Published in final edited form as:

Methods Mol Biol. 2012 ; 882: 215–244. doi:10.1007/978-1-61779-842-9_13.

Analytical Methods for Immunogenetic Population Data

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

Abstract

In this chapter, we describe analyses commonly applied to immunogenetic population data, along with software tools that are currently available to perform those analyses. Where possible, we focus on tools that have been developed specifically for the analysis of highly polymorphic immunogenetic data. These analytical methods serve both as a means to examine the appropriateness of a dataset for testing a specific hypothesis, as well as a means of testing hypotheses. Rather than treat this chapter as a protocol for analyzing any population dataset, each researcher and analyst should first consider their data, the possible analyses, and any available tools in light of the hypothesis being tested. The extent to which the data and analyses are appropriate to each other should be determined before any analyses are performed.

Keywords

Data analysis; Highly polymorphic; HLA; Immunogenetics; KIR; Population study

1. Introduction

The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis, but it has no power of anticipating any analytical revelations or truths. Its province is to assist us in making available what we are already acquainted with. For, in so distributing and combining the truths and the formulae of analysis, that they may become most easily and rapidly amenable to the mechanical combinations of the engine, the relations and the nature of many subjects in that science are necessarily thrown into new lights, and more profoundly investigated (Ada Augusta, Countess of Lovelace).

While data analysis is a central component of modern genetic and genomic research approaches, most analytical methods have not been developed specifically for immunogenetic data; the high level of polymorphism at the HLA and KIR loci necessitates analytical methods and software tools with the capacity to process 20 or more alleles per locus for many loci. In addition, the extensive linkage disequilibrium (LD) between immunogenetic loci (e.g., covering 3 MB in the MHC) requires the simultaneous computation of association measures between as many as 14 (e.g., in the KIR complex) highly polymorphic loci. Because there is no single tool that can carry out all analyses, and

because available tools are not perfect for immunogenetic data, specific concessions must sometimes be made to enable analysis. Researchers and analysts should always remain aware that even when an analytical application is appropriate to the data, analysis can still be confounded by variation in nomenclature and data-resolution (see Chap. 12).

As Ada Lovelace recognized in 1843, analytical tools do not provide new insights or reveal truths; they may allow the implementation of complex methods, but these methods and their assumptions should always be known and understood to the researcher. Therefore, in addition to the discussions in this chapter, researchers should be familiar with the literature describing each method and the manual describing the use of for each analytical tool. No analytical result should be accepted without being presented in the appropriate context.

In general, while some of the analyses described below *can* be performed on paper or in a spreadsheet application, all are best carried out by a dedicated software tool that has been specifically developed with that analysis in mind. This provides a facile means to describe what was done, minimizes error on the part of the analyst, and maximizes reproducibility both by the researcher and by other researchers. Software tools generally run in either a Microsoft Windows, Apple, or Linux/Unix environment (or in a web-browser), but because many tools do not run in all three of these operating systems, we do not recommend one over the others. Researchers should maintain access to all three operating systems in order to take advantage of all available tools.

Many of these calculations are “computationally expensive” in that they are CPU- and memory-intensive. In general, the faster the CPU and the more RAM on a system, the faster an analysis will progress, but contemporary computer systems should be sufficient to the task of running the applications described here. However, care should be taken to ensure that the most recently released version of any application is used for analysis; this will ensure that any known software issues have been addressed. All of the tools that we describe here are free, so there is no reason to rely on an outdated version. Finally, many of the analyses described here have been implemented as functions and packages written in R, the language and environment for statistical computing (1). While we recommend specific R packages for analysis, and generally recommend using the R language, a tutorial on R programming is beyond the scope of this chapter.

In the sections that follow, we may discuss one particular software tool or application in the context of a given analysis. Relationships between analytical methods and some available tools are summarized in Table 1. However, this table is not intended to be comprehensive; there are many more possible analyses than those described here, and there are many more tools available than are discussed. For example, the compilation of genetic analysis software at <http://www.nslj-genetics.org/soft/> (2) includes 520 applications. Finally, supplementary population datasets (derived from the master data set included in the supplementary materials for Chap. 12) and associated files are available online at <http://methods.immunogenomics.org>; these datasets demonstrate input formats and were used to generate the example figures and tables associated with each analysis.

2. Data Reporting

The verification of published findings through independent replication is a core element of scientific research. Whereas replication is often interpreted to pertain primarily to the generation of experimental data, it is equally important that analyses of those data be replicated. To facilitate replication of analytical outcomes, the data analyzed must be reported in an accurate and thorough manner. Data described in the body of an immunogenetic paper should be presented as both raw allele counts and the allele frequencies calculated from them; this will allow other investigators to perform additional analyses (using counts) and will permit easy identification of the extent of differences in frequencies. In addition, all alleles, genotypes, and haplotypes (including rare variants) should be made available, either within the body of the paper or as supplementary material.

3. Analyses

3.1. Calculation of Gene (Allele) Frequencies

1. Direct Counting—With the advent of molecular typing techniques, the need to estimate gene (allele) frequencies (GF) from phenotype data has diminished. In most cases, gene frequencies for HLA data can be obtained via direct counting, where the number of observations for a given allele is divided by the number of chromosomes ($2n$, where n = sample size) under study.

The PyPop (python for population genomics) application (3) can be used to calculate allele frequencies in this manner. Supplementary Tables S1, S2, S3, and S4 are PyPop-formatted synthetic genotype data files for four populations. Supplementary Table S5 is a PyPop configuration file specific for these data. Further examples of PyPop-formatted HLA allele-frequency data and configuration files, developed by Solberg et al. (4), can be found at <http://www.pyPop.org/popdata>. These calculations are also easily accomplished directly from a Microsoft Excel spreadsheet via the Genetic Analysis in Excel (GenAlEx) Excel add-on (5).

2. Estimation from carrier frequencies—Direct counting cannot be used for all immunogenetic data. A notable exception remains for the KIR loci, where much of the available data have been generated on a presence/absence basis for many KIR loci, yielding phenotypes for which only carrier frequencies (CF, the presence of one or two copies of the considered allele) can be obtained by direct counting. In these cases, it is necessary to estimate gene frequencies. This can be done with the assumption that the population under study is in Hardy–Weinberg equilibrium (HWE) (see the final equation in step 3). The most simple equation is given by:

$$GF = 1 - \sqrt{1 - CF}.$$

However, Lynch and Milligan (6) have shown that this provides a downwardly biased estimate and suggest that a better estimate is obtained by:

$$GF = 1 - (x^{1/2}(1 - (\text{var}(x)/8x^2))^{-1}).$$

where $x = 1 - CF$, and $\text{var}(x) = x(1 - x)/N$, where N is the number of sampled individuals.

Table 2 presents gene frequencies estimated using each of these methods (formula a and formula b) in comparison to frequencies calculated by direct counting.

3. Confidence intervals—Any estimated $GF^{\wedge}$ should be accompanied by a confidence interval (CI), a range of values that reveals the precision of the estimated frequency. The likelihood that the CI includes the actual population GF is given by the confidence level (usually, a 95% chance). For GF estimated from a sample of size n , the CI has a lower bound of:

$$GF - \varepsilon \sqrt{(GF(1 - GF)/n)}.$$

And an upper bound of:

$$GF + \varepsilon \sqrt{(GF(1 - GF)/n)},$$

where ε invokes the Normal distribution to determine the probability of the CI associated with the estimated GF. For a 95% CI, ε is 1.96, and for a 99% CI, ε is 2.576.

$\sqrt{(GF(1 - GF)/n)}$ is an approximation of the standard error of the estimated GF.

3.2. Hardy–Weinberg Testing

The Hardy–Weinberg (HW) principle provides a useful model for primary quality control (QC) verification of the integrity of genotype data, as genotyping errors may result in both individual genotype deviations and overall deviations from HW equilibrium (HWE) (see Note 1). In addition, HW testing is also useful for detecting sampling errors (see below) in population samples. Confidence in the accuracy of Hardy–Weinberg testing is therefore crucial for confidence in subsequent analyses, as many analytical methods (e.g., LD and haplotype estimation, Ewens–Watterson analyses of selection) are predicated on an assumption of HWE in the data set. In a Hardy–Weinberg test, observed genotype counts are

¹For the most part, publications that use HW testing in this manner, including this chapter, share a common approach: regardless of the method, all attempt to detect a statistically significant difference between the observed genotype frequencies and those expected under HWE; these expected frequencies derive from a Null hypothesis (H_0) that assumes the HW model is true. While much effort has gone into the development of these statistical methods, the approach itself is rarely questioned. Perhaps, rather than evaluating data on the basis of a failure to detect statistically significant HW deviations, it should be demonstrated that any detected departure from HWE is below a critical threshold, allowing one to assume that the HW model still applies.

The field of bioequivalence clinical trials has shown that classical difference testing may not be optimal for assessing absence of association when dealing with the modest effects that characterize departure from the HW model in genetic epidemiology studies. Whereas difference testing returns the probability of observing a difference by chance, equivalency testing returns the probability of observing a lack of departure (or a modest departure) by chance. Equivalence testing is often implemented as two one-sided tests: one returns the probability of observing a lack of difference if the actual departure is positive (e.g., homozygote excess), and the other returns the probability of a lack of difference if the actual departure is respectively negative (heterozygote excess). In the future, a more natural approach to the HW testing may better quantify the extent to which data do not depart from HWE.

compared to those expected under Hardy–Weinberg equilibrium proportions (HWE), as calculated by generating a table of all possible genotypes, using an appropriate statistical method. The relationship between the allele and genotype frequencies under HWE is given as:

$$f(A_i A_i) = p_i^2$$

and

$$f(A_i A_j) = 2p_i p_j,$$

where p_i is the allele frequency of A_i , and p_j is the allele frequency of A_j . When a population is in HWE, there will not be a significant departure from these allele and genotype frequencies and there will be no change in allele frequencies between generations.

Tests of *overall* locus-level HWE compute a p -value to estimate the significance of observed deviations across all genotypes. Significant deviation of observed genotype counts from expected HWE can result from factors that include sampling errors (the sampling of admixed, stratified, or some other form of blended populations), inbreeding or other nonrandom mating, natural selection, and genotyping errors. Tests for deviation from HWE have low power, and significant deviation from HWE is *not* common. Genotyping errors (e.g., failure to detect a specific allele, resulting in an excess of homozygotes) are the first consideration when significant deviations from HWE are detected (especially when such deviations are detected only at a single locus in a multilocus analysis), rather than the operation of selection, admixture, or nonrandom mating, unless the sample is suspected to be from an unusual population.

1. Chi-square testing—Hardy–Weinberg testing can be particularly challenging for highly polymorphic datasets. Historically, the chi-square test has been the standard approach for testing fit to HWE at the overall locus-level (regardless of the level of polymorphism at that locus). However, this test can lead to false acceptance or rejection of the null hypothesis when individual expected genotype counts in the table of all possible genotypes are small or close to zero (also known as, “sparse cells” in the table of all genotypes, as represented by the shaded cells with low numbers of expected genotypes in the upper half of Fig. 1). Sparse cells can be problematic because the minimum number of observed genotypes must be 1, while the minimum number of *expected* genotypes will be the square of the frequency of the rarest allele. It is not unusual for 30–40 or more alleles to be observed at the highly polymorphic HLA loci, with a wide range of frequencies, resulting in many such sparse cells. As a result, the ratio of observed to expected (O/E) genotypes can be as large as 100 or 1000 for rare genotypes in large populations with no actual HWE deviation, while the O/E ratio for common genotypes will usually be much smaller (usually between 0.2 and 5), even in cases of actual deviation from HWE. Three approaches can be taken to increase the accuracy of Hardy–Weinberg tests of immunogenetic data; (1) rare alleles can be “lumped” together in a combined class (as in the lower half of Fig. 1) for the chi-square test to be effective; (2) a complete enumeration of all possible tables of all possible genotypes (an

exact test) can be undertaken, or (3) approximations to such complete enumerations can be made via resampling.

2. Exact tests—An exact test for HWEP was developed by Louis and Dempster (7). This test generated all possible tables of genotypes (based on observed allele frequencies) when the sample size and allele frequencies are held constant in accordance with the exact distribution. The p -value was given by the cumulative conditional probability of obtaining a table of genotypes (with sample size and allele frequencies equal to the observed sample) with a conditional probability less than or equal to that of the genotypes in the observed sample (8, 9). This test provides the exact p -value for every sample and it does not require input parameters that may affect the result. However, the number of possible tables of genotypes grows exponentially as either the sample size (n) or the number of distinct alleles (k) increases, reducing the feasibility of this test when n and k are large.

3. Resampling approximations—Resampling *approximations* to complete enumeration of all possible tables were developed for data sets with larger numbers of alleles, where the *asymptotic* chi-square test may be particularly problematic and *exact* tests with complete enumeration were not possible (10–13). While these *approximation* or resampling tests are often erroneously referred to as “exact” tests, they do not perform a true exhaustive search for all possible tables of genotypes. These methods generally use the Monte Carlo (MC) simulation method to approximate to the exact p -value and, therefore, represent an acceptable alternative to the *exact* test.

Guo and Thompson (10) developed the first conventional MC test of HWEP based on Levene’s conditional sampling distribution and also proposed a MC test that uses a finite and irreducible Markov Chain (MCMC) (14) to randomly generate tables of all possible genotypes. In these MC-based tests, the p -value is given by the fraction of randomly generated tables with a conditional probability less than or equal to the conditional probability of the observed genotypes. Resampling MC and MCMC tests perform very favorably when compared to the *exact* test and always outperform the chi-square test. However, the MCMC method may fail to approximate to the exact p -value in a few cases, and the MC test is preferred in cases where the *exact* test cannot be performed.

Chi-square tests, and MC and MCMC resampling approximations of the exact test are performed by PyPop, which has been designed specifically for the analysis of highly polymorphic immunogenetic data. For the chi-square test, PyPop automatically creates combined categories of rare alleles based on a user-defined “lumping” threshold, with a default value of 5.

4. Hardy–Weinberg testing of individual genotypes—Chen et al. (15) measured the goodness of fit of individual genotypes to expected HWEP in MC approximations of the exact test by comparing disequilibrium coefficients (16, 17). PyPop calculates two measures (Chen and Diff) of the goodness of fit of individual genotypes when the MC or MCMC test is implemented. In cases when locus-level deviations from HWEP are detected, these individual genotype tests may help identify the specific genotypes contributing to the deviation, but should be considered only when the number of expected genotypes is at least

1. As noted above, large p -values may result when a genotype with an expected count much less than 1 is observed once or twice; researchers should avoid making analytical inferences on the basis of these p -values.

3.3. Haplotype Estimation

Estimated haplotypes and haplotype frequencies 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 bone-marrow donor registries (BMDRs). Associations between markers and disease loci that are not evident with a single-marker locus may be identified in multilocus marker analyses using estimated haplotype frequencies (HFs). The design of studies and the recruitment of the samples are dependent on the possibility of identifying haplotypes by segregation analysis in families or estimating haplotypes from population samples of phase-unknown unrelated individuals (18). Haplotypes are used for disease association mapping, QTL mapping, and even imputing underlying genetic markers (19).

The term “haplotype” now includes any set of genetic polymorphisms (i.e., all DNA sequence variation including deletion/insertions) at contiguous loci. Except when recombination occurs, these neighboring genetic polymorphisms are cotransmitted by a single parental chromosome. Haplotypes may be represented as blocks of DNA sequence variants (e.g., SNP haplotype blocks), or groups of sequence variants can be abstracted into an allelic nomenclature at the level of a functional locus, as in the HLA and KIR systems.

3.3.1. Expectation-Maximization Algorithm—Early work on the estimation of haplotype frequencies from unrelated genotype data was based on the expectation-maximization (EM) algorithm with the assumption of HWEP at the locus-level (20–24). Later work refined, explored, and extended aspects of the algorithm (25–30). Application to haplotypes of SNPs (31–33) and Bayesian methods (34, 35) are commonly used. It remains unclear whether the Bayesian algorithms perform better than maximum likelihood implemented in EM algorithm (35).

Haplotypes can be estimated using a number of software tools, for example, in a standard implementation of the Expectation-Maximization (EM) algorithm in the haplo.stats package for R, the open source language, and environment for statistical computing (1). Although there is a great desire within the immunogenetic community for applications capable of analyzing very large (>million individuals) data sets, available HF and LD estimation software are generally limited in their capacity to a few thousands of individuals. For example, precompiled versions of PyPop are currently limited to 7 loci and 5,000 individuals when it comes to estimating haplotypes and calculating LD values. In contrast, haplo.stats can accommodate very large datasets, depending on the number of alleles at each locus; for example, haplo.stats estimates haplotypes for 240,000 individuals over four loci with an average of 25.5 alleles per locus, or 60,000 individuals over 50 loci with the same mean number of alleles. Supplementary Table S6 is a haplo.stats-formatted version of the synthetic genotype data in Supplementary Tables S1–S4. The “master data file” described in

Chap. 12 of this volume (Chap. 12 Supplementary Table S1) can also serve as a haplo.stats input file.

Population-level haplotype frequencies are estimated via EM using simultaneous maximum-likelihood estimation of n -locus haplotype frequencies. The expectation step determines the expected number of copies for each haplotype contributing to a given genotype. For a three locus haplotype, this is calculated as:

$$E[n_{abc}|P_i] = 2f_{abc}Sf_{abc}/P_r(P_i),$$

where S is the number of ambiguous haplotypes in P_i , $E[n_{abc}|P_i]$ is the expected number of copies of haplotype H_{abc} within P_i , and f_{abc} is the frequency of each other possible haplotype H_{abc} to form the genotype of frequency P_i . The maximization step determines new estimates for f_{abc} for the next iteration of the algorithm. At each iteration, the estimations globally improve.

3.3.2. Challenges to the Use of Estimated Haplotypes

1. Rare estimated haplotypes: The performance of haplotype frequency estimation algorithms is sensitive to various aspects of the population under study (35). Estimated frequencies for rare haplotypes ($n = 1$ or 2 in a dataset), which incorporate low-frequency alleles, are often incorrect, even when the EM algorithm finds the global maximum likelihood (27, 36, 37). The accuracy of haplotype estimates is critical for association and candidate gene studies, fine-mapping of disease genes, and for microsatellite, SNP, and protein level variation, and the presence or absence of specific low-frequency alleles and haplotypes must inform the robustness of associations. Analytical inferences should not be made on the basis of these rare haplotypes.

2. Haplotype estimation for immunogenetic data: The diversity and complexity of Immunogenetic data poses additional challenges for haplotype estimation. Over the past 30 years, the Immunogenetic community has seen an exponential increase of the number of HLA alleles leading to regular nomenclature revisions, and this phenomenon now extends to the KIR genes (38). In both the MHC 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 allele per loci (roughly >50), and high haplotype diversity (roughly >1,000). In addition, KIR and HLA data are very sensitive to ethnic background diversity. The potential for population substructure is particularly relevant for immunogenetic data due to the fact that MHC and KIR genes can reflect both the selective and demographic histories of populations. These issues are exacerbated in BMDRs where sample sizes for specific research questions are often very large (>100,000).

Little is known about the behavior of estimated haplotypes in the extreme situations described above for the HLA and KIR regions and little attention has been paid to the biases affecting haplotype frequency estimation. The frequency of the alleles, the sample size of the dataset, the various levels of missing information, and the various levels of linkage disequilibrium surely influence the accuracy of the estimation. Haplotype frequency

estimations are primarily affected by sampling fluctuation. In HLA and KIR, it is highly likely that the sample sizes are usually too small to cover the extent of the haplotype diversity. As a result, the haplotype frequencies and linkage disequilibrium between alleles are overestimated; such a bias would occur even if the chromosome phase was known.

3. HF 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 must be modified to account for this reduced combinatorial space, e.g., using an a priori list of known/possible haplotypes to constrain the EM algorithm (39, 40). 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.

Several recent KIR HF estimation studies have noted shortcomings in the use of such constraints, imposed by the need to specify predefined haplotype patterns (Fig. 2). Yoo et al. (40) found that accuracy measures related to haplotype identification were particularly low for fewer than 200 individuals and suggested that more than 500 individuals would provide acceptable estimation accuracy. In describing their HAPLO-IHP software, Yoo et al. noted that unusual haplotypes incompatible with constraints may be incorrectly rejected. When the a priori list of user-defined haplotypes does not span the observed genotypes, haplotypes that may not exist are “constructed” in an attempt to satisfy user-defined haplotype patterns.

3.4. Measures of Linkage Disequilibrium

Linkage disequilibrium is defined as the nonrandom association of alleles at two loci. High levels of LD combined with high levels of polymorphism are the defining characteristic of immunogenetic loci. Measurement of LD provides a means to assess the degree to which pairs of alleles are likely to be observed on the same haplotype and has important implications in analyzing immunogenetic data for population and disease association studies.

3.4.1. Haplotype-Level LD statistics

1. D_{ij} and D'_{ij} : Pairwise disequilibrium statistics can be calculated for each haplotype for polymorphic loci:

$$D_{ij} = x_{ij} - p_i \times q_j,$$

where x_{ij} is the estimated haplotype frequency (see previous section) and p_i and q_j are the i th and j th allele frequencies at the two loci. In order to account for differing allele frequencies at the loci, a normalized disequilibrium value can be used (41). This is given by:

$$D'_{ij} = D_{ij} / D_{\max},$$

where $D_{\max}$ is the lesser of $p_i q_j$ and $(1-p_i)(1-q_j)$, when D_{ij} is <0 and $p_i(1-p_i)$ and $q_j(1-q_j)$, when D_{ij} is >0 .

2. r^2 : The r^2 measure is another means of normalizing D_{ij} to account for differing allele frequencies. This is the square of the correlation coefficient (r) between the alleles at the p and q loci (42). Because r is given as:

$$r = (D_{ij} / p_i \times q_j (1 - p_i)(1 - q_j))^{1/2},$$

r^2 is therefore given as:

$$r^2 = D_{ij}^2 / (p_i \times q_j (1 - p_i)(1 - q_j)),$$

3.4.2. Global LD Statistics—For loci with more than two alleles, global LD statistics extend the haplotype-level statistics to account for all possible combinations of alleles at each locus (43).

1. W_n : W_n is a multiallelic extension of the correlation measure r . The chi-square value for testing the significance of LD can be written as $W/(2N)$ where:

$$W = (\sum \sum D_{ij} / p_i \times q_j)^{1/2},$$

and p_i and q_j are the observed allele frequencies at each of the two loci having k and l alleles, respectively. W_n , or Cramer's V statistic, is a normalized value that addresses differing numbers of alleles at the two loci (44, 45).

$$W_n = W / \sqrt{(\min(k, l) - 1)}.$$

The values of W_n fall between 0 and 1, and the significance of the overall disequilibrium is assessed using the abovementioned chi-square test. It should be noted that the W_n measure is always symmetric with respect to two loci, whereas the number of alleles reported at each locus can differ considerably. It is therefore important not to overinterpret values of W_n for locus pairs with highly asymmetric numbers of alleles. Finally, for biallelic loci, W_n is equivalent to r .

2. D'

D' is a second global disequilibrium statistic, which sums the absolute value of normalized D_{ij} values over all haplotypes, weighted by the frequencies of the alleles in each haplotype (46). As with W_n , D' values fall between 0 (equilibrium) and 1 (linkage). This is given as:

$$D' = \sum \sum p_i \times q_j |D'_{ij}|,$$

PyPop calculates D_{ij} , D'_{ij} , D' and W_n values.

3.4.3. Graphical Representation of LD Patterns—The interpretation of LD values between many markers can be facilitated through the graphical representation of LD patterns. Compared to a tabular presentation of LD values, such visual representations facilitate the identification of patterns and interesting subsets of the data. So-called “heat maps” are a common means of representing pairwise LD values across markers, as a half-matrix in which the strength of the LD (e.g., the log of the p -value) is represented by a color scale. However, most of the software tools developed for graphical LD presentations represent biallelic markers and can therefore only represent average LD between multiallelic loci. Popular software tools for this purpose include graphical overview of linkage disequilibrium (GOLD) (47), Haploview (48), MIDAS (49), and various packages in R (e.g., LDHeatmap (50)). While PyPop does not generate graphical LD representations, PyPop-generated LD data can be imported directly to R. To our knowledge, only MIDAS will simultaneously represent the interallelic component of LD.

3.5. Measurement of Selection

1. Ewens–Watterson homozygosity statistic—The expected proportion of homozygotes under Hardy–Weinberg, for an observed value of k and a given sample size (n), is used as a measure of the allele-frequency distribution and compared to the distribution expected under the neutral model for the same values of k and n (51). Allele-frequency distributions are used to calculate Watterson’s homozygosity F statistic (52). This is given by:

$$F = \sum p_i^2,$$

where p_i is the frequency of the i th allele at a locus. The homozygosity test can be accomplished using the exact test described by Slatkin (53, 54). For given values of n and k , all possible configurations of alleles are listed (each configuration is a distinct way of distributing the n sampled genes into k allelic categories). The probability of obtaining a particular configuration can be computed under the null hypothesis of neutrality using the Ewens sampling formula (51). The homozygosity value of each configuration along with its probability gives the sampling distribution for F under neutrality. This distribution is used to find the probability of obtaining homozygosity values equal to or larger than that observed, for a test of positive selection, by examining how many configurations result in homozygosities greater than this observed value (53). Similarly, a test of balancing selection is based on the probability of obtaining a homozygosity value as small as or smaller than the observed value. Significant p -values reject the null hypothesis that the sample came from a population that is undergoing neutral evolution.

2. Normalized deviate of homozygosity—Homozygosity values calculated for different values of n and k can be directly compared by calculating the normalized deviate of homozygosity (F_{nd}) (55). This is given by:

$$F_{nd} = (F_{obs} - F_{exp}) / \sqrt{\text{var}(F_{exp})},$$

where F_{obs} is the homozygosity value calculated for an observed frequency distribution, F_{exp} is the mean homozygosity expected under the neutral model. While F_{nd} is a normalized deviate (similar to a z-score), the sampling distribution for F_{nd} is not normally distributed, so that p -values cannot be inferred from a given F_{nd} value using traditional parametric methods. Statistical significance for an F_{nd} value is given by the significance of the corresponding F_{obs} value.

The normalized deviate of homozygosity can also be used to characterize homozygosity values that deviate significantly from the null hypothesis in terms of modes of evolution. F_{nd} values significantly lower than 0 result from allele-frequency distributions that are more “even” than expected and are consistent with the action of balancing selection. F_{nd} values significantly higher than 0 result from allele-frequency distributions that are more skewed than expected toward specific alleles and are consistent with either directional selection or an extreme demographic effect.

In addition, because F_{nd} is equal to 0 under the null hypothesis, a paired sign test (56) can be used to compare multiple F_{nd} values against the expectation of neutrality.

PyPop calculates F and F_{nd} values.

3.6. Measures of Genetic Diversity

1. Heterozygosity—The level of genetic diversity at a given locus is dependent upon the allele frequencies of the marker and the number of alleles observed in the sample. Within a given population, variation may be described by the heterozygosity (H), which ranges between 0 and 1:

$$H = 1 - \sum p_i^2,$$

where p_i is the frequency of the i th allele. For a population in HWE, this is the probability that a random individual in the population is a heterozygote. Heterozygosity will be maximized when all alleles are at an equal frequency.

2. Polymorphism information content—The polymorphism information content (PIC) value is an additional statistic based on allele frequencies at a locus that describes the ability of a marker to differentiate individuals within a population (57):

$$\text{PIC} = 2 \sum p_i \times p_j (1 - p_i \times p_j),$$

where p_i is the frequency of the i th allele, and p_j is the frequency of the $(i + 1)$ th allele. This is the probability that one of two individuals in a randomly mating population is a heterozygote and that the other is a different genotype. As with heterozygosity (H), PIC is

maximized when all alleles are at equal frequency. The values of H and PIC are very similar at high heterozygosities, but PIC will never exceed H and its values are less than H when heterozygosity is low.

3.7. Measures of Genetic Differentiation

The measures below are used to quantify genetic variation within and between populations and to determine subdivisions (subpopulations) of a single source (total population).

1. F_{ST} — F_{ST} values quantify levels of population differentiation by assessing the proportion of genetic variance in subpopulations relative to the total genetic variance (58). F_{ST} can be calculated based on a partitioning of heterozygosity:

$$F_{ST} = (H_t - H_s) / H_t,$$

where H_t is the heterozygosity of the total population, and H_s is the average heterozygosity of the subpopulations. Nei (59) has shown that this can be expressed in terms of allele frequencies:

$$F_{ST} = \sum \text{var}(p_i) / \sum \text{var}(1 - p_i),$$

where p_i is the allele frequency of the i th allele in the total population and $\text{var}(p_i)$ is the variance of the i th allele over subpopulations. F_{ST} is a qualitative measure, and Wright (60) has suggested guidelines for interpretation of these values:

0–0.05 indicates little genetic differentiation between subpopulations

0.05–0.15 indicates moderate genetic differentiation between subpopulations

0.15–0.25 indicates great genetic differentiation between subpopulations

0.25 and above indicates very great genetic differentiation between subpopulations

It has been shown that there is good concordance between F_{ST} and average divergence times within and between subpopulations, given neutral loci and assuming the infinite alleles mutation model (61). F_{ST} has been traditionally applied to data such as those obtained for allozyme variation. This measure may lose some power when applied to loci with a relatively high mutation rate, as has been suggested for microsatellite loci. Several different methods have been applied to estimate mutation rates for microsatellites (62, 63) and rates ranging from 10^{-3} to 10^{-5} have been reported.

2. R_{ST} —Additional measures related to F_{ST} have been described for application to microsatellite data. Slatkin (64) introduced R_{ST} , which has similar properties to F_{ST} , but assumes a stepwise mutation process, as well as a relatively high mutation rate. It is calculated as:

$$R_{ST} = (S - S_w) / S,$$

where S_w is the sum over all loci of twice the weighted mean of the within population variances, $V(A)$ and $V(B)$, and S is the sum over all loci of twice the variance of the combined populations, $V(A + B)$. In computer simulations, it was demonstrated that R_{ST} may provide a relatively more unbiased estimate of coalescence times compared to F_{ST} .

3. Population-pairwise F_{ST} —The degree of differentiation between pairs of populations (population-pairwise F_{ST}) can be used to investigate the existence of population structure (64–66). In accounting for small differences in subpopulation sample sizes, the population-pairwise F_{ST} calculation may result in small negative pairwise F_{ST} values; it is common practice to treat these negative values as being equivalent to zero. Pairwise standardized F_{ST} values ($F_{ST'}$ values) are generated using Hedrick's method of dividing each value by the maximum population-pairwise F_{ST} value (67), allowing comparison of genetic differentiation between loci with different mutation rates and between populations with different effective sizes. For n subpopulations, this population-pairwise approach results in a matrix of $n(n-1)$ F_{ST} or $F_{ST'}$ values.

Arlequin will calculate F_{ST} and population-pairwise F_{ST} values. Supplementary Table S7 includes Arlequin-formatted genotype data for the synthetic datasets discussed in this chapter.

3.8. Graphical Representations of Genetic Difference Data

There are a variety of methods for representing genetic difference data between subpopulations in a graphical (as opposed to tabular) format. In many cases, the graphical representation can be applied independently of the measure of differentiation, so that multiple different genetic differentiation measures can be compared using the same graphical representation and multiple graphical representations can be applied to the same genetic differentiation measure. Because the graphical representation usually depends on an additional analysis, we describe some of the commonly used representations here as individual analyses.

In general, these representations should not necessarily be thought of as providing the definitive answer to a question so much as they serve as aids for the interpretation of genetic differentiation data that may be too complex to present in a tabular format (as with LD values). As Ada Lovelace noted, “the Analytical Engine has no pretensions whatever to *originate* anything.” The results of these methods should always be interpreted critically, and the researcher who uses these methods should develop a set of criteria for accepting or rejecting the results of a method *before* using that method. Overinterpretation of any of these representations should be avoided when there is no obvious historical, functional, or biological basis for them.

1. Principal component analysis—Principal component analysis (PCA) is used for dimensionality reduction in a data set, identifying those elements that contribute most to its variance, and is particularly useful as an exploratory tool in a complex data set (68). For the representation of genetic differentiation analyses, it is common to present the results of a PCA via multidimensional scaling (MDS), where each range for a given component is presented along a corresponding MDS axis (69, 70). Each data-element (a population or an

individual) is represented by its position relative to each axis. This MDS approach allows the comparison of similarities and differences between populations, or the individuals that they comprise. For a 2D PCA MDS plot, distribution of points along the primary (x) axis will correspond to the greatest amount of variation in genetic distances in the data set (the first principal component); distribution along the y -axis corresponds to the next highest degree of the remaining variation that is not correlated with the x -axis (the second principal component). Because there can be many more than two principal components (as long as there remains variation that is not correlated with higher order components), comparisons can be related with multiple 2D PCA plots, representing the intersection of different components, or with multiple 3D visualizations. However, increasingly smaller percentages of the variance are represented by the higher numbered components, and these are usually not presented. In some cases, it may be necessary to present multiple plots for the same components. For example, when some populations display extensive genetic differentiation relative to others, it is often difficult to illustrate the differences between populations with relatively low degrees of differentiation; the PCA for these latter populations can be presented in a MDS plot with a smaller scale.

As noted above, PCA can be used to investigate differentiation between sampled individuals or between populations. The PCA-mediated comparison of individuals in multiple populations is in essence a population structure analysis, which is described below. For population-level analyses, genetic distances are first calculated in a pairwise fashion (e.g., as population-pairwise F_{ST} values) between populations, and PCA is performed to assess the variation in distance between populations. PCA MDS analysis is available in a wide variety of statistical software applications (e.g., GenAlEx, and R packages). A population-level PCA would proceed via the following steps:

- i. Calculate allele or haplotype frequencies in a set of subpopulations.
- ii. Use a differentiation measure to generate a genetic distance matrix for the subpopulations.
- iii. Calculate the principal components from the distance matrix.
- iv. Generate a (series of) MDS figure(s) representing two or three of the principal components.

A 2D population-level PCA MDS plot generated in GenAlEx for the synthetic datasets discussed in this chapter is presented in Fig. 3. GenAlEx formatted data are included in Supplementary Table S8. (see Note 2).

2. Population structure analysis—Population substructure and population admixture can be directly investigated by estimating the likelihood that a given genotype belongs to a specific population. Likelihood values can be calculated based on HWEP, allele frequencies,

²To generate a population-level PCA plot using GenAlEx, population allele frequencies are calculated and a genetic distance matrix is derived from the frequency distributions for each locus. In this case Nei's unbiased genetic distances (95) were used, but Nei's standard genetic distances (78) can also be used. The PCA is based on the genetic distance between population groups. For a PCA plot at the individual level (see Fig. 4), a matrix of genetic distances is computed from the raw genotype data, and the PCA plot is generated from the between-individual distance matrix.

and LD (if available) and are used to assign individual genotypes to specific groups or clusters, which can range from individual populations to geographic regions. As with phylogenetic trees, these clustering results can be displayed using multiple graphical representations. A 2D individual-level PCA plot generated in GenAIEx for the synthetic datasets discussed in this chapter is presented in Fig. 4.

While there are many clustering tools available, the most widely used is Structure (71, 72), which Rosenberg et al. (73) used to cluster the populations of the CEPH Human Genetic Diversity Panel, largely by geographic origin, on the basis of genotype data for a genome-wide set of 377 microsatellites. Structure iteratively resamples individuals into a number of user-defined clusters (K) and calculates the likelihood for each organization via Bayesian inference from expected HWEs. In addition, other types of clustering analyses can also be carried out with Arlequin, CLUTO (74), and various R packages.

Supplementary Table S9 contains Structure-formatted genotype data for the synthetic datasets discussed in this chapter. Figure 5 includes structure plots generated with these synthetic datasets for $K = 2$ and 4 (see Note 3).

3. Phylogenetic analysis—A phylogenetic tree (aka dendrogram) is a branching representation of the evolutionary history between populations, individuals, or gene/protein sequences (taxa) based upon similarities and differences in some characteristic (for our purposes, immunogenetic allele and haplotype frequencies) shared by all taxa. Trees generated using gene or protein sequences (sequence trees) often allow inferences regarding the relative “age” of sequence variants and the inference of ancestral sequences. We do not discuss sequence trees here. Trees generated using population-level allele-frequency data (population trees) can represent relative degrees of shared ancestry between populations. Where sequence trees can be interpreted as gene genealogies, population trees should be considered as graphs of the general trends in relationships between modern populations, which can change in ways (e.g., admixture, splitting, fusion, bottlenecks, etc.) that nucleotide and protein sequences cannot. In particular, the relationships represented by population trees are generally representative of the first (and sometimes second) principal components of the frequency data used to generate them. Population trees are generated via the following general steps.

- i. Calculate allele or haplotype frequencies in a set of subpopulations.
- ii. Use a differentiation measure to generate an estimated genetic distance matrix for the subpopulations.

³When creating a Structure project with this file, indicate the number of individuals (694), the ploidy of the data (2, diploid), the number of loci (2), the value provided for missing data (–1), the presence of a header row of marker names, the special format for the file including all data for each individual on a single line, the inclusion of a column for the sample ID of each sampled individual, and the inclusion of a column identifying the population origin for each sampled individual. Then, create a parameter set that defines the ancestry model (admixture), the allele-frequency model (independent), and the run-length in terms of the length of Burnin Period (50,000), and the number of Monte Carlo Markov Chain (MCMC) reps after Burnin (50,000). For other datasets, these last two values will need to be determined empirically, by observing the number of reps necessary for the alpha-value to converge. To start a Structure run, specify the number of clusters (K) assumed for the data (2 or 4). Group the resulting Bar Plot by population ID to generate results similar to those in Fig. 5. Because Structure clustering is accomplished via Bayesian inference, Structure should be run multiple times for datasets that include large numbers of individuals at many markers, and the results compared for overall trends.

- iii. Calculate the tree topology from the distance matrix.
- iv. Generate a tree figure representing that tree topology.

Allele-frequency-based genetic distance calculations do not take the sequence relationships between individual alleles into account, so that all alleles are considered to be equidistant from each other; for many immunogenetic alleles (e.g., those that diverged prior to the radiation of human populations from Africa), this should not be an issue, but when populations display large differences in frequency for alleles or haplotypes that have been relatively recently generated (e.g., DRB1*08:02:01 and DRB1*08:07), genetic distances may be overestimated, resulting in very large branch lengths. Similarly, alleles that may be reported differently depending on the typing method used (e.g., DRB1*14:01:01 and DRB1*14:54) may result in similar overestimates. In these cases, the datasets should be reviewed for consistency in the level of resolution of typing, and alternate names for what is potentially the same allele should be binned into a common category (e.g., DRB1*14:01:01G). Finally, trees including isolated populations with low values of k , in which a few alleles are subject to genetic drift, may suffer similar problems. It is oftentimes useful to establish threshold for sample size (e.g., $2n > 49$) and k (e.g., >5) to exclude populations that are too small or display too few alleles.

In general, phylogenies can be drawn as either “rooted” or “unrooted” trees, as illustrated in Fig. 6 ; rooted trees identify a common ancestor for all of the taxa, giving the tree directionality from root to twigs. We recommend presenting population trees constructed using immunogenetic allele and haplotype frequencies as unrooted, as it is difficult to know exactly how or where to place the root. For example, the influence of low values of k on branch lengths (discussed above) raises questions about the effectiveness and appropriateness of midpoint rooting. Similarly, because these trees are based on frequency distributions, the lack of a nonhuman population sharing allelic and haplotypic diversity with human populations makes outgroup rooting difficult.

PHYLIP: The PHYLogeny Inference Package (PHYLIP) (75, 76) is a software suite of applications for building phylogenetic trees using a variety of methods. Supplementary Table S10 is a PHYLIP GENDIST-formatted allele-frequency data file for the synthetic population datasets analyzed in this chapter. Figure 7 includes a pair of unrooted Neighbor-Joining (NJ) (77) trees generated using the data included in Supplementary Tables S7 and S10.

The tree in Fig. 7a is based on Nei’s standard genetic distances (SGD) (78) calculated in PHYLIP, whereas the tree in Fig. 7b is based on population-pairwise F_{ST} values calculated in Arlequin. A genetic distance scale should be included with every tree.

Steps (ii)–(iv) outlined above were carried out with PHYLIP to generate Fig. 7a using the GENDIST (for step ii.), NEIGHBOR (for step iii.), and DRAWTREE (for step iv) programs to draw unrooted NJ trees based on Nei’s Standard Genetic Distances. Figure 7b was generated using the same procedure for steps (iii) and (iv), but steps (i) and (ii) were carried out using Arlequin to generate population-pairwise F_{ST} values.

PHYLIP's GENDIST estimates genetic distance with three different measures—Nei's SGD, Cavalli-Sforza's chord distance (79), and Reynold's genetic distance (65)—and each measure is based on implicit assumptions that may not always apply to immunogenetic data. For example, all three measures assume that population differentiation derives from genetic drift, yet the HLA loci have been shown to be under balancing selection in numerous studies (4, 80, 81).

Nei's SGD assumes that new alleles arise by neutral mutation, and that the mutation rate is equal across all loci; again, the latter assumption clearly does not hold for HLA loci, where there are many more class I alleles than class II alleles, and where many HLA-B alleles are observed to be restricted to specific regions of the world, whereas most HLA-DQA1 and DQB1 alleles are observed in all populations (82).

However, the Cavalli-Sforza and Reynolds distance models assume no mutation; frequency differences between populations are assumed to be the result of genetic drift alone. This assumption of no mutation seems even a further departure from observed immunogenetic biology than the assumption of locus-identical neutral mutation, as unique HLA alleles are observed on a regular basis (83), and natural selection appears to have favored novel HLA allele variants over older variants in North and South American populations (84, 85).

Clearly, none of these models applies perfectly to immunogenetic data. Our empirical experience has been that HLA gene-trees conform best to expectations when generated using Nei's SGD; this distance estimate includes a mutational component, which clearly applies to HLA data.

PHYLIP's NEIGHBOR (see Note 4) builds trees with either of two clustering methods—NJ and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (86, 87). For the purposes of this discussion, the primary difference between these methods is the assumption of the rate of evolution. The NJ method does not assume a regular stochastic evolutionary rate (i.e., a molecular clock), while the UPGMA method does. This difference means that NJ trees are inherently unrooted (and should be drawn as such), while UPGMA trees are inherently rooted. The assumption of a molecular clock in the UPGMA method makes these trees particularly ill-suited for building immunogenetic gene-frequency trees, as exonic immunogenetic allelic differentiation does not conform to a molecular clock model; for the HLA loci, different modes of evolution are in effect for introns, non-ARS-encoding exon sequences, and ARS-encoding exon sequences (85, 88, 89).

For the generation of immunogenetic gene-frequency trees in PHYLIP, we recommend building NJ trees with Nei's SGD.

4. Criteria for rejection or acceptance of trees—As discussed above, it is important to develop criteria for the purpose of evaluating trees prior to analysis. Sequence trees can be evaluated via bootstrapping (90, 91), in which a population of trees is generated by resampling subsets of the primary data. The topologies of the resulting trees are compared

⁴When using NEIGHBOR, researchers should always use the Jumble (J) option to randomize the input order of taxa.

and each branch in the consensus tree is evaluated based on degree of sharing of topological features across the population of trees.

For sequence trees, resampled datasets are generated by randomly sampling and duplicating subsets of nucleotide or peptide positions, but resampled gene-tree datasets are generated by randomly sampling and duplicating entire loci; therefore bootstrapping cannot be used to evaluate gene-trees constructed for single loci, and bootstrapping performs poorly for gene-trees constructed with a small number of loci. For example, gene-trees constructed using two loci will yield boot-strap values of 0.0, 0.5, or 1.0.

Studies of population relationships at non-MHC loci have generally shown a close correlation between genetics and geography, so that populations tend to share similar allele frequencies with their neighbors on a local level (92, 93). Therefore, it is not unreasonable to expect that most populations will be represented as being more closely related to their neighbors (populations in the same global region) than to nonneighboring populations in gene-trees, and that most relationships in a tree will corroborate geographic, historical, anthropologic, and linguistic evidence. Mack and Erlich (94) proposed that HLA gene-trees be rejected as invalid if more than 6% of the intraregional population relationships in that tree do not meet this expectation. This criterion is necessarily conservative; if a tree meets no expectations, it is difficult to say which relationships are genuine, and which might be spurious. Trees such as these may reveal more about the diversification of the loci investigated than they do about population relationships. Finally, population relationships, however unexpected, that are repeatedly observed in trees derived from different markers, clearly merit serious consideration as reflecting actual relationships rather than as a spurious artifact of the tree-building process.

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 by NIH/NIAID contract AI40076 (RMS, GT). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases or the National Institutes of Health.

References

1. R Core Development Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008.
2. Li W, Hu Z, Jiang W. An alphabetic list of genetic analysis software. North Shore LIJ Research Institute. 2010 <http://www.nslj-genetics.org/soft/>.
3. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. PyPop update—a software pipeline for large-scale multilocus population genomics. *Tissue Antigens*. 2007; 69(s1):192–197. [PubMed: 17445199]
4. 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(7):443–464. [PubMed: 18638659]
5. Peakall R, Smouse PE. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol Ecol Notes*. 2006; 6:288–295.

6. Lynch M, Milligan BG. Analysis of population genetic structure with RAPD markers. *Mol Ecol*. 1994; 3:91–99. [PubMed: 8019690]
7. Louis EJ, Dempster ER. An exact test for Hardy-Weinberg and multiple alleles. *Biometrics*. 1987; 43(4):805–811. [PubMed: 3427165]
8. Levene H. On a matching problem arising in genetics. *Ann Math Stat*. 1949; 20(1):91–94.
9. Emigh TH. A comparison of tests for Hardy-Weinberg equilibrium. *Biometrics*. 1980; 36(4):627–642.
10. Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*. 1992; 48(2):361–372. [PubMed: 1637966]
11. Huber M, Chen Y, Dinwoodie I, Dobra A, Nicholas M. Monte Carlo algorithms for Hardy-Weinberg proportions. *Biometrics*. 2006; 62:49–53. [PubMed: 16542228]
12. Yuan A, Bonney GE. Exact test of Hardy-Weinberg equilibrium by Markov chain Monte Carlo. *Math Med Biol*. 2003; 20:327–340. [PubMed: 14969383]
13. Ebrahimi N, Bilgili D. A new method of testing for Hardy-Weinberg equilibrium and ordering populations. *J Genet*. 2007; 86:1–7. [PubMed: 17656843]
14. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. *J Chem Phys*. 1953; 21:1087–1092.
15. Chen JJ, Hollenbach JA, Trachtenberg EA, Just JJ, Carrington M, Rønningen KS, Begovich A, King MC, McWeeney SK, Mack SJ, Erlich HA, Thomson G. Hardy-Weinberg testing for HLA class II (DRB1, DQA1, DQB1 and DPB1) loci in 26 human ethnic groups. *Tissue Antigens*. 1999; 54:533–542. [PubMed: 10674966]
16. Hernández JL, Weir BS. A disequilibrium coefficient approach to Hardy-Weinberg testing. *Biometrics*. 1989; 45(1):53–70. [PubMed: 2720060]
17. Chen JJ, Thomson G. The variance for the disequilibrium coefficient in the individual Hardy-Weinberg test. *Biometrics*. 1999; 55:1269–1272. [PubMed: 11315081]
18. Barnette T, Gourraud PA, Cambon-Thomsen A. Strategies in analysis of the genetic component of multifactorial diseases; biostatistical aspects. *Transpl Immunol*. 2005; 14(3–4):255–266. [PubMed: 15982570]
19. Guan Y, Stephens M. Practical issues in imputation-based association mapping. *PLoS Genet*. 2008; 4(12):e1000279. [PubMed: 19057666]
20. Piazza, A. Haplotypes and linkage disequilibria from three-locus phenotypes. In: Kissmeyer-Nielsen, F., editor. *Histocompatibility testing*. Copenhagen: Munksgaard; 1975. p. 923–927.
21. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc*. 1977; 39:1–38.
22. Ott J. Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis. *Ann Hum Genet*. 1977; 40:443–454. [PubMed: 879713]
23. Yasuda N. Estimation of haplotype frequency and linkage disequilibrium parameter in the HLA system. *Tissue Antigens*. 1978; 12:315–322. [PubMed: 371070]
24. Morton NE, Simpson SP, Lew R, Yee S. Estimation of haplotype frequencies. *Tissue Antigens*. 1983; 22(4):257–262. [PubMed: 6648948]
25. Hawley ME, Kidd KK. HAPLO: a program using the EM algorithm to estimate the frequencies of multisite haplotypes. *J Hered*. 1995; 86:409–411. [PubMed: 7560877]
26. Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet*. 1995; 56:799–810. [PubMed: 7887436]
27. Fallin D, Schork NJ. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet*. 2000; 67:947–959. [PubMed: 10954684]
28. Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK. The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. *Am J Hum Genet*. 2000; 67:518–522. [PubMed: 10859209]
29. Kirk KM, Cardon LR. The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur J Hum Genet*. 2002; 10:616–622. [PubMed: 12357332]

30. Single RM, Meyer D, Hollenbeck J, Nelson M, Noble JA, Erlich HA, Thomson G. 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.* 2002; 22:186–195.
31. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 2001; 68:978–989. [PubMed: 11254454]
32. Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet.* 2002; 70:157–169. [PubMed: 11741196]
33. Qin ZS, Niu T, Liu JS. Partition-ligation-expectation maximization algorithm for haplotype inference with single nucleotide polymorphisms. *Am J Hum Genet.* 2002; 71:1242–1247. [PubMed: 12452179]
34. Stephens M, Donnelly P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet.* 2003; 73:1162–1169. [PubMed: 14574645]
35. Niu T. Algorithms for inferring haplotypes. *Genet Epidemiol.* 2004; 27(4):334–347. [PubMed: 15368348]
36. Slatkin M, Excoffier L. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity.* 1996; 76:377–383. [PubMed: 8626222]
37. Tishkoff SA, Kidd KK. Implications of biogeography of human populations for ‘race’ and medicine. *Nat Genet.* 2004; 36(Suppl 1):S21–S27. [PubMed: 15507999]
38. Robinson J, Waller MJ, Fail SC, Marsh SG. The IMGT/HLA and IPD databases. *Hum Mutat.* 2006; 27:1192–1199. [PubMed: 16944494]
39. Gourraud PA, Gagne K, Bignon JD, Cambon-Thomsen A, Middleton D. Preliminary analysis of a KIR haplotype estimation algorithm: a simulation study. *Tissue Antigens.* 2007; 69(Suppl 1):96–100. [PubMed: 17445177]
40. Yoo YJ, Tang J, Kaslow RA, Zhang K. Haplotype inference for present-absent genotype data using previously identified haplotypes and haplotype patterns. *Bioinformatics.* 2007; 23(18):2399–2406. [PubMed: 17644820]
41. Lewontin RC. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics.* 1964; 49:49–67. [PubMed: 17248194]
42. Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet (Der Ziichter).* 1968; 38:226–231.
43. Klitz W, Stephen JC, Grote M, Carrington M. Discordant patterns of linkage disequilibrium of the peptide transporter loci within the HLA class II region. *Am J Hum Genetics.* 1995; 57:1436–1444. [PubMed: 8533774]
44. Cramer, H. *Mathematical methods of statistics.* Princeton, NJ: Princeton University Press; 1946.
45. Cohen, J. *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum; 1988.
46. Hedrick PW. Gametic disequilibrium measures: proceed with caution. *Genetics.* 1987; 117(2):331–341. [PubMed: 3666445]
47. Abecasis GR, Cookson WO. GOLD-graphical overview of linkage disequilibrium. *Bioinformatics.* 2000; 16:182–183. [PubMed: 10842743]
48. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 2005; 21(2):263–265. [PubMed: 15297300]
49. Gaunt TR, Rodriguez S, Zapata C, Day IN. MIDAS: software for analysis and visualisation of interallelic disequilibrium between multiallelic markers. *BMC Bioinformatics.* 2006; 7:227–237. [PubMed: 16643648]
50. Shin J-H, Blay S, McNeney B, Graham J. LD heatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J Stat Soft 16 Code Snippet 3.* 2006
51. Ewens WJ. The sampling theory of selectively neutral alleles. *Theor Popul Biol.* 1972; 3(1):87–112. [PubMed: 4667078]
52. Watterson G. The homozygosity test of neutrality. *Genetics.* 1978; 88:405–417. [PubMed: 17248803]
53. Slatkin M. An exact test for neutrality based on the Ewens sampling distribution. *Genet Res.* 1994; 64:71–74. [PubMed: 7958833]

54. Slatkin M. A correction to the exact test based on the Ewens sampling distribution. *Genet Res.* 1996; 68:259–260. [PubMed: 9062082]
55. Salamon H, Klitz W, Eastel S, Gao X, Erlich HA, Fernandez-Vina M, Trachtenberg EA. Evolution of HLA class II molecules: allelic and amino acid site variability across populations. *Genetics.* 1999; 152:393–400. [PubMed: 10224269]
56. Conover, W. Practical nonparametric statistics. New York: Wiley; 1980.
57. Chakravarti A. Information content of the Centre d'Etude du Polymorphisme Humain (CEPH) family structures for linkage studies. *Hum Genet.* 1991; 87:721–724. [PubMed: 1937475]
58. Wright S. The genetic structure of populations. *Ann Eugen.* 1951; 15:323–354. [PubMed: 24540312]
59. Nei M. F-statistics and analysis of gene diversity in subdivided populations. *Ann Hum Genet.* 1977; 41:225–233. [PubMed: 596830]
60. Wright, S. Evolution and the genetics of populations. Vol. 4. Chicago: The University of Chicago Press; 1978. Variability Within and Among Natural Populations
61. Slatkin M. Inbreeding coefficients and coalescence times. *Genet Res.* 1991; 58:167–175. [PubMed: 1765264]
62. Weber JL, Wong C. Mutation of human short tandem repeats. *Hum Mol Genet.* 1993; 2(8):1123–1128. [PubMed: 8401493]
63. Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, Bernardi G, Lathrop M, Weissenbach J. The 1993 – 1994 Genethon human genetic linkage map. *Nat Genet.* 1994; 7:246–339. [PubMed: 7545953]
64. Slatkin M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics.* 1995; 139:457–462. [PubMed: 7705646]
65. Reynolds J, Weir BS, Cockerham CC. Estimation for the coancestry coefficient: basis for a short-term genetic distance. *Genetics.* 1983; 105:767–779. [PubMed: 17246175]
66. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution.* 1984; 38:1358–1370.
67. Hedrick PW. A standardized genetic differentiation measure. *Evolution.* 2005; 59:1633–1638. [PubMed: 16329237]
68. Pearson K. “On lines and planes of closest fit to systems of points in space” (PDF). *Phil Mag.* 1901; 2(6):559–572.
69. Cox, TF.; Cox, MAA. Multidimensional Scaling. 2nd edn.. Boca Raton, FL: Chapman and Hall; 2001.
70. Borg, I.; Groenen, P. Modern multidimensional scaling: theory and applications. 2nd edn.. New York: Springer; 2005.
71. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000; 155(2):945–959. [PubMed: 10835412]
72. Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour.* 2009; 9(5):1322–1332. [PubMed: 21564903]
73. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic structure of human populations. *Science.* 2002; 298(5602):2381–2385. [PubMed: 12493913]
74. Karypis, G. Technical Report 02-017. Minneapolis, MN: University of Minnesota; 2002. CLUTO: a clustering toolkit.
75. Felsenstein J. PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics.* 1989; 5:164–166.
76. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6 (Distributed by the author). Seattle: Department of Genome Sciences, University of Washington; 2005.
77. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 1987; 4(4):406–425. [PubMed: 3447015]
78. Nei M. Genetic distance between populations. *Am Nat.* 1972; 106:283–292.
79. Cavalli-Sforza LL, Edwards AFW. Phylogenetic analysis: models and estimation procedures. *Am J Hum Genet.* 1967; 19:233–257. [PubMed: 6026583]

80. Hughes AL, Nei M. Pattern of nucleotide substitution at MHC class I loci reveals overdominant selection. *Nature*. 1988; 335:167–170. [PubMed: 3412472]
81. Hughes AL, Yeager M. Natural selection at major histocompatibility complex loci of vertebrates. *Ann Rev Genet*. 1998; 32:415–435. [PubMed: 9928486]
82. Sanchez-Mazas A, Fernandez-Viña M, Middleton D, Hollenbach JA, Buhler S, Di D, Rajalingam R, Dugoujon JM, Mack SJ, Thorsby E. Immunogenetics as a tool in anthropological studies. *Immunology*. 2011; 133(2):143–164. [PubMed: 21480890]
83. Middleton D, Gonzalez F, Fernandez-Vina M, Tiercy JM, Marsh SG, Aubrey M, Bicalho MG, Canossi A, Carter V, Cate S, Guerini FR, Loiseau P, Martinetti M, Moraes ME, Morales V, Perasaari J, Setterholm M, Sprague M, Tavoularis S, Torres M, Vidal S, Witt C, Wohlwend G, Yang KL. A bioinformatics approach to ascertaining the rarity of HLA alleles. *Tissue Antigens*. 2009; 74:480–485. [PubMed: 19793314]
84. Cadavid LF, Watkins DI. Heirs of the jaguar and the anaconda: HLA, conquest and disease in the indigenous populations of the Americas. *Tissue Antigens*. 1997; 6:702–711. [PubMed: 9458134]
85. Erlich HA, Mack SJ, Bergström 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(1–2):19–24. [PubMed: 9420466]
86. Sokal R, Michener C. A statistical method for evaluating systematic relationships. *Univ Kansas Sci Bull*. 1958; 38:1409–1438.
87. Murtagh F. Complexities of hierarchic clustering algorithms: the state of the art. *Comput Stat Quart*. 1984; 1:101–113.
88. Bergström TF, Josefsson A, Erlich HA, Gyllensten UB. Analysis of intron sequences at the class II HLA-DRB1 locus: implications for the age of allelic diversity. *Hereditas*. 1997; 127(1–2):1–5. [PubMed: 9420464]
89. Bergström TF, Josefsson A, Erlich HA, Gyllensten U. Recent origin of HLA-DRB1 alleles and implications for human evolution. *Nat Genet*. 1998; 18(3):237–242. [PubMed: 9500545]
90. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*. 1985; 39:783–791.
91. Penny D, Hendry MD. Testing methods of evolutionary tree construction. *Cladistics*. 1985; 1:266–278.
92. Cavalli-Sforza, LL.; Menozzi, P.; Piazza, A. The history and geography of human genes. Princeton, NJ: Princeton University Press; 1994.
93. Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. *Nat Genet*. 2003; 33:266–275. [PubMed: 12610536]
94. Mack, SJ.; Erlich, HA. Population relationships as inferred from classical HLA genes. 13th International histocompatibility workshop anthropology/human genetic diversity joint report. In: Hansen, JA., editor. Immunobiology of the human MHC: Proceedings of the 13th international histocompatibility workshop and conference. Vol. I. Seattle: IHWG; 2007. p. 747-757.
95. Nei M. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*. 1978; 89:583–590. [PubMed: 17248844]

Fig. 1.

Sparse cells and the "lumping" of alleles into combined classes. The effect of creating a combined, "lumped" allele class on expected genotype counts is illustrated in the upper and lower halves. In the upper half, expected genotype counts are calculated for all eight alleles; the expected counts for the ten genotypes comprising alleles 5, 6, 7, and 8 (*shaded*) are much less than 1. In the lower half, alleles 5–8 have been lumped into the "5–8" allele class, and no genotype in the table has an expected count less than 0.6.

Fig. 2.

Overview of HLA haplotype estimation and KIR haplotype estimation strategies. The EM algorithm for estimating KIR haplotype frequencies (HFs) can be modified from the standard approach applied to HLA genotypes (*upper-left box*) to account for a reduced combinatorial space using a set of reference haplotypes as an a priori list of known/possible haplotypes to constrain the algorithm (*upper-right box*).

Fig. 3.

Population-level principal component analysis multidimensional scaling plot generated using the supplementary data. Population-level principal component (PC) analysis of the synthetic data in Supplemental Table S8. The internal axes represent values of 0.0 for each component. The first PC in this plot represents 70% of the variance in these data, and the second PC represents 26%. Higher order PCs describe only 4% of the variance and do not need to be presented.

Fig. 4.

Individual-level principal component analysis multidimensional scaling plot generated using the supplementary data. Individual-level principal component (PC) analysis of the synthetic data in Supplemental Table S8. The internal axes represent values of 0.0 for each component. The first PC in this plot represents 21% of the variance in these data, and the second PC represents 19%. Because higher order PCs describe 60% of the variance (evident from the extensive clustering of individuals in the second PC dimension), additional PC dimensions should be presented.

Fig. 5.

Structure bar plot generated using the supplementary data. Structure analysis of the synthetic data in Supplementary Table S9 with the number of clusters (K) set to 2 or 4. Vertical bars represent each individual included in the analysis, and each tone (or color in the electronic version) indicates the extent to which that individual's genotype is derived from one of the K clusters, with each tone (or color) corresponding to a cluster. Because of the low number of loci and the extensive sharing of alleles between populations, very few individuals are assigned to a single cluster (tone/color). However, the relative relatedness of each population can be inferred from the tone/color compositions of their constituents.

Fig. 6.

Examples of phylogenetic trees. Three representations of the same phylogeny for four taxa (A–D). *Black dots* indicate nodes in each tree. The branches between each taxon and the nearest node are known as “twigs” or “leaves”. *Grey dots* in the rooted trees indicate the position of the root node. Taxa A and C are more similar to each other than either is to taxon B or D, and taxa B and C are more similar to each other than either is to taxon A or D. In the unrooted tree and the midpoint rooted tree, A and C are in one clade, and B and D are in a second clade. In the outgroup rooted tree, A, B, and C are in a single clade, to the exclusion of D.

Fig. 7.

Phylogenetic trees generated using the supplementary data. **(a)** Unrooted neighbor-joining tree generated in PHYLIP using Nei's standard genetic distances (included in Supplemental Table S10) generated for the synthetic data in Supplementary Tables S1–S4. Inset bar shows a genetic distance of 0.082. **(b)** Unrooted neighbor-joining tree generated in PHYLIP using population-pairwise F_{ST} distances generated in Arlequin for the synthetic data in Supplementary Table S7. Inset bar shows a population-pairwise F_{ST} distance of 0.007.

Table 1

Matrix of population genetic analyses and available software tools

Analysis	General statistical software (e.g., SAS)	MS Excel	GenAlEx ^a	PyPop ^b	Arlequin ^c	EstiHaplo ^d	PHYLIPE ^e	Structure ^f	CLUTO ^g	PLINK ^h	R package ⁱ	GenePop ^j
Carrier frequency estimation	+	+	++	-	-	-	-	-	-	-	-	-
Hardy-Weinberg	-	-	+	+++	++	-	-	-	+	+	haplo.stats	++
Haplotype estimation	-	-	+	+++	++	+++	-	-	+++	+	-	++
Linkage disequilibrium	-	-	-	+++	++	-	-	-	+++	+	-	++
Measures of selection	-	-	++	+++	++	-	-	-	-	-	-	-
Measurement of genetic differentiation	-	-	++	-	++	-	++	++	+++	-	++	++
Principal component analysis	-	-	++	-	-	-	-	-	-	-	++	-
Phylogenetic analysis	-	-	-	-	-	-	++	+	+++	-	++	-
Population structure analysis	-	-	-	-	++	-	-	++	+++	-	-	-

- = analysis is not possible with that tool; + = analysis is possible, but the tool is not recommended for this analysis; ++ = analysis is possible with this tool; +++ = this tool has been optimized for the analysis of immunogenetic data

^a <http://www.anu.edu.au/BoZo/GenAlEx/>
^b <http://www.pyPop.org>
^c <http://cmpg.unibe.ch/software/arlequin3/>
^d <http://www.methodomics.com/estihaplo>
^e <http://evolution.genetics.washington.edu/phylip.html>
^f <http://pritch.bsd.uchicago.edu/structure.html>
^g <http://glaros.dtc.umn.edu/gkhome/views/cluto/>
^h <http://pngu.mgh.harvard.edu/~purcell/plink/>
ⁱ <http://www.r-project.org/>
^j <http://genepop.curtin.edu.au/>

Table 2

Comparison of methods for determining gene frequencies for present/absent data

Locus	Count	Carrier frequency	Gene frequency estimate from carrier frequency		Direct counting ^a	
			Formula a	Formula b	Count	Gene frequency
Present	109	0.4225	0.2401	0.2398	62	0.2403
Absent	243	0.9419	0.7589	0.7570	196	0.7597
Total ^b	352	1.3643	0.9989	0.9968	258	1

^a Assuming a molecular method that can distinguish homozygotes from heterozygotes

^b $2n = 258$