary loci that have been identified), that is, all putative disease PRE loci in the region, and alleles or haplotypes thereof by A_i , $i = 1, 2, \dots, iA$, and a linked marker locus by B , with alleles B_k , $k = 1, 2, \dots, kB$, then, under the null that the A locus defines all disease predisposition in the region:

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

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

Inequality of these relative frequencies in patients and controls is expected if the allele or SNP or haplotypes thereof under study *does not include all* genes involved in the disease process and in LD with the marker loci. While fit to these expectations does not exclude the possibility that other genes in the region are involved in disease, lack of fit unequivocally shows that all disease-PRE genes in the region have *not* been identified (provided that stratification effects have not produced spurious results).

4. Amino acid-level analyses: In some cases, the analytical challenges posed by rare alleles can be avoided by analyzing the data at the amino-acid level. Immunogenomic studies have generally focused on allele- and locus-level polymorphisms, considering individual amino-acid or nucleotide polymorphisms only as they pertain to specific hypotheses. However, HW, LD, association, and selective neutrality analyses can be performed entirely at the amino-acid level, considering each amino-acid polymorphism independently of the others that constitute an allele sequence. The issues surrounding rare alleles are attenuated in these analyses, as the level of polymorphism at any individual amino-acid position is not as high as at the allele or locus level. However, variation in the number of variants at each amino-position poses challenges for LD estimates and attempts to correlate statistics between positions.

5. Sequence feature variant type analysis: Recently, a novel approach to genetic association analyses has been developed by subdividing genes/proteins into biologically relevant structural (e.g., beta-strand 1) and functional (e.g., peptide binding site) sequence features (SFs), and their “alleles,” called variant types (VTs) (44). In SFVT analyses, association tests are performed on the VTs of these biologically relevant SFs.

2.1.2. Odds Ratios and Relative Risk—Odds ratios (OR) are used in disease association studies as means to describe the strength of association between two variables, and reflect the ratio of odds of an outcome (e.g., disease) occurring in one group (e.g., individuals *with* allele *a*) to the odds of it occurring in another group (individuals *without* allele *a*). Relative risk (RR) is a related but different measure that describes the risk of having the outcome of interest relative to exposure (e.g., positive or negative for allele *a*). In case-control studies, the OR is the preferred measure, due in part to its use in logistic regression (see next section), allowing a more ready comparison of results derived from different methods. Another important advantage to the OR is that it is by its mathematical nature invertible, i.e., the OR for predisposition to disease is the direct inverse of the OR for protection; this is extremely important in terms of immunogenetic studies, where alleles may be associated in either a PRO or PRE role. It is not unusual to find both PRO and PRE alleles in a given disease study for a given locus; the ability to invert ORs allows more direct comparison of the relative strengths of predisposition vs. protection. In contrast, RR is not invertible. RR is more applicable to prospective studies, but for retrospective studies OR is preferred.

1. Calculation of OR and RR: Both OR and RR are calculated for each allele of interest in a 2×2 table. The OR is a simple cross-product of a 2×2 table. A generic example is given in Table 4.

The OR for Table 4 is given by:

$$OR = \frac{ad}{bc}.$$

The RR is given by

$$RR = \frac{a/(a+b)}{c/(c+d)}.$$

For the example given in Table 2, OR = 2.22 and RR = 1.34.

2. Confidence intervals: Values for OR and RR are only meaningful when presented in terms of a corresponding confidence interval (CI). A confidence interval provides a likely range of values for an estimate of a particular parameter within a particular level of significance. The standard probability level is generally set at 95%, meaning that there will be a 95% chance that the value of interest (here, the OR or RR) will fall within that range. For OR and RR, a CI that includes one is not considered statistically significant, as the null hypothesis in this test is OR or RR = 1.

Derivation of the CI is approached in a variety of ways, most of which require the use of computer software, and may include resampling (bootstrap), likelihood or other statistical methods. A full discussion of these methods is not possible here, and their application will ultimately depend on the type of data and analytical methods and software used by the investigator. Most statistical software packages compute basic CIs according to Clopper and Pearson (45), often known as the “exact” method.

2.1.3. Logistic Regression—Analysis using logistic regression can also be used in case-control studies. Logistic regression is a form of *generalized linear modeling* for data with a dichotomous, or binary, outcome variable, such as case-control data where the outcome is either “affected” or “unaffected.” Among its advantages, logistic regression provides a means to develop association models that include the contribution of both quantitative and qualitative covariables, which can involve cumbersome stratification procedures in contingency table testing; logistic regression is a critical tool in the analysis of complex multivariate datasets. Example of potential types of input variables for a logistic regression analysis with immunogenetic data are given in Table 5. Attention must be paid to how both quantitative and qualitative variables are coded, because interpretation of the ORs can greatly vary. For example in Table 5, including number of years smoking as a covariate will provide a single OR associated with each increment of 1 year of smoking and assumes that the risk increment is fully proportional to years smoking. This model may not make sense, so in some cases it is more appropriate to recode quantitative variables as binary (e.g., smoking yes/no in Table 5) or discrete, ordinal (e.g., smoking duration, Table 5), rather than continuous, variables.

A disadvantage of logistic regression analysis is a tendency to overestimate ORs (46) in sample sizes <500, possibly leading to erroneous conclusions. Therefore, as with any statistical method, care must be taken in both the application and interpretation of results in logistic regression analyses. In addition, it is important that the data be reported in a manner that allows readers to reanalyze the data (see below, data reporting).

Logistic regression analysis is always performed using computer software. While many commercial (e.g., SAS, SPSS, STATA) and free (e.g., PLINK (47)) software packages can

be used for logistic regression, many are not suitable for polymorphic immunogenetic data. For example, PLINK performs logistic regression for SNP or copy number variant (CNV) data and will not handle the level of polymorphism typical of HLA or KIR data. HLA or KIR data may be recorded for conditional analysis of SNP or CNV data in PLINK, but this will depend on the specific research question. There are also several packages and standard functions for the R language (e.g., the “glm” function in the base package) that can perform logistic regression, but extreme care must be taken in the coding of immunogenetic data, as many of these functions are not designed to handle high levels of polymorphism.

A regression analysis produces a model that includes all of the variables that are useful in predicting the (dichotomous) outcome variable. Logistic regression provides an OR for each variable involved in predicting outcome, and can be particularly useful in immunogenetic studies that must account for a number of cofactors in assessing associations. Several options are available in building the most parsimonious model, i.e., the one explaining the maximum of the variance with the minimal number of variables. A backward stepwise approach can be particularly productive in an exploratory analysis; independent variables are removed from the full model and the fit of the new model is tested at each step. This procedure allows the investigator to eliminate noncontributing variables and build a more parsimonious model. A forward stepwise approach can be more robust when large numbers of variables are studied. In this procedure, the most significant variables are successively added to an empty model, until the addition of variables does not contribute to significantly increase the variance explained by the model.

2.2. Family-Based Studies

Family studies have long been the bread and butter of genetic epidemiology. Family studies provide a means to assess genetic linkage and/or association by testing within a nuclear family.

2.2.1. Pedigree Types in Family Studies—The simplest type of family study involves (1) “trios,” sets of “simplex” families consisting of a proband (an individual with the disease phenotype of interest) and both of the proband’s parents (Fig. 1a). Additional family collections may include (2) multiplex sib pairs (MSP) pedigrees with two or more affected siblings (Fig. 1c), (3) multiplex parent child (MPC), with one affected parent and child (Fig. 1b) and (4) multiplex parent sibs (MPS) with at least one affected parent and two or more affected siblings (Fig. 1d). The type of analyses available will depend on the nature of the pedigrees collected.

2.2.2. Affected Sib Pair Method—When considering related individuals, the loci linked to disease susceptibility are expected to be found in the region of the genome inherited from the same ancestors. Affected sib pair methods can be used in the study of HLA-associated diseases. Deviations from the Mendelian random expectations of 25, 50, and 25% that two affected sibs will on average share 2, 1, and 0 parental chromosomes in common identical by descent (IBD) suggest a disease PRE gene in the region. Several statistics can be used, a goodness-of-fit χ^2 is commonly used and tests the null hypothesis of the absence of linkage with the disease locus.

2.2.3. Transmission Disequilibrium Test and Affected Family-Based Controls

—Originally developed as a test for genetic linkage, the Transmission Disequilibrium Test (TDT) is applicable as an association test as well, since it always tests a compound hypothesis of linkage and association: only linked markers that are associated with disease can give a significant result. The fundamental basis for the TDT involves the likelihood that if a marker is linked to disease, there will be a bias in transmission from parents to the affected individual. In the TDT test, only parents who are heterozygous for the marker of interest are considered. The TDT test has the advantage of being robust with respect to population structure, thereby sidestepping the possible pitfall of population stratification often encountered in case-control studies. The TDT test looks at deviations from the expected 50:50 ratio of transmitted and nontransmitted alleles from parents heterozygous for the marker alleles.

The use of nuclear family data (two parents and their children) to estimate control marker allele frequencies (and similarly haplotype frequencies) was introduced, with application to HLA data, by Rubinstein et al. (48), and Falk and Rubinstein (49). In families ascertained for the presence of at least one affected child, termed S (simplex) pedigrees (also referred to as *trio* families), the two parental marker alleles *not* transmitted to the affected child (the proband) are used as population (control) alleles (see Fig. 1a). These are referred to as affected family-based controls (AFBACs). This matched design for patient (parental transmitted) and “control” (parental nontransmitted) marker alleles avoids ethnic confounding in the case of a stratified population (50–52).

Field (53) and Thomson (54)—extended this AFBAC approach to nuclear pedigrees ascertained for the presence of at least two affected sibs, termed MSP pedigrees (although referred to then as MS pedigrees, for multiplex sibs); using the alleles from both parents that are *never* transmitted to either sib in the affected sib pair as the “control” population (Fig. 1c). Additionally, MPC and MPS ascertainment schemes were considered: these extend the S and MSP ascertainment schemes to also include at least one affected parent. The non- or never-transmitted alleles, respectively, from the *unaffected* parent form the respective AFBAC populations (Fig. 1b, d). Similar extensions apply to other ascertainment schemes, e.g., with ascertainment based on the presence of at least three affected sibs in a nuclear family, the AFBACs are the parental alleles *never* transmitted to any of the three affected sibs.

AFBAC is similar to the TDT in that it examines the transmission of marker alleles in family trios. The AFBAC method has often mistakenly been equated to a heterogeneity test of association of patient and AFBAC frequencies. The AFBAC method per se is for obtaining an unbiased estimate of marker allele frequencies, and does not refer to any specific test of association/linkage. The TDT is the appropriate statistical test to apply. However, note that in the case of a homogenous population sample, the results from a TDT and contingency table analysis will be very similar. Here, all nontransmitted parental alleles are considered, regardless of whether the parents are homozygous or heterozygous.

Statistical tests to apply to the AFBAC data (alleles or haplotypes) are as follows (55):

- a. Equality of AFBAC maternal and paternal frequencies (a significant difference may imply a maternal/fetal interaction effect, in which case only the paternal AFBACs should be used).
- b. Test for equality of AFBAC frequencies with control marker frequencies available from any other sources, e.g., population data, other disease studies, and AFBACs ascertained under different criteria. With case/control data, one must be careful to understand the ascertainment scheme for the controls, i.e., whether it is from a random *or* nondiseased population. For rare diseases, there will be little difference, but for common diseases this is not the case.
- c. Test for multiplicative structure of the paternal transmitted and nontransmitted alleles for S ascertainment, with appropriate modifications for other ascertainment schemes, and similarly for the maternal alleles. A significant difference in the latter case may reflect a maternal/fetal interaction effect.
- d. Test for a multiplicative structure of the paternal nontransmitted alleles and the child's genotype for S ascertainment, with appropriate modifications for other ascertainment schemes, and similarly for the maternal alleles. A significant difference in the latter case may reflect a maternal/fetal interaction effect.
- e. Test for fit to HWP of the AFBAC genotypes created, with S ascertainment, from the maternal and paternal nontransmitted alleles in each family. For MS ascertainment, only the families where both sibs share the same two parental alleles (AC) (Fig. 1c) are used in this test. A significant deviation from HWP may reflect a maternal/fetal interaction effect.

3. KIR Data: Special Considerations

While KIR data are in many ways typical of immunogenetic data, with high levels of polymorphism and linkage disequilibrium through the region, as well as a clear history of extensive recombination and gene conversion events, the limitations of most typing techniques mean that in many cases investigators will have to deal with large amounts of missing data. Likewise, most typing techniques cannot detect or distinguish between loci that may be duplicated on a KIR haplotype. In addition, the relationship between the KIR and their HLA ligands dictates consideration of interaction effects between these two important polymorphic regions when analyzing for disease associations. Table 6 illustrates potential units of analysis in disease association studies with KIR data. This list is by no means exhaustive, and the ultimate choice of analyses performed should be determined by the specific research hypotheses being tested in the study.

Issues specific to the analysis of KIR data for disease association studies include:

1. *Treatment of presence/absence or gene-content typing data.* As allele level KIR genotyping improves, datasets will contain a mixture of gene-content and allele level typing data. Analysis of KIR data at the gene-content level may at times be more robust than at the allele level, or vice-versa. In gene-content data analyses, allele frequencies must be estimated using expected frequencies under Hardy–Weinberg Equilibrium Proportions (HWEP) or phenotype frequencies. Allele- and

locus-level experimental data can be compared to published allele-level KIR typing datasets.

2. *Incorporation of HLA interaction effects.* It is not clear how best to model the KIR interaction with its HLA ligand; analyses could define HLA at a broad functional level or use molecular (HLA allele) designations. Current research depends almost exclusively on broad designations, but recent work suggests some crossover in HLA-C-group KIR ligand specificity. We will assess the impact of HLA-ligand definition using both broad designations and allelic designations for HLA ligands using published data and model.
3. *Controlling for other HLA associations.* Many disease associations with KIR also show strong HLA associations (e.g., autoimmune disorders). Accounting for primary HLA associations while accounting for HLA's function as a KIR ligand involves consideration of the extensive linkage disequilibrium throughout the HLA region.
4. *The use of estimated haplotypes vs. the standard A/A and B/x designations.*
5. *The treatment of haplotypic data.* It is unclear how chromosomal location should be considered for otherwise identical loci (e.g., KIR2DS3), and there are no standards for (locus level) CNV.

4. Data Reporting

A basic premise in scientific research involves the necessity of replication of studies. Although this is usually taken to mean replication of the laboratory findings, it is important that data analyses be replicated as well. In order to facilitate replication of analytical outcomes, it is imperative that there be a careful and thorough reporting of data. For immunogenetic studies, the data should be listed as both raw counts and frequencies; this will allow other investigators to perform additional analyses (using counts) as well as meta-analysis. We recommend the authors provide either within the body of the paper or as supplemental material, a list of all alleles, haplotypes and genotypes, including rare variants.

Acknowledgments

This work was supported by National Institutes of Health (NIH) grants U01AI067068 (JAH, SJM) and U19 AI067152 (PAG) awarded by the National Institute of Allergy and Infectious Diseases (NIAID) and 1R01DK061722 (JAH) awarded by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Glossary

Bonferroni correction

The most commonly used method of correcting for multiple comparisons. Generally, the test significance level is divided by the number of comparisons made in the study, thereby increasing the overall stringency of the significance testing.

Confidence interval	A likely range of values for an estimate of a particular parameter within a particular level of significance.
Contingency table	Also known as a <i>cross-tabulation</i> . A $2 \times n$ table used to analyze heterogeneity between two sets of observations of two or more categorical variables.
Genetic association	The occurrence within a population, greater than that expected by chance, of a genetic trait with a particular phenotype.
Logistic regression	A statistical method to determine which in a set of independent variables has a predictive relationship to a binary-dependent outcome variable.
Multiple comparisons	Also known as <i>multiple testing</i> . Performing a statistical test multiple times in the same analysis, thereby increasing the number of chances that the null hypothesis will be incorrectly rejected, leading to false positive associations.
Odds ratio	The ratio of odds of an outcome occurring in one group to the odds of it occurring in another group.
Population stratification	Also referred to as <i>population substructure</i> . Allele frequency differences between subpopulations within a study population due to ancestry differences or selection biases.
Relative risk	A measure that describes the risk of having the outcome of interest relative to exposure.
Yates correction for continuity	An adjustment to the χ^2 test statistic performed by subtracting 0.5 from the (O-E) value for each cell in a contingency table. The purpose of this correction is to account for sparse cells in the table which may introduce discontinuity with regard to the χ^2 distribution.

References

1. Vandiedonck C, Knight JC. The human major histocompatibility complex as a paradigm in genomics research. *Brief Funct Genomic Proteomic*. 2009; 8(5):379–394. [PubMed: 19468039]
2. Mather, KA.; President, M.; Eastel, S.; Hollenbach, J.; Klitz, W.; Huttley, G.; Thomson, G. Long distance linkage disequilibrium between HLA microsatellites in two human populations. In: Hansen, JA., editor. *Immunobiology of the human MHC: proceedings of the 13th international histocompatibility workshop and congress*; Seattle: IHWG Press; 2007.
3. Ann, B.; Begovich, WK.; Steiner, LL.; Grams, S.; Suraj-Baker, V.; Hollenbach, J.; Trachtenberg, E.; Louie, L.; Zimmerman, PA.; Hill, AVS.; Stoneking, M.; Sasazuki, T.; Rickards, O.; Titanji, VPK.; Konenkov, VI.; Sartakova, ML. HLA DQ haplotypes in 15 different populations. In: Kasahara, M., editor. *Major histocompatibility complex: evolution, structure and function*. Tokyo: Springer; 1999.
4. Petzel-Erler, ML.; Gorodezky, C.; Layrisse, Z.; Klitz, W.; Fainboim, L.; Vullo, C.; Bodmer, J.; Egea, E.; Navarrete, C.; Infante, E.; Alaez, C.; Olivo, A.; Debaz, H.; Bautista, N.; de la Rosa, G.; Vazquez, MN.; Navarro, JL.; Pujol, MJ.; Duran, C.; Schafhauser, C.; Faucz, FR.; Janzen, M.; Maciag, P.; Boldt, ABW.; Souza, PSA.; Probst, CM.; da Silva, GF.; Makhatadze, N.; Dominguez, E.; Montagnani, S.; Matos, M.; Martinez, A.; Herrera, F.; Hollenbach, J.; Thomson, G.; Pando, M.;

- Satz, L.; Larriba, J.; Fernandez, G.; Pesoa, SA.; Borosky, A.; Garavito, G.; Angel, L.; Brown, J.; Llop, E. Anthropology component report for region Latin-America: Amerindian and admixed populations. In: Charron, D., editor. Genetic diversity of HLA: functional and medical implications. Paris: EDK Press; 1997.
5. Trachtenberg, EA.; Hayes, E.; Hollenbach, JA.; Keyeaux, G.; Bernal, J.; Klitz, W. HLA class II variation and linkage disequilibrium in nine Amerindian and three African American tribes from Columbia. Results of expedicion humana. In: Charron, D., editor. Genetic diversity of HLA: functional and medical implications. Paris: EDK Press; 1997.
 6. Trachtenberg E, Vinson M, Hayes E, Hsu YM, Houtchens K, Erlich H, Klitz W, Hsia Y, Hollenbach J. HLA class I (A, B, C) and class II (DRB1, DQA1, DQB1, DPB1) alleles and haplotypes in the Han from southern China. *Tissue Antigens*. 2007; 70:455–463. [PubMed: 17900288]
 7. Begovich AB, Moonsamy PV, Mack SJ, Barcellos LF, Steiner LL, Grams S, Suraj-Baker V, Hollenbach J, Trachtenberg E, Louie L, Zimmerman P, Hill AV, Stoneking M, Sasazuki T, Koenkov VI, Sartakova ML, Titanji VP, Rickards O, Klitz W. Genetic variability and linkage disequilibrium within the HLA-DP region: analysis of 15 different populations. *Tissue Antigens*. 2001; 57:424–439. [PubMed: 11556967]
 8. Bergstrom TF, Engkvist H, Erlandsson R, Josefsson A, Mack SJ, Erlich HA, Gyllensten U. Tracing the origin of HLA-DRB1 alleles by microsatellite polymorphism. *Am J Hum Genet*. 1999; 64:1709–1718. [PubMed: 10330359]
 9. Bergstrom, TF.; Mack, SJ.; Gyllensten, U.; Ehrlich, HA. Evolution of HLA-DRB loci, DRB1 lineages, and alleles: analyses of intron-1 and-2 sequences. In: Kasahara, M., editor. Major histocompatibility complex: evolution, structure and function. Tokyo: Springer; 2000.
 10. Crawford MH, Reddy BM, Martinez-Laso J, Mack SJ, Erlich HA. Genetic variation among the Golla pastoral caste subdivisions of Andhra Pradesh, India, according to the HLA system. *Hum Immunol*. 2001; 62:1031–1041. [PubMed: 11543904]
 11. Ellis JM, Mack SJ, Leke RF, Quakyi I, Johnson AH, Hurley CK. Diversity is demonstrated in class I HLA-A and HLA-B alleles in Cameroon, Africa: description of HLA-A* 03012, *2612, *3006 and HLA-B*1403, *4016, *4703. *Tissue Antigens*. 2000; 56:291–302. [PubMed: 11098929]
 12. Erlich HA, Mack SJ, Bergstrom T, Gyllensten UB. HLA class II alleles in Amerindian populations: implications for the evolution of HLA polymorphism and the colonization of the Americas. *Hereditas*. 1997; 127:19–24. [PubMed: 9420466]
 13. Hildesheim A, Apple RJ, Chen CJ, Wang SS, Cheng YJ, Klitz W, Mack SJ, Chen IH, Hsu MM, Yang CS, Brinton LA, Levine PH, Erlich HA. Association of HLA class I and II alleles and extended haplotypes with nasopharyngeal carcinoma in Taiwan. *J Natl Cancer Inst*. 2002; 94:1780–1789. [PubMed: 12464650]
 14. Johansson A, Ingman M, Mack SJ, Erlich H, Gyllensten U. Genetic origin of the Swedish Sami inferred from HLA class I and class II allele frequencies. *Eur J Hum Genet*. 2008; 16:1341–1349. [PubMed: 18478041]
 15. Leffell MS, Fallin MD, Erlich HA, Fernandez-Vijna M, Hildebrand WH, Mack SJ, Zachary AA. HLA antigens, alleles and haplotypes among the Yup'ik Alaska natives: report of the ASHI Minority Workshops, Part II. *Hum Immunol*. 2002; 63:614–625. [PubMed: 12072196]
 16. Mack SJ, Bugawan TL, Moonsamy PV, Erlich JA, Trachtenberg EA, Paik YK, Begovich AB, Saha N, Beck HP, Stoneking M, Erlich HA. Evolution of Pacific/Asian populations inferred from HLA class II allele frequency distributions. *Tissue Antigens*. 2000; 55:383–400. [PubMed: 10885559]
 17. Mack SJ, Erlich HA. HLA class II polymorphism in the Ticuna of Brazil: evolutionary implications of the DRB1*0807 allele. *Tissue Antigens*. 1998; 51:41–50. [PubMed: 9459502]
 18. Mack SJ, Tu B, Lazaro A, Yang R, Lancaster AK, Cao K, Ng J, Hurley CK. HLA-A, -B, -C, and -DRB1 allele and haplotype frequencies distinguish Eastern European Americans from the general European American population. *Tissue Antigens*. 2009; 73:17–32. [PubMed: 19000140]
 19. MacKenzie J, Perry J, Ford AM, Jarrett RF, Greaves M. JC and BK virus sequences are not detectable in leukaemic samples from children with common acute lymphoblastic leukaemia. *Br J Cancer*. 1999; 81:898–899. [PubMed: 10555765]

20. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. Signatures of demographic history and natural selection in the human major histocompatibility complex loci. *Genetics*. 2006; 173:2121–2142. [PubMed: 16702436]
21. Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, Thomson G. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum Immunol*. 2008; 69:443–464. [PubMed: 18638659]
22. Tang TF, Hou L, Chen M, Belle I, Mack S, Lancaster A, Ho GY, Hwang WY, Alsagoff F, Ng J, Hurley CK. HLA haplotypes in Singapore: a study of mothers and their cord blood units. *Hum Immunol*. 2007; 68:430–438. [PubMed: 17462511]
23. Tu B, Mack SJ, Lazaro A, Lancaster A, Thomson G, Cao K, Chen M, Ling G, Hartzman R, Ng J, Hurley CK. HLA-A, -B, -C, -DRB1 allele and haplotype frequencies in an African American population. *Tissue Antigens*. 2007; 69:73–85. [PubMed: 17212710]
24. Cao K, Hollenbach J, Shi X, Shi W, Chopek M, Fernandez-Vina MA. Analysis of the frequencies of HLA-A, B, C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. *Hum Immunol*. 2001; 62:1009–1030. [PubMed: 11543903]
25. Hollenbach JA, Thomson G, Cao K, Fernandez-Vina M, Erlich HA, Bugawan TL, Winkler C, Winter M, Klitz W. HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives. *Hum Immunol*. 2001; 62:378–390. [PubMed: 11295471]
26. Khakoo SI, Chloe LT, Martin MP, Brooks CR, Gao X, Astemborski J, Cheng J, Goedert JJ, Vlahov D, Hilgartner M, Cox S, Little AM, Alexander GJ, Cramp ME, O'Brien S, Rosenberg WMC, Thomas DL, Carrington M. HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. *Science*. 2004; 305:872–874. [PubMed: 15297676]
27. Williams AP, Bateman AR, Khakoo SI. Hanging in the balance. KIR and their role in disease. *Mol Interv*. 2005; 5:226–240. [PubMed: 16123537]
28. Li Y, Zhang T, Ho C, Orange JS, Douglas SD, Ho WZ. Natural killer cells inhibit hepatitis C virus expression. *J Leukoc Biol*. 2004; 76:1171–1179. [PubMed: 15339939]
29. Hollenbach JA, Ladner MB, Saetern K, Taylor KD, Mei L, Haritunians T, McGovern DPB, Erlich HA, Rotter JI, Trachtenberg EA. Susceptibility to Crohn's disease is mediated by KIR2DL2/KIR2DL3 heterozygosity and the HLA-C ligand. *Immunogenetics*. 2009; 61(10):663–671. [PubMed: 19789864]
30. Gaudieri S, Desantis D, McKinnon E, Moore C, Nolan D, Witt CS, Mallal SA, Christiansen FT. Killer immunoglobulin-like receptors and HLA act both independently and synergistically to modify HIV disease progression. *Genes Immun*. 2005; 6:683–690. [PubMed: 16121209]
31. Khakoo SI, Carrington M. KIR and disease: a model system or system of models? *Immunol Rev*. 2006; 214:186–201. [PubMed: 17100885]
32. Kunert K, Seiler M, Mashreghi MF, Klippert K, Schonemann C, Neumann K, Pratschke J, Reinke P, Volk HD, Kotsch K. KIR/HLA ligand incompatibility in kidney transplantation. *Transplantation*. 2007; 84:1527–1533. [PubMed: 18091530]
33. Gedil MA, Steiner NK, Hurley CK. KIR3DL2: diversity in a hematopoietic stem cell transplant population. *Tissue Antigens*. 2007; 70:228–232. [PubMed: 17661911]
34. Sun JY, Gaidulis L, Dagis A, Palmer J, Rodriguez R, Miller MM, Forman SJ, Senitzer D. Killer Ig-like receptor (KIR) compatibility plays a role in the prevalence of acute GVHD in unrelated hematopoietic cell transplants for AML. *Bone Marrow Transplant*. 2005; 36:525–530. [PubMed: 16025153]
35. Lanier LL. Natural killer cells fertile with receptors for HLA-G? *Proc Natl Acad Sci USA*. 1999; 96:5343–5345. [PubMed: 10318882]
36. Moffett A, Hiby SE. How does the maternal immune system contribute to the development of pre-eclampsia? *Placenta*. 2007; 28(suppl A):S51–S56. [PubMed: 17292469]
37. Hu LP, Tao LX. (Statistical analysis using SAS software package for data of RxC contingency table (part two)). *Zhong Xi Yi Jie He Xue Bao*. 2009; 7:878–882. [PubMed: 19747446]
38. Horstmann GA. Product of the month: SPSS statistical package for the sciences. *Radiologe*. 1988; 28:438–439. [PubMed: 3186979]

39. R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2008.
40. Thomson G, Valdes AM, Noble JA, Kockum I, Grote MN, Najman J, Erlich HA, et al. Relative predispositional effects of HLA class II DRB1-DQB1 haplotypes and genotypes on type 1 diabetes: a meta-analysis. *Tissue Antigens*. 2007; 70:110–127. [PubMed: 17610416]
41. Hollenbach JA, Thompson SD, Bugawan TL, Ryan M, Sudman M, Marion M, Langefeld CD, Thomson G, Erlich HA, Glass DN. Juvenile idiopathic arthritis and HLA class I and class II interactions and age-at-onset effects. *Arthritis Rheum*. 2010; 62:1781–1791. [PubMed: 20191588]
42. Thomson G, Marthandan N, Hollenbach JA, Mack SJ, Erlich HA, Single RM, Waller MJ, Marsh SG, Guidry PA, Karp DR, Scheuermann RH, Thompson SD, Glass DN, Helmberg W. Sequence feature variant type (SFVT) analysis of the HLA genetic association in juvenile idiopathic arthritis. *Pac Symp Biocomput*. 2010:359–370. [PubMed: 19908388]
43. Thomson G, Valdes AM. Conditional genotype analysis: detecting secondary disease loci in linkage disequilibrium with a primary disease locus. *BMC Proc*. 2007; 1(suppl 1):S163. [PubMed: 18466509]
44. 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 DB, Helmberg W. Sequence feature variant type (SFVT) analysis of the HLA genetic association in juvenile idiopathic arthritis. *Proceedings of the 2010 pacific society of biocomputing conference*. *Pac Symp Biocomput*. 2009:359–370.
45. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934; 26:404–413.
46. Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. *BMC Med Res Methodol*. 2009; 9:56. [PubMed: 19635144]
47. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007; 81:559–575. [PubMed: 17701901]
48. Rubinstein P, Walker M, Carpenter C, Carrier C, Krassner J, Falk C, Ginsberg F. Genetics of HLA disease associations: the use of the haplotype relative risk (HRR) and the “haplo-delta” (Dh) estimates in juvenile diabetes from three racial groups. *Hum Immunol*. 1981; 3:384.
49. Falk CT, Rubinstein P. Haplotype relative risk: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet*. 1987; 51:227–233. [PubMed: 3500674]
50. Khoury MJ, Beaty TH. Applications of the case-control method in genetic epidemiology. *Epidemiol Rev*. 1994; 16:134–150. [PubMed: 7925722]
51. Schaid DJ, Sommer SS. Comparison of statistics for candidate-gene association studies using cases and parents. *Am J Hum Genet*. 1994; 55:402–409. [PubMed: 8037216]
52. Thomson G. Mapping disease genes: family-based association studies. *Am J Hum Genet*. 1995; 57:487–498. [PubMed: 7668275]
53. Field LL. Non-HLA region genes in insulin dependent diabetes mellitus. *Baillieres Clin Endocrinol Metab*. 1991; 5:413–438. [PubMed: 1909861]
54. Thomson G. AGFAP method: applicability under different ascertainment schemes and a parental contributions test. *Genet Epidemiol*. 1993; 10:289–310. [PubMed: 8224808]
55. Thomson G. Analysis of complex human genetic traits: an ordered-notation method and new tests for mode of inheritance. *Am J Hum Genet*. 1995; 57:474–486. [PubMed: 7668274]

Fig. 1.

The four family-based ascertainment schemes are selected in the presence of at least: (a) one affected child (S), (b) one affected child and one affected parent (multiplex parent child (MPC)), (c) two affected sibs (multiplex sib pairs (MSP)), and (d) two affected sibs and one affected parent (multiplex parent sibs (MPS)). As noted in the text, additional family members (parent or sib) may be affected within each of these schemes; they occur with frequencies expected under the disease model and are retained in analyses. For the MSP and MPS ascertainment schemes, using the ordered notation of Thomson (55) the first affected

sib ascertained is always denoted by genotype AC, with the four possible genotypes for the second affected sib listed.

Table 1

A 2×k contingency table for case-control data

Allele	Case (count)	Control (count)	Row total
0101	75	29	104
0102	5	5	10
0301	59	56	115
0401	6	15	21
0701	9	25	34
0701	6	11	17
0801	73	39	112
1101	39	35	74
1103	9	4	13
1104	40	24	64
Binned	33	25	57
Total	354	268	622

Table 2

A 2×2 contingency table for allele 0101

Allele	Case (count)	Control (count)	Row total
0101	75	29	104
Others (<i>not</i> 0101)	279	239	517
<i>Total</i>	<i>354</i>	<i>268</i>	<i>622</i>

Table 3

JIA-OP HLA-DRB1 allele data ranked by odds ratio (OR)

RPE ^a	A ^b	B ^c	DRB1	Patient	Control	Chi-square (χ^2)	p-Value ^d	Effect	OR	C ^e	C ^e
	Ix		*1103	12	1	6.80	0.009	PRE	9.40	1.22	72.49
1	I	Ix	*0801	102	13	48.61	3.1E-12	PRE	6.90	3.83	12.43
2	I	Ix	*1104	57	11	20.71	5.3E-06	PRE	4.26	2.21	8.20
			*0403	9	3	1.68	0.20		2.33	0.63	8.65
3	II	Ix	*1301	90	38	9.99	0.002	INT	1.95	1.31	2.90
			*0102	9	5	0.35	0.55	INT	1.39	0.46	4.18
	II	Ix	*1101	60	36	1.42	0.23	INT	1.31	0.85	2.02
			*0901	9	6	0.08	0.78	INT	1.16	0.41	3.28
	II	Ix	*0101	74	50	0.52	0.47	INT	1.16	0.79	1.69
	II	Ix	*0301	89	61	0.50	0.48	INT	1.14	0.81	1.62
			*1201	10	8	0.006	0.94	INT	0.96	0.38	2.46
	II	Ix	*1302	28	23	0.05	0.82		0.94	0.53	1.64
			*1303	10	9	0.11	0.74		0.86	0.34	2.12
			Binned ^f	27	27	0.92	0.34		0.76	0.44	1.31
			*1601	6	8	1.05	0.30		0.58	0.20	1.67
			*1401	11	18	4.05	0.04		0.46	0.22	0.99
			*1502	5	10	3.26	0.07		0.38	0.13	1.12
	III	Ix	*0404	7	16	6.34	0.01	PRO	0.33	0.14	0.81
1	III	Ix	*1501	38	80	28.24	1.1E-07	PRO	0.33	0.22	0.49
1	III	Ix	*0701	30	65	23.92	1.0E-06	PRO	0.33	0.21	0.51
2	III	Ix	*0401	21	47	18.10	2.1E-05	PRO	0.33	0.19	0.55
3		Ix	*0103	4	11	5.42	0.02	PRO	0.28	0.09	0.87
			Total	708	546	182.1	1.1E-27	PRO			

^aNumbers denote the order of removal due to largest effect(s) in the relative predispositional effect (RPE) analysis

^bSet A: Based on pairwise allele comparisons, the common alleles are divided into mutually exclusive, and significantly different, predisposing (PRE) (I), neutral (intermediate (INT)) (II), and protective (PRO) (III) categories for use later in amino acid comparisons (described below)

^cSet B: The sets I, II, and III above are expanded (indicated by Ix, IIX, and IIIX) to include rare alleles, while excluding those alleles which do not clearly fall into one of the three risk categories

^dThe individual p -values are biased as the assumption of a 1 df χ^2 is incorrect, and conservative; the p -values can be used however for a relative ranking of the allelic effects

^eThe upper and lower 95% confidence intervals (CIs) for the odds ratio (OR) are given (see below)

^fThe binned category consists of all alleles with an expected value <5 under the χ^2 test of heterogeneity of patient and control allele counts

Table 4Generic 2×2 table

Allele	Case (count)	Control (count)
Y	a	b
Not Y	c	d

Table 5

Example of input variables for logistic regression analysis

Independent variable (HLA or KIR) ^{a,b}	Covariable				
	Dependent variable (disease status) ^c	Risk factor: smoking (yes/no)-Binary ^d	Risk factor: smoking (years)-Quantitative ^e	Risk factor: smoking (duration)-Ordinal ^f	Genetic: <i>snp</i> -Categorical ^g
0101,0401	0	1	15.8	1	1
0401,0501	0	1	10	2	2
0301,0401	1	1	2.5	3	3
0101,0501	1	0	1	2	4

^aData can be analyzed at the allele, genotype, haplotype, or other user-defined level

^bUser will need to specify the genetic model, e.g., recessive, codominant, etc.

^cBy convention, 0 = control; 1 = case

^dHere, risk factor (smoking) is analyzed as present or absent. 0 = nonsmoker, 1 = smoker

^eHere, risk factor (smoking) is analyzed as a continuous quantitative variable, number of years smoked

^fHere, risk factor (smoking) is analyzed as an ordered discrete variable, number of years smoked categorized 1 = 0–5 years; 2 = 6–10 years; 3 = 11–20 years; 4 = more than 20 years

^gHere, SNP data categorized by genotype, e.g., 1 = AG; 2 = AT; 3 = GG; 4 = GT

Table 6

Potential units of analysis for killer cell immunoglobulin-like receptors (KIR) data in disease association

Analytical unit	Examples
KIR ligands	HLA-C1 carrier; HLA-C1 homozygous; HLA-C2 carrier; HLA-C2 homozygous; HLA-Bw4 carrier; HLA-Bw4 homozygous; HLA-Bw6 carrier; HLA-Bw6 homozygous; HLA-Bw4_Thr80 carrier; HLA-Bw4_Ile80 carrier
KIR genes	KIR2DL1; KIR2DP1; KIR2DL2; KIR2DL3; KIR2DL5; KIR3DL1; KIR2DS1; KIR2DS2; KIR2DS3; KIR2DS4; KIR2DS5; KIR3DS1
KIR diplotypes	AA; AA (with no activating); AB; BB; centromeric AA; telomeric AA; centromeric AB; telomeric AB; centromeric BB; telomeric BB
KIR haplotypes	A; B; B variants
KIR and ligands	KIR2DL1 with HLA-C2; KIR2DL1 without HLA-C2; KIR2DL2 with HLA-C1; KIR2DL2 without HLA-C1; KIR2DL3 with HLA-C1; KIR2DL3 without HLA-C1; KIR3DL1 with HLA-Bw4; KIR3DL1 with HLA-Bw4_Ile80; KIR3DL1 with HLA-Bw4 homozygous; KIR2DL3 with HLA-C1; KIR3DS1 with HLA-Bw4; KIR3DS1 with HLA-Bw4_Ile80; KIR3DS1 with HLA-Bw4 homozygous