Online Supplement for:

SARCOIDOSIS SUSCEPTIBILITY AND RESISTANCE HLA-DQB1 ALLELES IN AFRICAN AMERICANS

METHODS

Introduction to Family-based Association Tests

The idea of family-based association testing originated with haplotype relative risk proposed by Falk and Rubinstein (1), and was later improved upon with the transmission disequilibrium test (TDT) (2). The TDT as it was first proposed was restricted to testing a single allele in nuclear families of two parents (where at least one parent was homozygous for the marker being tested) and one affected offspring. Various extensions of the original TDT formulation have included methods that allow for missing parents (3, 4), multiallelic markers (5), and inclusion of covariates (6) and unaffected sibs (7). Although these extensions have been useful, individually they generally only addressed a single issue and could not incorporate aspects of other methodological extentions of the TDT. A recently proposed unified approach to family-based association testing (8) brings together most of the afore-mentioned methodological extensions of the TDT into one statistic that is described below.

The Unified Family-based Association Statistic

To determine whether one or more alleles at the locus of interest was associated with the sarcoidosis phenotype, we used the family-based association test statistic (8), *S*, calculated using the FBAT software (9). The *S* statistic is a summation of the following parameters over *i* offspring in *j* families,

$S = \sum_{ij} T_{ij} X_{ij}$.

In the equation above, X_{ij} is defined as a function of the genotype of the ij^{th} offspring at the locus being tested and T_{ij} is a function of the phenotype of interest. X_{ij} was parameterized as a count of *A* alleles (where *A* represents a specific allele) with possible values 0, 1, or 2. T_{ij} was parameterized as a 0,1 variable where 0 represented unaffected status and 1 represented affected status.

The expected value of *S (E(S))*, was based on the genetic model postulated and the expected value of X_{ij} under that model. Computation of the variance of *S*, (*V*), allowed estimation of a significance level for a large number of families based on a Z statistic normally distributed with mean 0 and variance 1. For each allele, the Z statistic takes on the form:

$$Z = [S - E(S)] / \sqrt{V}$$

Once the distribution of *S* over all *k* alleles is known (i.e., S_k), an overall score, U, can be defined as the following,

$$U = \sum_k \{S_k - E(S_k)\}$$

This score formed a multiallelic test statistic with a chi-squared distribution where the number of degrees of freedom are equal to the rank of the variance matrix, V, i.e.,

$$\chi^2 = U^T V^- U.$$

Gene– Environment Interaction Modeling

Gene–environment (GxE) interactions were evaluated using a general estimating equations analytic approach (10). The general model that was fit is shown below:

$$Y_{ij} = \beta_o + \beta_1 g_{ij} + \beta_2 e_{ij} + \beta_3 g_{ij} e_{ij}$$

In this model, the dependent term, Y_{ij}, represents the sarcoidosis history status (1 = positive history; 0 = negative history) of the *i*th sib in the *j*th sibship. On the right-hand side of the

equation, g_{ij} and e_{ij} represent the terms for the genetic and environmental factors, respectively. For an additive model, g_{ij} was coded as 0, 1, or 2 dependent upon the number of copies of the allele of interest found in the *ij*th sib. For the dominant model, g_{ij} was coded as 0 or 1 based on whether the allele of interest was absent or present in the *ij*th sib. The term for the environmental factor, e_{ij}, was coded in a similar manner as the genetic dominant model. The estimated terms were: β_1 the effect estimate of the genetic factor, β_2 the effect estimate of the environmental factor, β_3 the effect estimate of the cross product of the genetic and environmental factors (i.e., the GxE interaction term), and β_0 the model intercept term. All models were implemented using an autoregressive correlation structure.

Phenotyping. All probands had physical examinations and biochemical profiles at time of presentation. Criteria for organ system involvement have been previously described (36). Three broad phenotype categories were used: Mild, Moderate, or Severe. Mild: asymptomatic or normal organ systems without treatment, or spontaneous resolution to normal within 2 years of presentation; Moderate: symptomatic disease course with minimal residual abnormality in any organ with or without treatment, or disease course for more than 3 years, with some residual abnormality despite therapy; Severe: progressive organ dysfunction over three or more years despite therapy.

The citations in this online supplement refer to the reference list in the main article.