We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

186,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 1

Statistic and Analytical Strategies for HLA Data

Fang Yuan and Yongzhi Xi

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57493

1. Introduction

To date, the HLA system is the most complex and polymorphic human gene system identified. Although the research history of HLA is not very long, we have made rapid advancements in our understanding of the HLA system during this short time. Research in the HLA field involves elucidating the structure and various biological functions of genes and proteins associated with the HLA system; in addition, it can be directly applied in the study of basic medicine, clinical medicine, anthroposociology, and other fields. HLA research has led to not only revolutionary reforms in basic medical disciplines, such as biology, immunology, heredity, genetics, and anthroposociology, but also unprecedented breakthroughs in many clinical medicine specialties, including organ transplants, oncology, transfusion science, forensic medicine, ecsomatics, genesiology, and vaccination, as well as in disease-related fields of internal medicine. Therefore, it is critical to organize and process HLA study data using appropriate statistical analysis.

Undoubtedly, the proper use of statistics can directly affect the scientific nature, truth, and objectivity of HLA-related studies. Moreover, in addition to the principles and methods of biomedical statistics commonly used in other life sciences, the statistical analysis of HLA study data has its own specific requirements, which integrate the theories and methods of modern bioinformatics. Bioinformatics is a significant research frontier in biomedical statistics and an important field of biomedical research, expanding from macrocosm to microcosm. It integrates numerous methods of biotechnology, computer technology, mathematics, and statistics and is gradually becoming a major discipline yielding discoveries of the secrets of biology, thereby playing an irreplaceable role in organizing and processing relative HLA study data. However, these methods are not within the scope of basic statistical and analytical strategies used for evaluating HLA study data. Thus, due to the limited space and contents of this book, this Chapter will not discuss them. If appropriate, we will describe these methods in a specific chapter of a new monograph about the progress of HLA basic research in the future.

INTECH
open science | open minds

© 2014 The Author(s). Licensee InTech. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Basic concepts of HLA genetic statistics

2.1. Genetics basis for statistical analysis of HLA data

Hardy-Weinberg law: The Hardy-Weinberg law is also referred to as the hereditary equilibrium law or genetic equilibrium law. The basis of the Hardy-Weinberg law is as follows: in an infinite, randomly mating group, when there is no migration, mutation, selection, or genetic drift, the genotype frequency and gene frequency at a locus in the group will remain unchanged generation by generation, achieving a genetic equilibrium state, known as the Hardy-Weinberg equilibrium. This law was proposed by G.H. Hardy, a British mathematician, and W. Weinberg, a German medical scientist, in 1908.

The factors that influence the Hardy-Weinberg equilibrium are as follows:

1. Mutation: Under natural conditions, the rate of gene mutation caused by the reparation effects of DNA replicase is 1×10^{-6}–10^{-8}/gamete/locus/generation in higher animals, demonstrating that the frequency of natural mutation is very low.
2. Selection: a) Reproductive fitness: This is a measure of the ability of providing genes for progeny, i.e., the relative capability of a certain genotype to survive and produce progeny in comparison with other genotypes; in HLA studies, the normal fitness 1 is often used as a reference. b) Heterozygote dominance: In some recessive hereditary diseases and under certain conditions, the heterozygote may be more favorable to survival and progeny reproduction in comparison to homozygous normal individuals.
3. Random genetic drift: The random fluctuation of gene frequency in a small or separated group is referred to as genetic drift.
4. Migration: Gene frequencies may vary among individuals of different races and nationalities. Migration makes different populations intermate, and foreign genes are mutually introduced, which leads to gene flow and thus alters the gene frequency of the original group.
5. Genetic heterogeneity: Individuals with consistent phenotypes or identical clinical symptoms of a specific type of disease may have different genotypes. If they are not strictly distinguished, the Hardy-Weinberg equilibrium will likely become complex.
6. Founder effect: This is a form of genetic drift and refers to a new group established by minor individuals with some alleles of the parent group. The population size of this new group may increase later; however, its gene variance is very small because there is no mating or proliferation between this group and other biological groups. This situation generally occurs in an isolated island or a self-enclosed, newly established village.

Generally, the circumstances meeting the criteria of ideal populations do not exist in practical applications. However, the Hardy-Weinberg equilibrium is still the basis for studies of gene distribution because it is impossible to model all of the factors influencing the investigated group, and various factors can counteract each other (e.g., mutation and selection).

Now we will explain this concept with an example.

Assume that there is an autosomal locus, in brief, alleles A and A′. If the frequencies of genes A and A′ are p_m and q_m in males and p_f and q_f in females, then sperm frequencies with genes A and A′ are p_m and q_m, respectively, and ovum frequencies with genes A and A′ are p_f and q_f, respectively. Obviously, $p_m+q_m=1$ and $p_f+q_f=1$. If mating is completely random, the genotype frequency of the next generation will be as shown in Table 1.

		Sperm	
		A(p_m)	A′(q_m)
Ovum	A(p_f)	AA(p_m*p_f)	AA′(q_m*p_f)
	A′(q_m)	AA′(p_m*q_f)	A′A′(q_m*q_f)

Table 1. Genotype frequencies of progeny generated by random combinations of sperm and ovum

The investigated genes are in autosomes and are unrelated to genotypes; therefore, the frequencies of the three genotypes are identical in male and female progeny. Assume that the frequencies of the three genotypes AA, AA′, and A′A′ are *P*, *Q*, and *R*, respectively. From the table above, we can obtain:

$$P = p_m \times p_f$$

$$Q = q_m \times p_f + p_m \times q_f$$

$$R = q_m \times q_f$$

If we assume that the frequencies of genes A and A′ in progeny are p and q, respectively, then $p+q=1$, $p=P+1/2Q=p_m * p_f+1/2(q_m * p_f+ p_m * q_f)=1/2p_m+1/2p_f$; similarly, $q=1/2q_m+1/2q_f$.

That is to say, when gene frequencies are different between males and females, they will be averaged in the next generation and thus become equal in both sexes. Therefore, when mating is completely random, and selection, mutation, and migration are absent, the gene frequencies and the frequencies of the three genotypes will maintain unchanged generation by generation. If the frequency series of genes A and A′ in gamete is expressed as:

$$(p_A + q_A')$$

then the genotype frequency series in progeny is:

$$\left(p^2{}_{AA} + 2pq_{AA'} + q^2{}_{A'A'}\right)$$

By generalizing the results above, if we assume that the frequencies of n alleles "A_1, A_2…A_n" in a group are p_1, p_2 …p_n, then ($\sum_{i=1}^{n} p_i=1$), and it may be proved that the genotype frequency series in progeny can be expressed as

$$(p_{1A1} + p_{1A1} + \cdots \cdots + p_{nAn})^2$$

This is the presentation formula of the Hardy-Weinberg equilibrium. From this formula, we can see that the frequency of homozygotes AA or A′A′ is equal to the square of the gene

frequency, while the heterozygote frequency is twice the product of the corresponding two gene frequencies. We will explain this concept using ABO blood groups as an example. ABO blood groups are known to be controlled by three alleles A, B, and O, found at the same locus. We can assume that the gene frequencies are p, q, and r, respectively. According to the presentation formula of the Hardy-Weinberg equilibrium, various genotype frequencies of ABO blood groups are expressed with the expansion equation of $(pA+pB+pO)^2$. See the following table.

Phenotype	Genotype	Genotype frequency
A	AA	p^2
	AO	$2pr$
B	BB	q^2
	BO	$2qr$
O	OO	r^2
AB	AB	$2pq$

Table 2. Genotype frequency of ABO blood groups

2.2. Statistical basis of HLA data analysis

2.2.1. Population and sample

The study subjects of HLA statistical data analysis are mostly specific groups, such as individuals with a disease, of the same race, or from the same region, etc. However, due to the limitations of the study method, it is usually impossible to investigate every individual in the group, and the features of the whole group can only be presumed by analyzing some individuals of the group. Thus, two concepts should be defined, i.e., population and sample. The core issue of statistical data analysis is how to deduce the population from a sample.

Population refers to all subjects in a study. The population can also be divided into the infinite population and the limited population. For example, we want to investigate the distribution of a certain HLA phenotype in Asian individuals; because it is difficult to estimate the total number of Asian individuals, we can assume that this population is infinite. Alternatively, if we want to study the recombination characteristics of the HLA system in a specific family, this population is limited. In HLA data analysis, most populations are infinite. Every member constituting the population is referred to as an individual.

A sample is a part of the population, and the number of individuals contained in a sample is the sample size. The core issue of statistical data analysis is that we presume the characteristics of a population from a sample. In order to accurately estimate the population parameters, an appropriate sample size is the foundation of data analysis.

Many factors need to be considered when determining the sample size, such as study objectives, precision, degree of confidence, reliability of statistical testing, sampling method, basic information of the population, study protocol, and study funds. Determination of the appro-

priate sample size fully reflects the repeatability rule in statistical analysis. Now, we will discuss how to determine the sample size in several common cases of HLA statistical analysis.

1. Determination of sample size when estimating population parameters

For example, if we want to understand the distribution of HLA-B*27 in healthy residents of a certain region and the frequency of the HLA-B*27 gene in patients with ankylosing arthritis, how many individuals should be included in the sample? According to the principle of the hypothesis test, if the sample size is too small, then the pre-existing differences cannot be shown; thus, it is hard to obtain correct study results, and the conclusion lacks sufficient basis. Conversely, an oversized sample can increase the practical difficulties of such analyses and unnecessarily waste labor, materials, financial resources, and time; in addition, sample excess may cause inadequate investment and decrease quality control during the scientific research process, thereby introducing potential interference with the study results.

When determining the sample size, the first thing to do is to define the test level or significance level "α", i.e., specifying in advance the allowable probability (α) of false-positive errors in this test (generally, α=0.05); additionally, you should decide whether a one-sided test or two-sided test will be used. The smaller α is, the larger the sample size must be.

The test power should also be defined. The higher the test power is, the larger the sample size must be. The test power is determined by the probability of type-II errors (β). In the design of scientific studies, the test power should be not lower than 0.75; otherwise, it is possible that the test results will not reflect true differences in the population, thereby yielding false-negative results.

In this example, the population represents healthy residents in a certain region, and the individuals investigated in the study constitute the sample. The frequency of gene HLA-B*27 in the population is presumed from the distribution proportion of HLA-B*27 in the sample. If we assume that the distribution frequency of gene HLA-B*27 is P, then the minimal sample (n) meeting the statistical conditions is calculated with the following formula:

When P is close to 0.5:

$$n = \left(\frac{u_{1-\alpha/2}}{\delta}\right)^2 P(1-P)$$

When P is close to 0 or 1:

$$n = \left[\frac{57.3u_{1-\alpha/2}}{\sin^{-1}\left(\delta/\sqrt{P(1-P)}\right)}\right]^2$$

When P is unknown:

$$n = 0.25\left(\frac{u_{1-\alpha/2}}{\delta}\right)^2$$

In the formulas above, u indicates "u distribution", and δ indicates permissible error.

In this example, we want to investigate the frequency of HLA-B*27 in healthy residents of a certain region. We assume that P in the previous investigation is 10%, the permissible error of

this investigation is 1%, and α=0.05 (two-sided), and we can attempt to estimate the number of individuals required for the study. From the critical value form of the u distribution or the u distribution function, we know that $u_{(1-0.05/2)}$=1.96, and the calculated n is about 3457 cases.

2. Estimation of the sample size when comparing the ratio of two populations

When comparing a certain ratio between two populations, for example assessing the differences in the morbidities of cardiovascular diseases between blue collar and white collar workers in a city, HLA analysis often involves determining the distribution differences of a certain gene in diseased and control groups. We can assume that at least two samples with samples sizes of n_1 and n_2, respectively, will be sampled from each population, and the estimated values of the population ratio obtained from the two samples are p_1 and p_2, respectively.

If n_1 is equal to n_2, then:

$$n_1 = n_2 = \frac{[u_{1-\alpha/2}\sqrt{2p(1-p)} + u_{1-\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)}]^2}{(p_1 - p_2)^2}$$

If n_1 is not equal to n_2, then n_2=k×n_1:

$$n_1 = \frac{[u_{1-\alpha/2}\sqrt{2p(1-p)(1+k)/k} + u_{1-\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)/k}]^2}{(p_1 - p_2)^2},$$

where u = u distribution, β = test power, and p = integrated rate of both groups.

2.2.2. Sampling: The process of obtaining a sample from the population

The purpose of sampling is to determine the characteristics of a population by studying a sample (subset) of the population. For example, we want to determine the distribution of genotype HLA-B*07 in a marrow bank from the gene frequencies of 1000 individuals in the bank. This requires that the sample can maximally represent the population features. Therefore, every individual in the population should have the same chance to be sampled, and the sample should be free from bias. For example, in a study investigating a certain HLA phenotype and disease, we generally hope to determine the relationship between the disease and the specific HLA phenotype. In order to do this, researchers must be careful not to deliberately exclude cases without the specific HLA phenotype during sampling. The resulting sample would then not be representative of the total population; this is a bias sample and would not represent the total population profile. The sample we use should be a miniature, accurate representation of the population. In order to achieve this goal, we should use the method of random sampling to obtain samples.

Many randomization methods are commonly used. Initially, drawing lots, casting coins, and casting lots were used; later, researchers adopted random number form, random arrangement form, and the computer-based methods to generate random numbers. For sampling studies in medical science and the grouping of trial subjects, random number form and random arrangement form are relatively convenient. They both perform random sampling and work

out a tool table according to the equal probability principle of mathematical statistics, and the sampling results are better than those obtained by drawing lots or casting coins. Study subjects should be randomly and uniformly assigned into each treatment group (all control and trial groups), thereby preventing various objective factors from intervening with the study results. The greater the number of study subjects, the higher the randomization level. However, it is unnecessary to maximize the amount of study subjects; we should select an appropriate randomization method depending on the trial features. Some common randomization methods are detailed below.

1. Drawing lots: This method is easy to perform. For example, if we want to divide 12 animals into two groups, we should number the animals with 1, 2, 3, …, 12 and prepare the 12 lots, each having a number from 1 to 12. The lots are then mixed, and 6 lots are drawn as per prior specifications; the animals with these 6 lots are assigned into Group 1, and the remaining animals are assigned into Group 2.

2. Random number form: The random number form is carried out according to the principle of random sampling. It can be used for both random assignment and random sampling. All of the numbers in the form are mutually independent. Regardless of horizontal, longitudinal, or slant order, the numbers can randomly occur; therefore, random numbers can be obtained in order by starting from any direction and any location. Some examples are given below.

a. Dividing into two groups: We planned to observe 20 patients with gastric ulcers (patient No. 1–20); one group uses an effective drug ranitidine as a control, and the other group uses a lily decoction. Twenty2-digit random numbers are generated by looking up the random number form, and the random numbers are arranged from small to large, allowing us to obtain the grouping order number "R". If R is between 1 and 10, then the patient is assigned into Group A; if R is between 11 and 20, then the patient is assigned into Group B. The grouping results are presented in Table 3 (reference: TianheXu, Jiu Wang. Design of Medical Experiments: Lecture 2 – Rules of randomization and blinding method. *Chinese Medical Journal*, 2005, 40(8): p.54).

Patient No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Random number	93	22	53	64	39	7	10	63	76	35	87	3	4	79	88	8	13	85	51	34
Grouping order number (R)	20	7	12	14	10	3	5	13	15	9	18	1	2	16	19	4	6	17	11	8
Group	B	A	B	B	A	A	A	B	B	A	B	A	A	B	B	A	A	B	B	A

Table 3. Randomized grouping results of 20 patients

b. Randomized division of three or more groups: If we want to randomly divide 15 animals into 3 groups, we should number the animals from 1 to 15.Then, fifteen 2-digit random numbers are generated by looking up the random number form, and the random numbers should be arranged from small to large. The order number "R" can then be obtained. If R is between 1 and 5, then the animal is assigned into Group A. If R is between 6 and 10,

then the animal is assigned to Group B. If R is between 11 and 15, then the animal is assigned to Group C. The grouping results are presented in Table 4 (TianheXu, Jiu Wang. Design of Medical Experiments: Lecture 2 – Rules of randomization and blinding method. *Chinese Medical Journal*, 2005, 40(8): p.54).

Animal No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Random number	33	35	72	67	47	77	34	55	45	70	8	18	27	38	90
Grouping order number (R)	4	6	13	11	9	14	5	10	8	12	1	2	3	7	15
Group	A	B	C	C	B	C	A	B	B	C	A	A	A	B	C

Table 4. Randomized grouping of 15 animals

2.3. Definitions of relative terms in HLA statistical data analysis

2.3.1. Definitions of HLA phenotype, haplotype, and genotype

HLA antigens have their own allele code on the chromosome; generally, the HLA antibody-antigen specificity of an individual can be detected using available typing reagents and committed cells. The antigen-specific type obtained by this method is referred to as the phenotype. However, the antigen phenotype does not reflect the individual's allele combination pattern on the chromosome. The combination of HLA alleles on the chromosome is referred to as the haplotype. If this combination expands from type-I and type-II alleles to type-III genes or adjacent loci, it is often referred to as an extended haplotype. Two haplotypes form the HLA genotype of an individual, i.e., the pattern of the HLA allele combination on two chromosomes in the individual (Table 5). Generally, the haplotype and genotype can only be determined by performing phenotype analysis of all the members of a family or by using special experimental methods, such as monospermal analysis. The phenotype of every individual has many potential combinations that depend on different genotypes. It is therefore important to understand an individual's haplotype and genotype in allogeneic organ transplant, transplantation of hematopoietic stem cells, and forensic identification.

Individual	**Individual 1**	**Individual 2**	**Individual 3**
Typing results	A*11:01 A*24:01 B*07:02 B*27:04	A*11:01 B*07:02 B*27:04	A*11:01 B*27:04
Phenotype	HLA-A11, A24, B7, B27	HLA-A11, B7, 27	HLA-A11, B27
Genotype	HLA-A*11:01,A*24:01 HLA-B*07:02, B*27:04	HLA-A*11:01, A*11:01 HLA-B*07:02, B*27:04	HLA-A*11:01, A*11:01 HLA-B*27:04, B*27:04
Haplotype	HLA-A*11:01, B*07:02 & HLA-A*24:01, B*27:04 or HLA- A*11:01, B*27:04 & HLA-A*24:01, B*07:02	HLA- A*11:01, B*07:02 & HLA- A*11:01, B*27:04	HLA- A*11:01, B*27:04 & HLA- A*11:01, B*27:04

Table 5. Differences in HLA phenotypes, haplotypes, and genotypes

2.3.2. Genetics features of HLA

1. Haplotype genetic mode: An HLA complex is a group of closely linked genes. Crossing-over between homologous chromosomes rarely occurs in these alleles, which are linked in the same chromosome, i.e., they form a haplotype. During reproduction, the HLA haplotype is inherited from parent to progeny as a complete genetic unit. The progeny can randomly obtain an HLA haplotype from both parents, thereby forming the progeny's new genotype. In the siblings of the same family, the probability of having two identical haplotypes is 25%, the probability of having one identical haplotype is 50%, and the probability of having two different haplotypes is 25%. Therefore, when seeking an appropriate donor for allogeneic organ transplant or transplant of hematopoietic stem cells in clinical practice, it is much easier to find the matched HLA antigen (the matched HLA haplotype in particular) in the patient's family than in nonsibling donors. However, it should be noted that when the haplotype is inherited from parent to progeny, homologous crossing-over between both haplotypes may occur (see details in the recombination section).

2. Codominant inheritance: This means that antigens encoded by each pair of alleles are expressed on the cell membrane, and there is no recessive gene. Allele rejection does not exist. If the haplotypes of an individual's two chromosomes are HLA-A*11:01, B*27:04 and HLA-A*24:01, B*07:02, then four different HLA molecules, A11, A24, B27, and B7, will be expressed on the cytomembrane surface of the individual.

3. Linkage disequilibrium: Various HLA alleles at different loci occur in the group at a specific frequency. In a group, if the frequencies of two alleles at different loci occurring in the same chromosome are higher than the expected random frequencies, i.e.,the haplotype frequency (observed data) is significantly higher or lower than theoretical value (the product of allele frequencies at different loci), then this non-free combination phenomenon is referred to as linkage disequilibrium. For example, A1 and B8 in Caucasians and A2 and B46 in southern Chinese individuals always occur together, and the resulting haplotypes A1-B8 and A2-B46, respectively, exhibit linkage disequilibrium.

3. Estimation of HLA population genetic parameters

3.1. Genetic structure

Studies of the genetic parameters of the HLA system actually start from the loci "HLA-A" and "HLA-B"; these two loci are often used as examples in HLA data analysis. To provide a simple description, we will expand this model to an autosomal double-loci multiple-alleles genetic model and use the following symbols.

Assume that there are two linkage loci (I and J) on a human chromosome; each locus has multiple codominant alleles. The alleles at locus I are labeled as i_1, i_2, i_3, i_n, and i_0. "i_0" represents the undetected blank gene at locus I; therefore, the allele number at locus I is calculated as $l = n + 1$.

Similarly, the alleles at locus J are expressed as j_1, j_2, j_3,…j_m and j_0. "j_0" represents the undetected blank gene at locus J; therefore, the allele number at locus J is calculated as $k = m + 1$.

Because the alleles at loci I and J can randomly combine, the number of all the possible haplotypes is $l*k$. These haplotypes can form various genotypes, calculated as $lk(lk+1)/2$. Considering that the phenotype of genes in a homozygosis state is the same as the phenotype of a gene hybridized with a blank gene, i.e., the phenotype of genotypes i_2i_2 and i_2i_0 is i_2 (+), the number of all possible phenotypes from any allele combination at locus I is $[l+(l-1) \times (l-2)/2]$; similarly, the number of all possible phenotypes from any allele combination at locus J is $[k+(k-1) \times (k-2)/2]$. Thus, the number of all possible phenotypes at loci I and J is $[l+(l-1) \times (l-2)/2] \times [k+(k-1) \times (k-2)/2]$.

Loci I and J are linked, so the gene frequency of each locus correlates to the frequency of haplotypes formed with genes at both loci; the antigen distribution in the population also correlates at both loci. This relation can be fully shown in a 2 × 2 four-space form. Unless specified otherwise, all the populations mentioned in this chapter are Mendelian populations achieving Hardy-Weinberg equilibrium, i.e., this population undergoes completely random mating, and there are no effects of selection, mutation, or migration.

The population distribution of antigens at both loci is presented in the table below. Symbols in the table indicate that in a population with a total number of individuals "*N*", "a" individuals have antigens i and j, "b" individuals have antigen i but without antigen j, "c" individuals have antigen j but without antigen i, and "d" individuals do not have antigens i and j. The marginal values A, B, C, and D in this table are respectively equal to the sum of the corresponding two spaces, and *N* is equal to the sum of four spaces.

		Antigen j		Total
		+	-	
Antigen i	+	a	b	C=a+b
	-	c	d	D=c+d
Total		A=a+c	B=b+d	***N***=a+b+c+d

Table 6. Population distribution of antigens i and j

Table 6 shows the relation between the frequencies of genes i and j and the haplotype frequency. The genes at loci I and J can form four haplotypes, ij, j0, i0, and 00; "0" represents the blank gene. The frequencies of the four genes are expressed as *s*, *t*, *u*, and *v* respectively. The frequency of gene i is expressed as "p_i", and the frequency sum of the other alleles at locus I is expressed as "q_i"; obviously, $p_i+q_i=1$. Similarly, the frequency of gene j and the frequency sum of the other alleles at locus J are expressed as "p_j" and "q_j", respectively. From the table below, we can see that the frequency of each gene can be expressed as the frequency sum of the corresponding haplotypes.

		Gene j		Total
		+	-	
Gene i	+	*s*	*u*	$p_i=s+u$
	-	*t*	*v*	$q_i=t+v$
Total		$p_j=s+t$	$q_j=u+v$	1

Table 7. Relation between the gene frequencies at loci I and J and the haplotype frequencies

3.2. Hardy-Weinberg equilibrium test

According to the gene or haplotype frequency, the expected values of all genotype frequencies and phenotype frequencies can be obtained by combination as per the Hardy-Weinberg equilibrium law; the coincidence degree of the expected value and the corresponding actual observed value is referred to as the Hardy-Weinberg coincidence test. This test is mainly used in two cases: 1) As a prompt for supporting or excluding a certain genetic mode. For example, in an assumed Mendelian genetic system, the gene or haplotype frequency is calculated on the basis of the assumed genetic mode, and recombination is then performed as per the Hardy-Weinberg equilibrium law to obtain the expected value of the phenotype. If the expected value coincides with the observed value of this phenotype, the genetic mode may be true; otherwise, the genetic mode may be excluded. The conclusion obtained by application of the Hardy-Weinberg equilibrium to test a genetic mode cannot be confirmatory because sometimes increasing the assumed loci may give better coincidence results. 2) For reliability estimation of the population survey data. For some genetic systems with well-established genetic modes, such as the HLA system discussed in this book, if the population can perform fully random mating, and there are no effects caused by selection, mutation, or migration, the population distribution should be in good Hardy-Weinberg equilibrium. Poor coincidence of both values shows that the population survey data are not reliable, which can help us identify the causes of errors in aspects sampling, typing technology, etc.

3.2.1. Measuring method for determination of the coincidence degree

The coincidence degree of a phenotype's expected value and observed value is generally measured with χ^2. The χ^2 is calculated for every phenotype, and the values are added to obtain the total χ^2. The P value is calculated by looking up the form. The χ^2 calculation formula is:

$$\chi^2=\sum \frac{(Expected\ value - Observed\ value)^2}{Expected\ value}$$

In the Hardy-Weinberg equilibrium test, the expected value of the phenotype is often less than 5. In this case, some authors will incorporate several phenotypes and calculate χ^2 again when the phenotype value is more than 5; however, this method has obvious subjective factors, and the calculated χ^2 value after incorporation will be reduced. In addition, due to variations of incorporation methods, it is difficult to compare data between studies. Therefore, we think that it is unnecessary to incorporate items with the phenotype expected values of less than 5,

and the χ^2 should be calculated as in other cases; although this may increase theχ^2 value, the resulting coincidence conclusion is more reliable.

Determination of the degrees of freedom in the χ^2 test: Assume that a genetic system consists of n alleles and Φ phenotypes, and the sample size is N. Because the gene frequencies $p_1+p_2+\ldots\ldots p_n=1$, the number of parameters estimated from the sample is (n-1); in addition, a degree of freedom is lost because the sample is too small. Therefore, the degrees of freedom remaining for other tests are:

$$d_f = \Phi - (n-1) - 1 = \Phi - n$$

In the Hardy-Weinberg equilibrium, $p \geq 0.5$ is generally used as the criterion to judge whether there are significant differences between the expected and observed values.

3.2.2. Hardy-Weinberg equilibrium test for separated loci

In a genetic system containing one or more loci, the Hardy-Weinberg equilibrium test can be performed for every locus. According to the Hardy-Weinberg equilibrium law, the expected frequency of the homozygous genotype is the product of the corresponding two gene frequencies, and the expected frequency of the heterozygous genotype is twice the product of the corresponding gene frequencies. The expected value of each phenotype is equal to the sum of the expected values of the corresponding genotypes. After multiplying the phenotype frequency by sample size "N" to calculate the expected value of the phenotype, the coincidence degree between this expected value and the observed value of the phenotype can be tested.

The table (Table 8) below shows a Hardy-Weinberg coincidence test of the antigen phenotype at locus HLA-C in Chinese individuals with the Han nationality. The expected values and observed values coincide well, demonstrating that the distribution of these alleles at locus C is in the Hardy-Weinberg equilibrium state. In this table, the HLA-Cw1 phenotype includes two genotypes, "HLA-C*01/HLA-C*01" and "HLA-C*01/blank"; therefore, the expected value of the phenotype is calculated as

$$106 \times (0.1442 \times 0.1442 + 0.1442\ 0.3793 \times 2) = 13.5779.$$

The phenotype expected value of HLA-Cw1, 2 is calculated as

$$106 \times 0.1442 \times 0.0143 \times 2 = 0.4311.$$

3.2.3. Hardy-Weinberg equilibrium test of haplotypes

The haplotype Hardy-Weinberg equilibrium test can be performed in a multiple-loci, multiple-alleles genetic system, and the allelic and linkage relationships of all the genes in the system can also be tested. If not considering recombination, haplotypes and alleles also comply with the same genetic rules; therefore, according to Hardy-Weinberg equilibrium, the expected value of the phenotype containing two identical haplotypes should be equal to the square of the haplotype frequency, and the expected value of the phenotype containing two different haplotypes should be equal to twice the product of the frequencies of the two haplotypes. the expected value of the phenotype can be calculated by sorting various haplotype frequencies

http://dx.doi.org/10.5772/57493

Phenotype	Genotype	Observed value	Expected value	Gene frequency
HLA Cw1	HLA-C*01/HLA-C*01 or HLA-C*01/blank	10	13.5779	HLA-C*01=0.1422
HLA Cw1,2	HLA-C*01/HLA-C*02	0	0.4311	HLA-C*02=0.0143
HLA Cw1,3	HLA-C*01/HLA-C*03	17	11.6275	HLA-C*03=0.3857
HLA Cw1,4	HLA-C*01/HLA-C*04	1	2.3665	HLA-C*04=0.0785
HLA Cw1,5	HLA-C*01/HLA-C*05	0	0	HLA-C*05=0
HLA Cw2	HLA-C*02/HLA-C*02 or HLA-C*02/blank	1	1.1716	HLA-C*blank=0.3793
HLA Cw2,3	HLA-C*02/HLA-C*03	1	1.1693	
HLA Cw2,4	HLA-C*02/HLA-C*04	1	0.2380	d_f=16-6=10
HLA Cw2,5	HLA-C*02/HLA-C*05	0	0	P>0.5
HLA Cw3	HLA-C*03/HLA-C*03 or HLA-C*03/blank	43	46.788	
HLA Cw3,4	HLA-C*03/HLA-C*04	5	6.4188	
HLA Cw3,5	HLA-C*03/HLA-C*05	0	0	
HLA Cw4	HLA-C*04/HLA-C*04 or HLA-C*04/blank	9	6.9655	
HLA Cw4,5	HLA-C*04/HLA-C*05	0	0	
HLA Cw5	HLA-C*05/HLA-C*05 or HLA-C*05/blank	0	0	
Blank	Blank/blank	18	15.2501	
Total		106	106.004	

Table 8. Hardy-Weinberg equilibrium test of the antigen phenotype at locus HLA-C in Chinese individuals with the Han nationality

into the corresponding phenotypes, and a coincidence test is then performed with the observed value of the phenotype. Table 9 shows the calculation method for the expected value of phenotype "HLA-A2, B15", where A* is blank, representing the set of all alleles at locus HLA-A except HLA-A*02, and B* is blank, representing the set of all alleles at locus HLA-B except HLA-B*15. Haplotype frequencies would be calculated as HLA-A*02 B*15=0.0113; HLA-A*02 B*blank=0.0559; HLA-A*blank B*15=0.0098; HLA-A*blank B*blank=0.0053.

Haplotype combination mode	Phenotype frequency
A*02 B*blank /A*blank B*15	2 × 0.0559 × 0.0098=0.001096
A*02 B*15 /A*blank B*blank	2 × 0.0113 × 0.0053=0.000120
A*02 B*15 /A*blank B*15	2 × 0.0113 × 0.0098=0.000221
A*02 B*15 /A*02 B*blank	2 × 0.0113 × 0.0059=0.000133
A*02 B*15 /A*02 B*15	0.0113 × 0.0113 × 0.000128

Table 9. The haplotype composition and expected frequency of phenotype "HLA-A2, B15"

3.3. Estimation of genetic parameters

Genetic parameters are estimated by assessing a quantity-limited sample, and thus, sampling error must exist. The size of sampling error is expressed with the standard error σ, where σ is equal to the root extraction of variance "*V*".

3.4. Antigen frequency

Antigen frequency is defined as the ratio or percentage of individuals with the antigen phenotype in the population. If *N* = total individuals and *C* = individuals with the antigen phenotype i, then the frequency of antigen i is calculated as:

$$f_i = C / N$$

The frequencies of antigens i and j can be easily obtained from the four-space form above:

$$f_i = (a + b) / N = C / N$$

$$f_j = (a + c) / N = A / N$$

and the standard error is calculated as: $\sigma fi = \frac{1}{N}\sqrt{\frac{CD}{N}} = \sqrt{\frac{fi*(1 - fi)}{N}}$

When the antigen frequency f_i is fixed, the greater the sample size *N*, the lower the standard error.

3.5. Gene frequency

Assuming that i_1 is an allele at locus I, the ratio or percentage of gene i_1 in all the genes of this locus is referred to as the gene or genotype frequency of i_1. The frequency sum of all alleles at a single locus is 1. Gene frequency can be obtained from family or population surveys.

3.5.1. Calculation of gene frequency by direct genotype count

If the genotyping results of *N* individuals are known, then the frequency of gene i in the population can be obtained by a simple counting method. Assuming that this value is *X*, the frequency of gene i is:

$$p_i = X / 2N$$

Assume that there are two alleles at autosomal locus I, i_1 and i_2. For diploids, it is possible to form three genotypes: i_1i_1, i_1i_2, and i_2i_2. After surveying 100 individuals, possible count results are presented per genotype in the following table (Table 10). In total, there are 200 genes at locus I: 36 individuals have i_1i_1, and the count of gene i_1 is 72; 48 individuals have i_1i_2, and the count of genes i_1 and i_2 is 48; 16 individuals have i_2i_2, and the count of gene i_2 is 32. Therefore, the gene frequency of i_1 is calculated as (72+48)/200=0.6; similarly, the gene frequency of i_2 is calculated as (32+48)/200=0.4; the sum of both frequencies is 1.

In the HLA system, each locus usually has several alleles. If we want to calculate the frequency of a certain allele at the locus, which can be expressed as i1, then the meaning of i_2 is non-i_1

genes. For example, if we want to calculate the gene frequency of HLA-B*27 at locus HLA-B, which is expressed as i_1, then the frequency sum of all the alleles except HLA-B*27 is expressed as i_2.

Genotype	i_1i_1	i_1i_2	i_2i_2
Observed individuals	36	48	16
Amount of gene i_1	72	48	0
Amount of gene i_2	0	48	32

Table 10. Distribution of genotype I in 100 random individuals

3.5.2. Estimation of gene frequency according to phenotype frequency

Currently, HLA typing technology is developing rapidly. With the popularization of high-throughput sequencing technology, the calculation of HLA gene frequency can be mostly completed using the counting method. However, due to limited technical conditions in some population surveys, we can only obtain the corresponding phenotypes. Therefore, how should be best analyze gene frequency? There are two main methods used in practical work: one is the root method, which involved simple arithmetic and easy to perform; the other is the maximal likelihood algorithm, which is highly efficient in estimating gene frequencies, but required specialized computer software (see details in the next section). A description of how to use the root method for estimation of gene frequency according to phenotype results is given below.

If the frequency of the dominant gene i is p_i, then the frequency sum of all the other alleles at this locus is

$q_i = 1 - p_i$;

The relationship between the phenotype frequency and the corresponding genotype frequency can be obtained according to the Hardy-Weinberg law (see the table below).

Phenotype	Phenotype frequency	Corresponding genotype	Frequency of corresponding genotype
I (+)	f_i	i homozygote "ii", i heterozygote "i-"	$p_i^2+2p_iq_i$
I (-)	$1-f_i$	Non-i combination "-/-"	q_i^2

Table 11. Relationship between phenotype and genotype frequencies

We can deduce from the table

$p_i = 1 - \sqrt{1 - fi}$

where p_i is the gene frequency of gene I, and f_i is the frequency of the phenotype or antigen containing gene i. This formula is often used for estimation of HLA gene frequency, and its form can be changed.

$$p_i = 1 - \sqrt{1 - C / N}$$

Or $p_i = 1 - \sqrt{D / N}$

The definitions of *C*, *D*, and *N* in the formula are the same as above, and the standard error of p_i is:

$$\sigma p_i = \frac{1}{2}\sqrt{\frac{f_i}{N}} = \frac{\sqrt{N - D}}{2N}$$

It should be noted that when the pi value is small, it can be calculated by the following formula:

$$p_i \approx \frac{f_i}{2}$$

3.6. Haplotype frequency

For double-loci multiple-allele genetic systems, each chromosome has two alleles belonging to two different loci, and the combination of these different alleles forms variant haplotypes. The ratio or percentage of each haplotype in the population is referred to as the frequency of this haplotype. The sum of all haplotype frequencies is 1.

3.6.1. Calculation of haplotype frequency by direct haplotype count

When an individual's haplotype is known, the haplotype frequency can be calculated by a simple counting method, and the calculation method and technology are the same as those for calculation of gene frequency. However, haplotypes can often only be obtained by family surveys, and HLA haplotypes cannot be fully determined in some families. During data analysis, rejection of these individuals may cause error. In this case, the relative haplotype frequency can be estimated by referring to the population survey results. For example, in Table 12 below, whether the mother's haplotype is A9-B13/A2-B13 or A9-B13/A2-B- cannot be fully determined by family analysis. The haplotype can only be determined by estimation of relative frequency. Assume we know that the frequency of haplotype A2-B13 is 0.0356 and that of haplotype A2-B- is 0.0559 from population survey data. Because the mother can only have these two haplotypes, the relative frequency of A2-B13 is 0.0356/(0.0356+ 0.0559)=0.39 and that of A2-B- is 0.0559/(0.0356+ 0.0559)=0.61. During counting, these haplotypes should be counted as 0.39 A2-B13 and 0.61 A2-B-, respectively.

In practical applications, due to advances in HLA genotyping methods, especially the widespread use of sequencing-based typing methods, high-resolution HLA results are comprehensively adopted; when there is only one allele that is detected at the locus of a certain gene, it is often considered a homozygous allele.

	Phenotype	Haplotype combination mode
Father	A10,11; B5,15	A11-B5 & A10-B15
Mother	A2,9; B13	A9-B13 % A2-B13 or A9-B13 % A2-B-
Child 1	A9,11; B5,13	A11-B5 & A9-B13
Child 2	A9,10; B13,15	A10-B15 & A9-B13
Child 3	A9,11; B5,13	A11-B5 & A9-B13

Table 12. A family's HLA typing results

3.6.2. Estimation of haplotype frequency according to phenotype frequency

Assume that codominant genes i and j are located at two different loci, and other blank genes at these two loci are expressed as 0. Therefore, it is possible to have four haplotypes (ij, j0, i0, and 00) in the population, and the frequencies of these four haplotypes are expressed as s, t, u, and v, respectively. The relationship between haplotype frequency and gene frequency is illustrated in Table 7. The actual observed value of the distribution of antigen ij in the population is presented in Table 6.

The expected distribution of antigens i and j in the population can be expressed with the pattern in Table 13, and the expected values in the table are obtained according to Hardy-Weinberg equilibrium. Thus, we can expand $(s+t+u+v)^2$ and then incorporate items with identical phenotypes. N is the sample size.

		Antigen j		Total
		+	-	
Antigen i	+	$(2s-s^2+2tu)N$	$(u^2+2uv)N$	$N[1-(t+v)]^2$
	-	$(t^2+2tv)N$	v^2N	$N(t+v)^2$
Total		$N[1-(u+v)^2]$	$N(u+v)^2$	N

Table 13. Expected values related to the distribution of antigen ij

After changing the formula, the gene frequency is expressed as:

Frequency of haplotype ij, $s = p_j - q_i + \sqrt{d / N}$

Frequency of haplotype j0, $t = q_i - \sqrt{d / N}$

Frequency of haplotype i0, $u = q_j - \sqrt{d / N}$

Frequency of haplotype 00, $v = \sqrt{d / N}$

If expressing as antigen frequency,

$$s = \sqrt{d/N} + 1 - \sqrt{1 - f_j} - \sqrt{1 - f_i}$$

$$t = \sqrt{1 - f_i} - \sqrt{d/N}$$

$$u = \sqrt{1 - f_j} - \sqrt{d/N}$$

$$v = \sqrt{d/N}$$

If expressing as phenotype amount,

$$s = \sqrt{d/N} + 1 - \sqrt{B/N} - \sqrt{D/N}$$

$$t = \sqrt{D/N} - \sqrt{d/N}$$

$$u = \sqrt{B/N} - \sqrt{d/N}$$

$$v = \sqrt{d/N}$$

The standard error for this equation can be calculated as:

$$\sigma s = \sqrt{\left[\left(1 - \sqrt{d/B}\right)\left(1 - \sqrt{d/D}\right) + s - s^2/2\right]/2N}$$

$$\sigma t = \sqrt{\left[\left(1 - \sqrt{d/D}\right) - t^2/2\right]/2N}$$

$$\sigma u = \sqrt{\left[\left(1 - \sqrt{d/B}\right) - u^2/2\right]/2N}$$

$$\sigma v = \frac{1}{2}\sqrt{(1 - v^2)/N}$$

and the standard error of "*s*" can be expressed as:

$$\sigma s = \sqrt{[(1 - v/(t+v))(1 - v/(u+v)) + s - s^2/2]/2} = \sqrt{\left[\left(\frac{p_i - s}{1 - p_j}\right)\left(\frac{p_j - s}{1 - p_i}\right) + s - s^2/2\right]/2}$$

If a haplotype contains a blank gene, such as A1-B-, then the frequency is equal to the sum of the gene frequency of A1 and the haplotype frequencies of the other alleles at locus B. The haplotype frequency calculated according to the above calculation formula for phenotype data may be a negative value sometimes; this is caused by inadequate sample size and sampling error.

3.7. Linkage disequilibrium parameter

3.7.1. Linkage disequilibrium

Linkage disequilibrium is controlled by inconsistency of the observed and expected values about the appearance of antigens at two linked loci in the same haplotype. Assuming that the genes at two linked loci are i and j, the linkage disequilibrium parameter is defined as the difference between the actual observed haplotype frequency of ij and the product of gene

http://dx.doi.org/10.5772/57493

frequencies of i and j, which is expressed as Δ. If the observed frequency of haplotype ij is "s", then $\Delta_{ij}=s-p_ip_j$.

When the genotypes and haplotypes of every individual in the population are known, the Δ value can be easily obtained by the counting method. However, it is generally necessary to estimate the Δ value from population survey data, i.e., $\Delta_{ij}=s-p_ip_j=sv-tu$.

The following formulas are commonly used:

$$\Delta ij=\sqrt{d/N}-q_iq_j$$

$$\Delta ij=\sqrt{d/N}-\sqrt{(1-f_j)(1-f_i)}$$

$$\Delta ij=\sqrt{d/N}-\sqrt{\mathrm{BD}}/N$$

The standard error of Δ can be calculated as:

$$\sigma\left(\Delta\right)=\sqrt{\frac{\alpha}{4N^2}-\frac{\Delta}{N}\left(\frac{B+D}{2\sqrt{BD}}-\frac{\sqrt{BD}}{N}\right)}$$

or as:

$$\sigma(\Delta)=\sqrt{\frac{1}{4N}+\frac{1}{4N^2}(\mathrm{B}+\mathrm{D}+\mathrm{d})-\frac{\sqrt{d}}{2N\sqrt{N}}\left(\sqrt{\frac{D}{B}}+\sqrt{\frac{B}{D}}\right)-\frac{BD}{N^3}+\frac{\sqrt{BDd}}{N^2\sqrt{N}}}$$

$\Delta / \sigma(\Delta) \geq 1.96$ is generally considered to be of significant linkage disequilibrium.

3.7.2. Relative Δ value

There is no comparability among absolute Δ values, so relative Δ values, i.e., "$\Delta_{(r)}$", are generally calculated for comparison.

$$\Delta_{ij(r)}=\Delta_{ij(r)}/\Delta_{ij(Max)}$$

From the calculation formula of

$$\Delta_{ij}=s-p_ip_j=sv-tu$$

we can see:

If tu=0, Δ_{ij} has the maximal value; if t or u is 0, Δ_{ij} is equal to p_jq_i or p_iq_j, and the lower value between p_i and p_j is used to calculate $\Delta_{ij(\mathrm{Max})}$, i.e.:

$p_i<q_j: \Delta_{ij(Max)}=p_i(1-p_j)$;

$p_i>q_j: \Delta_{ij(\mathrm{Max})}=p_j(1-p_i)$;

If s = 0, the negative Δ_{ij} has the maximal value, and

$$\Delta_{ij(Max)}=-p_jp_i.$$

3.8. Genetic distance

In order to quantitatively describe the process of generating genetic differences between two populations due to selection, mutation, migration, and random drift, the concept of genetic distance has been introduced.genetic distance is a measure of the gene frequency differences between populations, and it is used to describe interpopulation variance.

In 1977, Cavalli-Sforza and Bodmerthe defined the genetic distance (d) as:

$$d = \sqrt{1 - \sum_i \sqrt{p_{i1} p_{i2}}}$$

where p_{i1} and p_{i2} are the frequencies of gene i in populations 1 and 2, respectively.

4. Software analysis of HLA data

To conveniently implement haplotype frequency estimation, linkage disequilibrium, Hardy-Weinberg equilibrium, pairwise genetic distances, etc., of HLA data, computer software is usually required. There are severalprofessional statistical software and genetic analysis software programs.This chapter will introduce some common problems encountered when using software for HLA data analysis.

4.1. The processing method of HLA data analysis using *Arlequin* software

Arlequin is the French translation of "Arlecchino," a famous Italian character from "Commedia dell'Arte." Arlecchinois a multi-faceted character, but he has the ability to switch among his various character assets very easily according to his needs and necessities. This polymorphic ability is symbolized by his colorful costume, from which the *Arlequin* icon was designed (Figure 1).

The goal of *Arlequin* is to provide the average user in population genetics with a large set of basic methods and statistical tests to extract information on genetic and demographic features of a collection of population samples.

Arlequin can handle several types of data either in haplotypic or genotypic form.

- DNA sequences
- RFLP data
- Microsatellite data
- Standard data
- Allele frequency data

HLA data belong to "Standard data" in which the molecular basis of a polymorphism is not defined specifically, or when different alleles are considered mutationally equidistant from

http://dx.doi.org/10.5772/57493

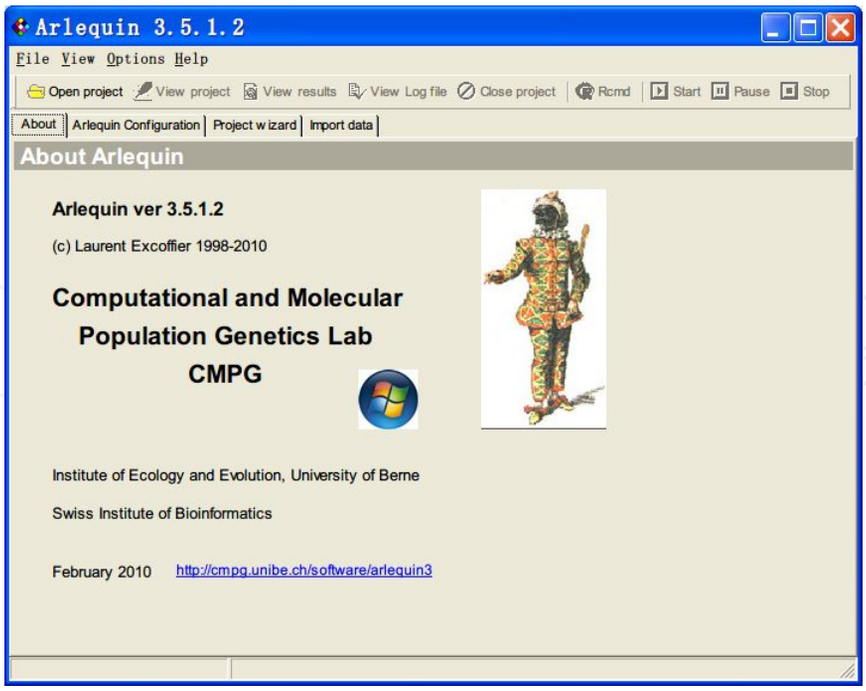

Figure 1. *Arlequin* software

each other. Therefore, standard data haplotypes are compared for their content at each locus, without regarding the nature of the alleles, which can either be similar or different.

4.1.1. Structure of an Arlequin input file

4.1.1.1. Input data file

The first step for the analysis of your data isto prepare an input data file (project file) for *Arlequin*. Because *Arlequin* is a versatile program that is able to analyze several data types, you must include information about the property of your data together with the raw data into the project file. A text editor can be used to define your data using reserved keywords.

Arlequin project files contain a description of the data properties as well as the raw data themselves. The project file may also refer to one or more external data files.

Input files are structured into two main sections with additional subsections that must appear in the following order (Figure 2):

1) Profile section (mandatory)

2) Data section (mandatory)

2a) Haplotype list (optional)

2b) Distance matrices (optional)

2c) Samples (mandatory)

2d) Genetic structure (optional)

2e) Mantel tests (optional)

4.1.1.2. Profile section

The data properties must be described in the profile section. The beginning of the profile section is indicated by the keyword [Profile] (within brackets).

The user must specify the following parameters:

- The title of the current project (used to describe the current analysis)
  - Notation: Title=
  - Possible value: Any string of characters within double quotation marks
  - Example: Title="An analysis of haplotype frequencies in two populations"
- The number of samples or populations present in the current project
  - Notation: NbSamples =
  - Example: NbSamples =3
  - The type of data to be analyzed. Only one type of data isallowed per project.
  - Notation: DataType =
  - Possible values: DNA, RFLP, MICROSAT, STANDARD, and FREQUENCY
  - Example: DataType = DNA
  - The parameter of “STANDARD” is used here when dealing with HLA Data.
- Thetype of data that the project addresses
  - Notation: GenotypicData =
  - Possible values: 0 (haplotypic data), 1 (genotypic data)
  - Example: GenotypicData = 0

This parameter is used to demonstrate whether haplotypic or genotypicdata are being used for the HLA data analysis.Unless specified, the parameter used here is usually “1.”

Additionally, the user has the option to specify the following parameters:

- The character used to separate the alleles at different loci (the locus separator)
  - Notation: LocusSeparator =
  - Possible values: WHITESPACE, TAB, NONE, any character other than "#", or a character
  - specifying missing data
  - Example: LocusSeparator = TAB
  - Default value: WHITESPACE

- Thegametic phase of the genotype
  - Notation: GameticPhase =
  - Possible values: 0 (unknown gametic phase), 1 (known gametic phase)
  - Example: GameticPhase = 1
  - Default value: 1
  - For general HLA data analysis, the parameter is “0.” If approaches such as pedigree analysis are used, and the HLA haplotype of each individual sample are given, “1” is used as the parameter.In the data input, one haplotype should be entered in the same row.
- Indication of a recessive allele
  - Notation: RecessiveData =
  - Possible values: 0 (co-dominant data), 1 (recessive data)
  - Example: RecessiveData =1
  - Default value: 0

Because the HLA loci are codominant, “1” is used as the parameter when dealing with HLA Data

- The code for the recessive allele
  - Notation: RecessiveAllele =
  - Possible values: Any string of characters within double quotation marks. This character string can be used explicitly in the input file to indicate the occurrence of a recessive homozygote at one orseveral loci.
  - Example: RecessiveAllele ="xxx"
  - Default value: "null"
- The character used to code for missing data
  - Notation: MissingData =
  - Possible values: A character used to specify the code for missing data, which can be
  - entered between single or double quotes.
  - Example: MissingData ='$'
  - Default value: '?'
- The absolute or relative values of haplotype or phenotype frequencies
  - Notation: Frequency =

- Possible values: ABS (absolute values), REL (relative values: absolute values will be found by multiplying the relative frequencies by the sample sizes)
- Example: Frequency = ABS
- Default value: ABS

- The number of significant digits for haplotype frequency outputs
  - Notation: FrequencyThreshold =
  - Possible values: A real number between 1e-2 and 1e-7
  - Example: FrequencyThreshold = 0.00001
  - Default value: 1e-5
- The convergence criterion for the EM algorithm used to estimate haplotype frequencies and linkage disequilibrium from genotypic data
  - Notation: EpsilonValue =
  - Possible values: A real number between 1e-7 and 1e-12.
  - Example: EpsilonValue = 1e-10
  - Default value: 1e-7

4.1.1.3. Data section

In this obligatory subsection, the user defines the haplotypic or genotypic content of the different samples to be analyzed. Each sample definition begins with the keyword SampleName and ends after the SampleData have been defined.

The user must specify the following parameters:

- A name for each sample
  - Notation: SampleName =
  - Possible values: Any string of characters within quotation marks
  - Example: SampleName= "A first example of a sample name"
- Thesample size
  - Notation: SampleSize =
  - Possible values: Any integer value
  - Example: SampleSize=732.

    Note: For haplotypic data, the sample size is equal to the haploid sample size. For genotypic data, the sample size should be equal to the number of diploid individuals present in the sample.

- The data
  - Notation: SampleData =
  - Possible values: A list of haplotypes or genotypes and their frequencies contained in the sample, which is entered within braces.
  - Example:

```
SampleData={
                MAN0102   1   A33  Cw10  B70    #pseudo-haplotypes
                              A33  Cw10  B7801  #the second pseudo-haplotype
}
```

If the gametic phase is known, the pseudo-haplotypes are treated as truly defined haplotypes. If the gametic phase is unknown, then only the allelic content of each locus is known.

4.1.1.4. Examples of input files

(1) Example of standard data (genotypic data, unknown gametic phase, recessive alleles)

In this example, the individual genotypes for 5 HLA loci are entered on two separate lines. In this example, the gametic phase between loci is unknown, and the data contains a recessive allele, which has been defined specifically as "xxx". Notably, with recessive data, all of the single locus homozygotes are considered potential heterozygotes with a null allele.

```
[Profile]
Title="Genotypic Data, Phase Unknown, 5 HLA loci"
NbSamples=1
GenotypicData=1
DataType=STANDARD
LocusSeparator=WHITESPACE
MissingData='?'
GameticPhase=0
RecessiveData=1
RecessiveAllele="xxx"
[Data]
[[Samples]]
SampleName="Population 1"
SampleSize=63
SampleData={
MAN0102  12 A33  Cw10  B70  DR1304  DQ0301
            A33  Cw10  B7801  DR1304  DQ0302
MAN0103  22 A33  Cw10  B70  DR1301  DQ0301
            A33  Cw10  B7801  DR1302  DQ0501
MAN0108  23 A23  Cw6  B35  DR1102  DQ0301
```

```
            A29  Cw7  B57  DR1104  DQ0602
MAN0109  6  A30  Cw4  B35  DR0801  xxx
            A68  Cw4  B35  DR0801  xxx
    }
```

(2) Example of standard data (genotypic data, known gametic phase)

In this example, three samples that consist of standard multi-loci data with known gametic phase have been defined. Therefore, the alleles listed on the same line constitute a haplotype on a given chromosome. For example, the genotype G1 consists of the following two haplotypes: A23-Cw6 on one chromosome and A29-Cw7 on the second.

```
[Profile]
Title="An example of genotypic data with known gametic phase"
NbSamples=3
GenotypicData=1
GameticPhase=1
RecessiveData=0
DataType=STANDARD
LocusSeparator=WHITESPACE
[Data]
[[Samples]]
SampleName="standard_pop1"
SampleSize=20
SampleData= {
G1  4  A23  Cw6
       A29  Cw7
G2  5  A30  Cw4
       A68  Cw4
}
```

4.1.2. The calculation of Hardy-Weinberg equilibrium and genetic parameters

4.1.2.1. The calculation of Hardy-Weinberg equilibrium

Performing an exact test of Hardy-Weinberg equilibrium (HWE) tests the hypothesis that the observed diploid genotypes are the product of a random union of gametes. This test is only possible for genotypic data, and separate tests are carried out at each locus. This test is analogous to Fisher's exact test on a two-by-two contingency table but extended to a contingency table of arbitrary size. If the gametic phase is unknown, then the test is only possible locus by locus. For data with a known gametic phase, the association at the haplotypic level within individuals can be tested.

The settings for the Hardy-Weinberg equilibrium test are displayed in Figure 2, and the output results are provided in Figure 3:

Locus by locus: Perform a separate HWE test for each locus.

Whole haplotype: Perform an HWE test at the haplotype level (if the gametic phase is available).

Locus by locus and whole haplotype: Perform both types of tests (if the gametic phase is available)

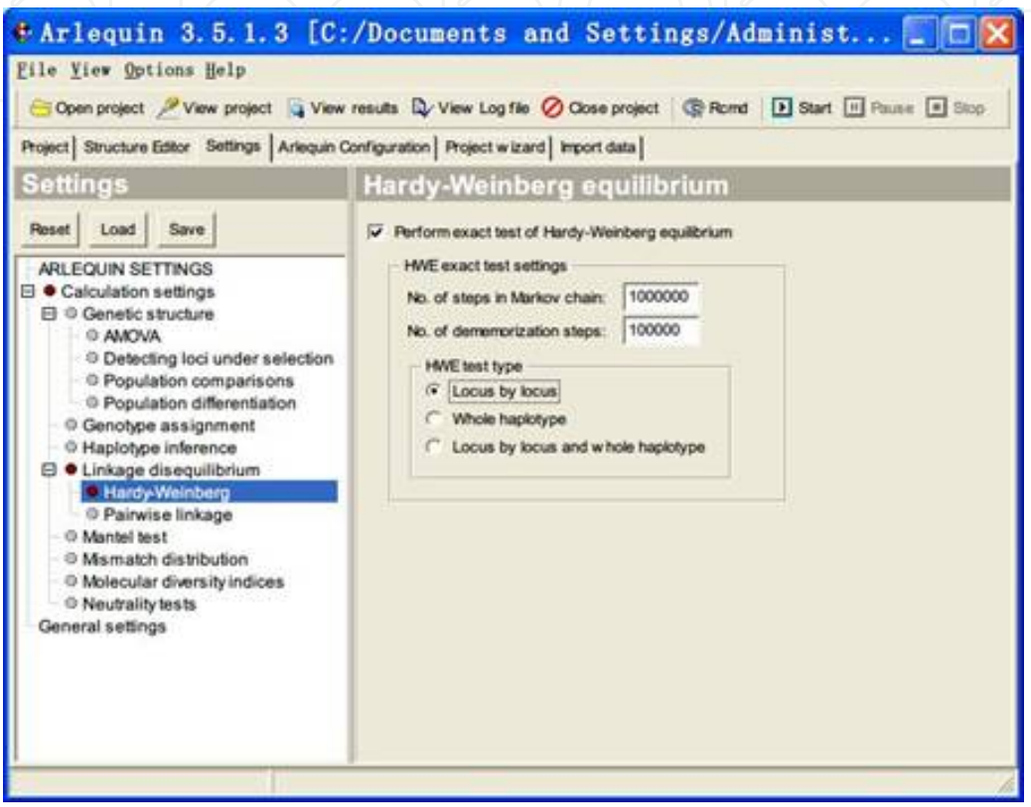

Figure 2. Settings for the Hardy-Weinberg equilibrium test

4.1.2.2. The calculation of genetic parameters

(1) Allele frequency, genotype frequency, and haplotype frequency

When genotypic data with an unknown gametic phase is being processed, two methods can be employed to infer haplotypes: the Expectation-Maximization (EM) algorithm (maximum-likelihood (ML)), which is the most commonly used, or the ELB algorithm (Bayesian).

When the gametic phase is not known or when recessive alleles are present, the ML haplotype frequencies are estimated from the observed data using an EM algorithm for multi-locus genotypic data. The settings are provided in Figure 4, and the results are shown in Figure 5.1, 5.2, and 5.3.

EM algorithms can be performed at the following levels:

Haplotype level: Estimate haplotype frequencies for haplotypes defined by alleles at all loci.

```
================================================================================
== Sample :     A test HLA sample
================================================================================


================================
== Hardy-Weinberg equilibrium : (A test HLA sample)
================================

          Guo, S. and Thompson, E. 1992.
          Levene H. (1949).

Exact test using a Markov chain (for all Loci):
Forecasted chain length     :1000000
Dememorization steps        :100000

--------------------------------------------------------------------------------
Locus  #Genot      Obs. Het.      Exp. Het.      P-value      s. d.      Steps done
--------------------------------------------------------------------------------
    1     19     0. 78947       0. 76671       0. 13879     0. 00037      1001000
    2     19     0. 26316       0. 80370       0. 00000     0. 00000      1001000
    3     19     0. 84211       0. 86486       0. 00322     0. 00005      1001000
    4     19     0. 73684       0. 90185       0. 00004     0. 00001      1001000
    5     19     0. 89474       0. 70697       0. 18591     0. 00048      1001000
--------------------------------------------------------------------------------
```

Figure 3. Results of the Hardy-Weinberg equilibrium test

Locus level: Estimate allele frequencies for each locus.

Haplotype and locus levels: The previous two options are performed in succession.

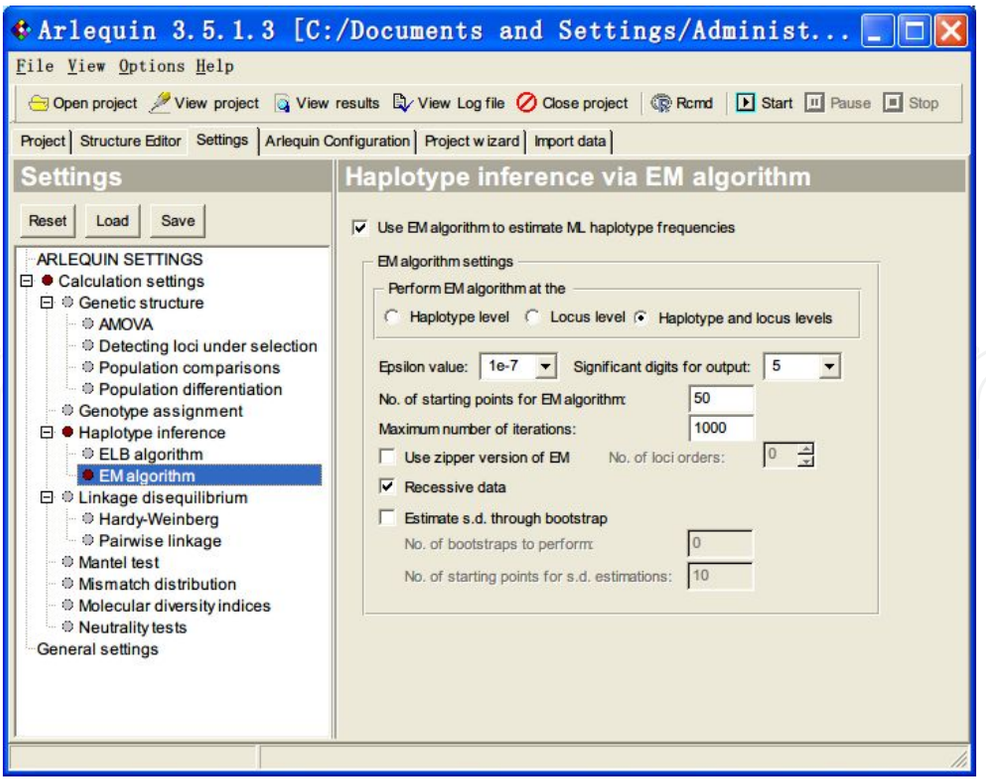

Figure 4. Settings for EM algorithmwith unknown gametic phase

```
==================================
== Haplotypes frequency estimation : (A test HLA sample)
==================================
----------------------------------------
(1) Conventional EM algorithm  (one pass)
----------------------------------------
No. of gene copies in sample                : 38
No. of random initial conditions for EM     : 50
No. of different maximum likelihoods found  : 12
Epsilon value for stopping iterations       : 1.000000e-07
LogLHood  : -99.66905848

Reference: Excoffier, L. and M. Slatkin. 1995.

Standard deviations not computed

------------------------------------------
Maximum-likelihood haplotype frequencies :
------------------------------------------

Total number of possible haplotypes      :           410
Minimum frequency to reach for output    :       1.00e-05

---------------------------------------
    #   Haplotype      Freq.       s.d.
---------------------------------------

    1      UNKNOWN  0.052632    0.000000 A2 NULL B7 DR1601 DQ0502
    2      UNKNOWN  0.026316    0.000000 A23 Cw4 B5102 DR0802 DQ0301
    3      UNKNOWN  0.026316    0.000000 A23 Cw6 B35 DR1104 DQ0602
    4      UNKNOWN  0.026316    0.000000 A23 Cw7 B49 DR1304 DQ0301
    5      UNKNOWN  0.026316    0.000000 A23 Cw7 B7 DR1101 DQ0301

-----------------------------------------
(2) Maximum-likelihood frequencies of genotypes
-----------------------------------------

(Haplotypes that are not listed above have a negative id)


===============
Phenotype #    1
===============
     Name          Abs. Freq.      Rel. Freq.
  MAN0102              1           0.004155

List of genotypes (   30)
-------------------------
Gen. #     Rel. Freq.       Exp. Freq.       Hapl. IDs    Haplotypes
    1       0.990730         0.004117              25    A33 Cw10 B7801 DR1304 DQ0301
                                                   21    A33 Cw10 B70 DR1304 DQ0302
    2       0.001324         0.000006              25    A33 Cw10 B7801 DR1304 DQ0301
                                                   47    NULL NULL B70 NULL DQ0302

(3) Allele frequencies :

(0 bootstrap replicates)
Allele frequencies for the locus 1
----------------------------------

No. of gene copies in sample                : 38
No. of random initial conditions for EM     : 50
No bootstrap confidence interval estimation
No. of different maximum likelihoods found  : 1
Epsilon value for stopping iterations       : 1.00000e-07
Logarithm of the sample maximum-likelihood  : -50.3215

------------------------------------------
Maximum-likelihood haplotype frequencies :
------------------------------------------

Total number of possible haplotypes      :             8
Minimum frequency to reach for output    :       1.00e-05

---------------------------------------
    #   Haplotype      Freq.       s.d.
---------------------------------------

    1      UNKNOWN  0.052632    0.000000 A2
    2      UNKNOWN  0.105263    0.000000 A23
    3      UNKNOWN  0.052632    0.000000 A29
    4      UNKNOWN  0.263158    0.000000 A30
```

Figure 5. Results of allele frequency, genotype frequencyand haplotype frequency

The settings when process in haplotypic data or genotypic (diploid) data with a known gametic phase are displayed in Figure 6, and the contents of the output results are provided in Figures 5.1, 5.2, and 5.3. The following parameters can be used in the process.

Use original definition: Haplotypes are identified according to their original identifier without considering that haplotype molecular definitions could be identical.

Infer from distance matrix: Similar haplotypes will be identified by computing a molecular distance matrix between haplotypes.

Haplotype frequency estimation:

Estimate haplotype frequencies by mere counting: Estimate the ML haplotype frequencies from the observed data using a mere gene counting procedure.

Estimate allele frequencies at all loci: Estimate allele frequencies at all loci separately.

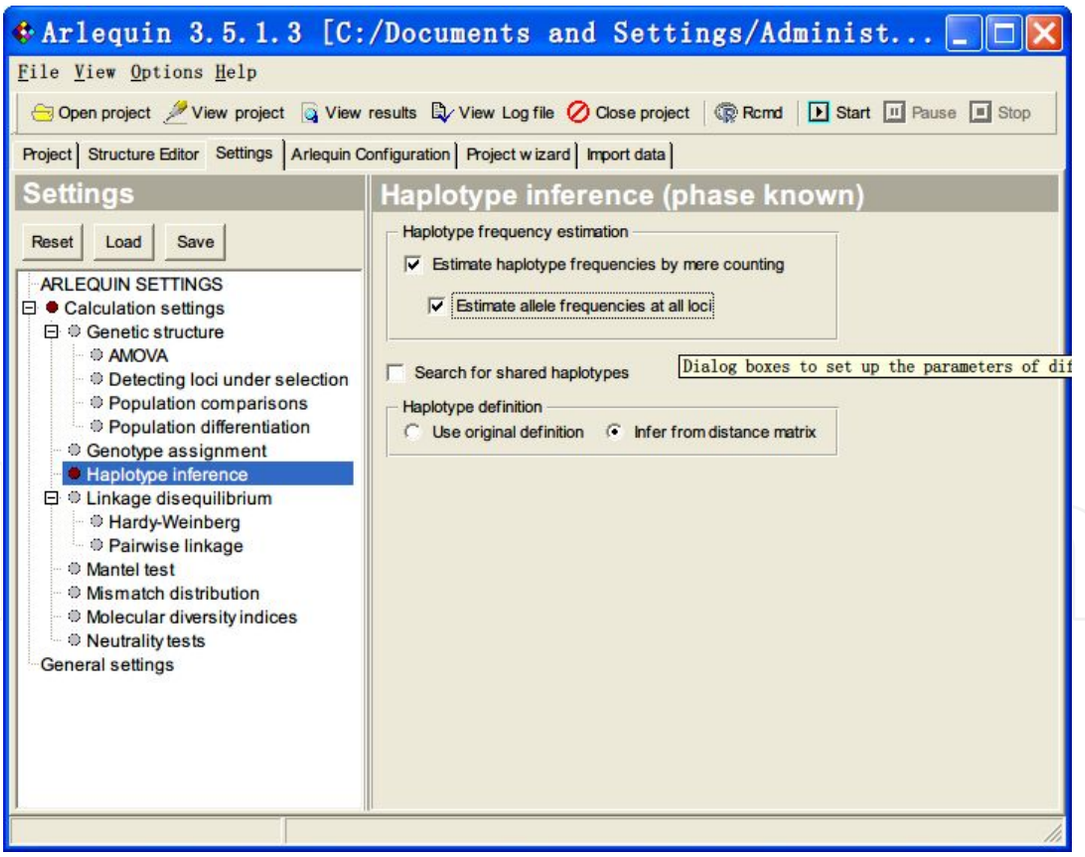

Figure 6. Settings for Haplotype inferencewith a known gametic phase

(2) The estimation of linkage disequilibrium parameters

The settings when processing data where the gametic phase is known are provided in Figure 7, and results of the calculation are shown in Figures 8.1 and 8.2.

Linkage disequilibrium between all pairs of loci: The user can test for the presence of a significant association between pairs of loci based on an exact test of linkage disequilibrium. The number of loci can be arbitrary, but if there are less than two polymorphic loci, this test is not applicable. The test procedure is analogous to Fisher's exact test on a two-by-two contingency table but extended to a contingency table of arbitrary size.

LD coefficients between pairs of alleles at different loci: Using this parameter, the D, D', and r^2 coefficients between all pairs can be calculated. D' is the most commonly used coefficientand represents the above section mentioned relative Δ value.

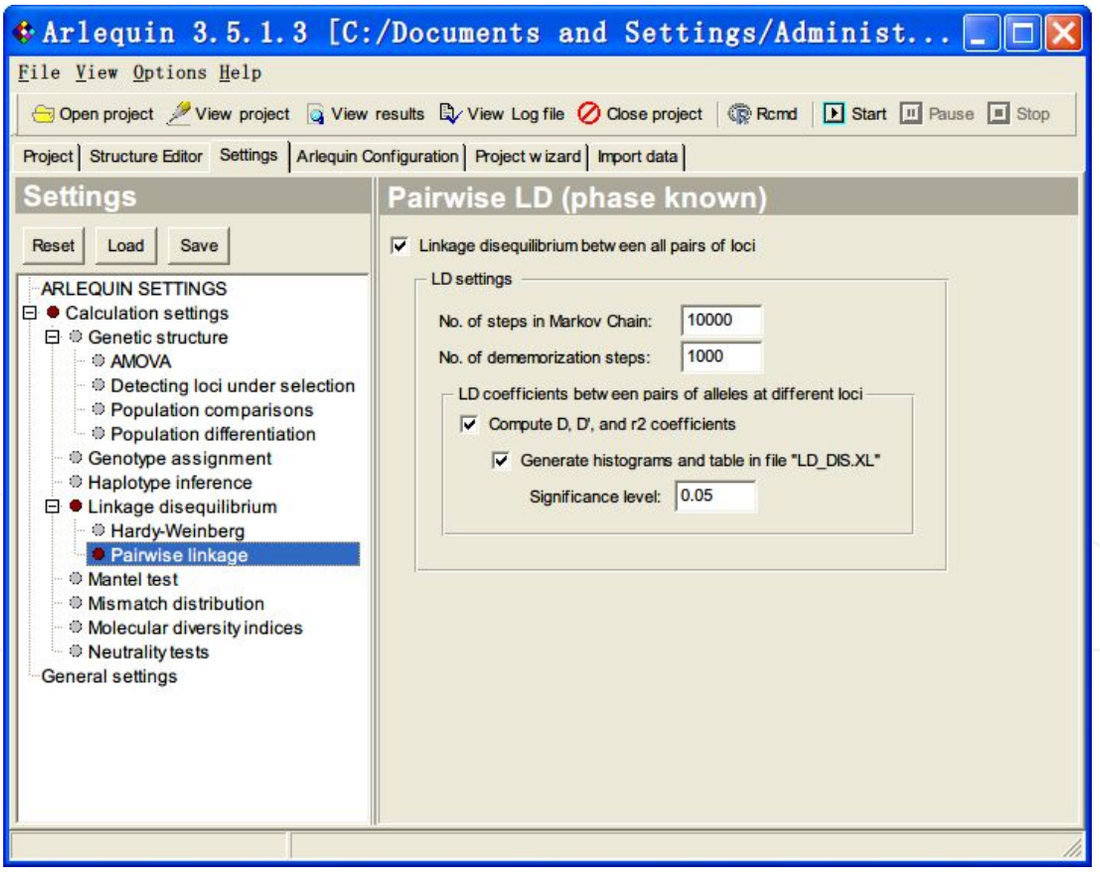

Figure 7. Settings for linkage disequilibrium

(1)

```
==================================
== Pairwise linkage disequilibrium : (A test HLA sample)
==================================

          Slatkin, M. 1994a.
          Slatkin, M. and Excoffier, L. 1996.
          Lewontin, R. C., and K. Kojima 1960.
Test of linkage disequilibrium for all pairs of loci:
-----------------------------------------------------

3: Table of disequilibrium values (D=pab-pa*pb) for all two-locus haplotypes

Locus0\Locus1|  Cw10|   Cw6|   Cw7|   Cw4|  NULL|   Cw2|   Cw5|
          A33|  0.12|  0.02| -0.03| -0.09| -0.01| -0.01| -0.01|
          A23| -0.03|  0.02|  0.01| -0.00| -0.02|  0.02| -0.00|
          A29| -0.02| -0.00|  0.02| -0.02| -0.01| -0.00|  0.02|
          A30| -0.02| -0.02|  0.02|  0.06| -0.02| -0.01| -0.01|
          A68| -0.03| -0.01| -0.01|  0.05|  0.01| -0.00| -0.00|
           A2| -0.02| -0.00| -0.01| -0.02|  0.04| -0.00| -0.00|
          A32| -0.01| -0.00| -0.00|  0.02| -0.00| -0.00| -0.00|

4: Table of standardized disequilibrium values (D'=D/Dmax)

Locus0\Locus1|  Cw10|   Cw6|   Cw7|   Cw4|  NULL|   Cw2|   Cw5|
          A33|  0.70|  0.45| -0.49| -0.77| -0.16| -1.00| -1.00|
          A23| -1.00|  0.25|  0.14| -0.14| -1.00|  1.00| -1.00|
          A29| -1.00| -1.00|  0.42| -1.00| -1.00| -1.00|  1.00|
          A30| -0.31| -1.00|  0.19|  0.30| -0.37| -1.00| -1.00|
          A68| -1.00| -1.00| -1.00|  0.65|  0.11| -1.00| -1.00|
           A2| -1.00| -1.00| -1.00| -1.00|  1.00| -1.00| -1.00|
          A32| -1.00| -1.00| -1.00|  1.00| -1.00| -1.00| -1.00|
```

(2)

```
5: Table of Chi-square values ( Chi2=sqr(D)*n/(pa*(1-pa)*pb*(1-pb)) )

Locus0\Locus1|  Cw10|   Cw6|   Cw7|   Cw4|  NULL|   Cw2|   Cw5|
          A33| 11.62|  1.01|  0.91|  5.98|  0.11|  0.67|  0.67|
          A23|  1.82|  1.80|  0.55|  0.03|  0.84|  8.73|  0.12|
          A29|  0.86|  0.18|  2.51|  0.86|  0.40|  0.06| 18.49|
          A30|  0.53|  1.16|  0.56|  2.92|  0.34|  0.37|  0.37|
          A68|  1.82|  0.38|  0.68|  4.61|  0.29|  0.12|  0.12|
           A2|  0.86|  0.18|  0.32|  0.86| 11.26|  0.06|  0.06|
          A32|  0.42|  0.09|  0.16|  2.52|  0.19|  0.03|  0.03|

6: Table of Chi-square P values (1 d.f.)

Locus0\Locus1|  Cw10|   Cw6|   Cw7|   Cw4|  NULL|   Cw2|   Cw5|
          A33|  0.00|  0.32|  0.34|  0.01|  0.74|  0.41|  0.41|
          A23|  0.18|  0.18|  0.46|  0.85|  0.36|  0.00|  0.73|
          A29|  0.35|  0.67|  0.11|  0.35|  0.53|  0.81|  0.00|
          A30|  0.47|  0.28|  0.46|  0.09|  0.56|  0.54|  0.54|
          A68|  0.18|  0.54|  0.41|  0.03|  0.59|  0.73|  0.73|
           A2|  0.35|  0.67|  0.57|  0.35|  0.00|  0.81|  0.81|
          A32|  0.52|  0.77|  0.69|  0.11|  0.66|  0.87|  0.87|
```

Figure 8. Results oflinkage disequilibrium

When the gametic phase is unknown, a different procedure for testing the significance of the association between pairs of loci is used. The procedure is based on a likelihood ratio test, where the likelihood of the sample evaluated under the hypothesis of no association between loci (linkage equilibrium) is compared with the likelihood of the sample when association is allowed. The significance of the observed likelihood ratio is found by computing the null distribution of this ratio under the hypothesis of linkage equilibrium, using a permutation procedure. The settings for this procedure are shown in Figure 9, and the output results are provided in Figure 10.

http://dx.doi.org/10.5772/57493

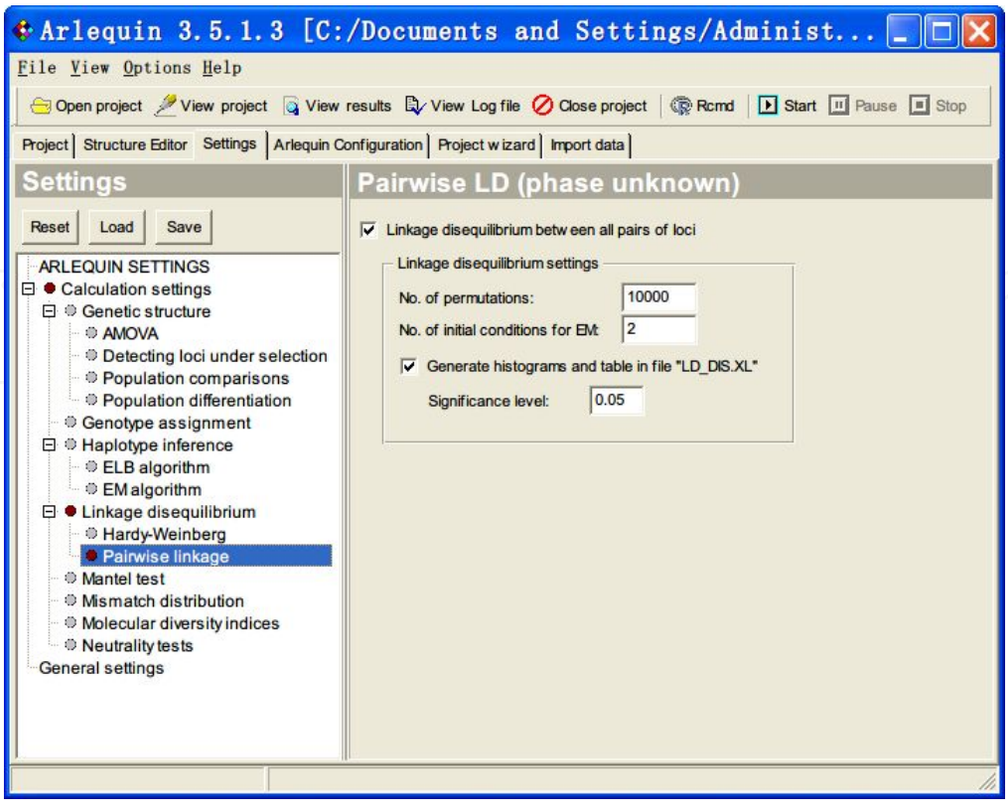

Figure 9. Settings for linkage disequilibrium with unkown phase

```
====================================
== Pairwise linkage disequilibrium : (A test HLA sample)
====================================

          Slatkin, M. 1994a.
          Slatkin, M. and Excoffier, L. 1996.
          Lewontin, R. C., and K. Kojima 1960.


Test of linkage disequilibrium for all pairs of loci:
-----------------------------------------------------


Permutation test using the EM algorithm
Number of permutations              : 10000
Number of initial conditions for EM  : 2

Pair(0, 1)
          LnLHood LD :  -85.13107          LnLHood LE : -110.23678
          Exact P=  0.00129  +-   0.00033 (10100 permutations done)     Chi-square test value=50.21141 (P =  0.05811,  36 d.f.)
Pair(0, 2)
          LnLHood LD :  -86.73838          LnLHood LE : -119.06953
          Exact P=  0.00059  +-   0.00024 (10100 permutations done)     Chi-square test value=64.66230 (P =  0.52355,  66 d.f.)
Pair(1, 2)
          LnLHood LD :  -89.88222          LnLHood LE : -128.66327
          Exact P=  0.00000  +-   0.00000 (10100 permutations done)     Chi-square test value=77.56209 (P =  0.15620,  66 d.f.)
Pair(0, 3)
          LnLHood LD :  -90.92872          LnLHood LE : -127.08772
          Exact P=  0.00010  +-   0.00010 (10100 permutations done)     Chi-square test value=72.31799 (P =  0.27729,  66 d.f.)
```

Figure 10. Results oflinkage disequilibrium with unkown phase

(3) The calculation of genetic distance

Arlequin provides several calculation methods to determine genetic distance, including *Reynolds'* distance, *Slatkin's* linearized coefficient, *Nei's* genetic distance, etc. *Nei's* genetic distance and the *Cavalli-Sforza* genetic distance calculating methods are the most commonly

used and produce the most similar results. The settings for calculating genetic distance are shown in Figure 11, and the output results are provided in Figure 12.

The calculation parameters are as follows:

Compute pairwise differences: Computes *Nei's* average number of pairwise differences within and between populations.

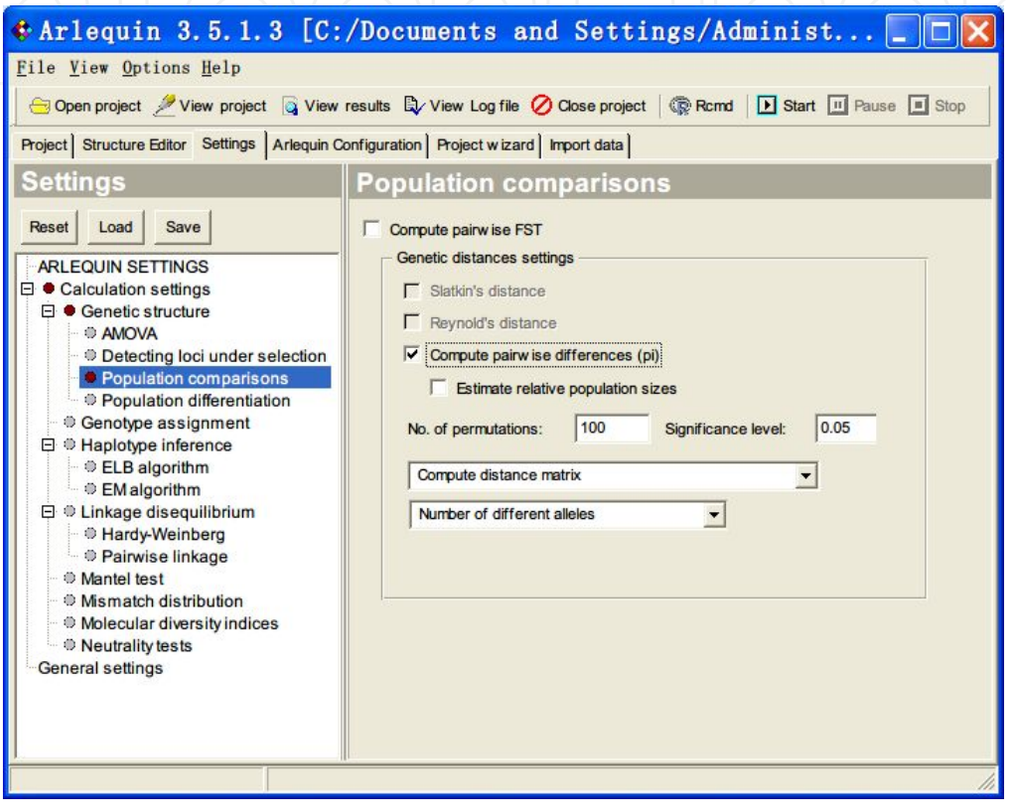

Figure 11. Settings for genetic distance

```
---------------------------------------
Population average pairwise differences
---------------------------------------

Above diagonal    : Average number of pairwise differences between populations (PiXY)
Diagonal elements : Average number of pairwise differences within population (PiX)
Below diagonal    : Corrected average pairwise difference (PiXY-(PiX+PiY)/2)

Distance method: Pairwise differences
                1          2          3          4          5          6          7
1    1.77562    1.82713    1.82553    1.80799    1.78133    1.82920    1.86023
2    0.05185    1.77493    1.83481    1.80863    1.76641    1.84444    1.83427
3    0.01880    0.02842    1.83784    1.83406    1.80991    1.85439    1.87138
4    0.02756    0.02855    0.02252    1.78524    1.78046    1.80833    1.85629
5    0.10834    0.09377    0.10581    0.10266    1.57035    1.83631    1.76463
6    0.04114    0.05673    0.03523    0.01547    0.15089    1.80049    1.87949
7    0.13367    0.10806    0.11371    0.12493    0.14071    0.14051    1.67749
```

Figure 12. Results of genetic distance

4.2. The requirements of data analysis on new-generation HLA typing techniques

HLA data analysis methods have always been closely related to the development of HLA genotyping techniques. In the 1980s and 1990s, HLA serotyping was the preferred technique. HLA phenotypes were determined first, and the square-root method was used commonly to predict HLA genotype frequencies. Currently, HLA genotyping techniques are more prevalent. Researchers tend to use the direct counting method to calculate the genotype. In previous HLA haplotype analyses, the haplotype was predicted using group analysis, and then the individual haplotype frequency was estimated. However, with the considerable cost decrease of genotyping techniques, more pedigree data are available for studies, such as the *Haplomap* program, where haplotypes can be studied directly using pedigree analysis. Moreover, with the development of new-generation gene sequencing techniques and the optimization of large-fragment high-throughput sequencing and fragment (reads) assembly algorithms, individual haplotypes would be distinguished directly. These methods contribute greatly to simplifying the data analyses process.

Currently, HLA data types are no longer limited to allele data. Other types of data, such as SNPs, microsatellite markers, short sequence repeats, etc., could be used for conjoint analysis with HLA data. This chapter has only introduced the application of *Arlequin* software in classic HLA data analysis. However, many other outstanding tools, such as *Phase* are more commonly used in studies of haplotype establishment using group genotype data and hot spot model recombination. *HapView* is more prominent in graphic linkage disequilibrium (LD) and haplotype studies, and the professional statistical software *SAS* is also used commonly for HLA data analysis. With the development of HLA typing techniques and analysis techniques, data processing methods will also become more in-depth and detailed.

Acknowledgements

Supported by grantsfrom the State Key Development Programfor Basic Research of China (No. 2003CB515509 and 2009CB522401) and from National Natural Scientific Foundation of China(No.81070450 and 30470751) to Dr. X.-Y.Z.

Author details

Fang Yuan and Yongzhi Xi*

*Address all correspondence to: xiyz@yahoo.com

Department of Immunology and National Center for Biomedicine Analysis, Beijing Hospital Affiliated to Academy of Medical Sciences, Beijing, PRC

References

[1] Edwards AW. Foundations of Mathematical Genetics, 2nd edition. Cambridge University Press. Cambridge. 2000.

[2] Crow JF. Hardy-Weinberg and Language Impediments. Genetics. 1999, 152: 821.

[3] Masel, Joanna. Rethinking Hardy–Weinberg and Genetic Drift in Undergraduate Biology. BioEssays. 2012, 34: 701.

[4] Cox DR. Principles of Statistical Inference. Cambridge University Press. Cambridge. 2006.

[5] Hu LP. Medical Statistics. People's Military Medical Press. Beijing. 2010.

[6] Xu TH, Wang J. Design of Medical Experiments: Lecture 2, Rules of Randomization and Blinding Method. Chinese Medical Journal. 2005, 40: 54.

[7] Marsh SG, Albert ED, Bodmer WF, et al. Nomenclature for Factors of the HLA System. 2010.

[8] Robinson J, Waller MJ, Fail SC, et al. The IMGT/HLA database. Nucleic Acids Research. 2009,37: 1013.

[9] Sharon L. Sampling: Design and Analysis, 2nd edition. Cengage Learning. 2009.

[10] Tan JM, Tissue Typing Technique and Clinical Application, 1st edition. People's Medical Publishing House. Beijing, 2002.

[11] Wang XZ. Principles of Population Genetics. Sichuan University Press. Chengdu. 1994.

[12] Guo J, Hu LP. Medical Genetics Statistics and SAS Application. People's Medical Publishing House. Beijing. 2012.

[13] Weir BS. Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sinauer Associates Inc. USA. 1996.

[14] Excoffier L, Slatkin M. Maximum-Likelihood Estimation Of Molecular Haplotype Frequencies In A Diploid Population. Molecular Biology and Evolution. 1995, 12:921.

[15] Excoffier L, Slatkin M. Incorporating genotypes of relatives into a test of linkage disequilibrium. The American Journal of Human Genetics. 1998, 171-180.

[16] Guo S, Thompson E. Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles. Biometrics. 1992, 48:361.

[17] Raymond M, Rousset F. An Exact Test for Population Differentiation. Evolution. 1995,49:1280.

[18] Gaggiotti O, Excoffier L. A Simple Method of Removing the Effect of a Bottleneck and Unequal Population Sizes on Pairwise Genetic Distances. Proceedings of the Royal Society London. 2000, 267: 81.

[19] Dempster A, Laird N, Rubin D. Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. 1977, 39:1.

[20] Cavalli-Sforza LL, Population structure and human evolution. Proceedings of the Royal Society London, 1966, 164, 362.

[21] Cavalli-Sforza LL, Bodmer WF. The Genetics of Human Populations. W.H. Freeman Publishers. San Francisco. 1971.

[22] Lange K, Mathematical and Statistical Methods for Genetic Analysis. Springer. New York. 1997

[23] Excoffier L, Lischer H. Arlequin Suite ver 3.5: A New Series of Programs to Perform Population Genetics Analyses under Linux and Windows. Molecular Ecology Resources. 2010, 10: 564.

[24] Stephens M, Scheet P. Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation. American Journal of Human Genetics. 2005, 76:449-462.

[25] Scheet P, Stephens M. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. American Journal of Human Genetics. 2006, 78: 629.

IntechOpen

IntechOpen

IntechOpen