Genetics of large populations and association studies

Lounès Chikhi, Pauline Garnier-géré

Genetics of Large Populations and Association Studies

Lounès Chikhi, *Université Paul Sabatier, Toulouse, France*

Pauline Hélène Garnier-Géré, *INRA Recherches Forestières, Gazinet, France*

Advanced article

Association studies aim to discover the genetic basis of genetic diseases or drug responses through the comparison of genotypes at marker loci with phenotypes of affected and nonaffected individuals in large populations assumed to be at Hardy–Weinberg equilibrium.

Article contents

doi: 10.1038/npg.els.0005446

Introduction

The distribution of genetic variability in natural populations is the result of a number of evolutionary factors. We present concepts on the genetics of large populations, with special reference to association studies. Particular issues related to selection, mutation and genetic load, or migration in spatial models are not treated here. Genetic drift is assumed to be negligible, unless otherwise stated. (***See*** Fitness and Selection; Genetic Drift; Genetic Load; Population Genetics: Historical Aspects; Mutation Rates: Evolution.)

Hardy–Weinberg (HW) Equilibrium

Population genetics is founded on a simple principle that was independently demonstrated in 1908 by the English mathematician G. H. Hardy and the German physician W. Weinberg. This principle states that, in a large, randomly mating population with nonoverlapping generations, the 'genotype' frequencies at one diploid locus (1) are the product of the 'allele' frequencies at that locus, and (2) remain constant between generations, hence the 'equilibrium', irrespective of the dominance of the alleles, and provided that migration, mutation and natural selection do not affect that locus. Another implicit assumption of the Hardy–Weinberg (HW) principle is that allele frequencies are the same in males and females (see **Table 1**).

Table 1 shows how HW equilibrium proportions can be obtained by considering a locus with two alleles A_1 and A_2 in respective frequencies p_1 and p_2. There are only three possible genotypes, namely two homozygotes, A_1A_1 and A_2A_2, and one heterozygote, A_1A_2. **Table 1** shows that, under random mating, the proportions of these genotypes will be ${p_1}^2$, ${p_2}^2$ and $2p_1p_2$ respectively.

Note that HW proportions are obtained by expanding $(p_1+p_2)^2$. For a locus with n alleles, A_1, $A_2, \ldots, A_n$, the frequencies of homozygotes A_iA_i and heterozygotes A_iA_j are obtained by expanding $(p_1+p_2+\cdots+p_n)^2$ and are thus ${p_i}^2$ and $2p_ip_j$ respectively. Note also that the HW equilibrium is reached after only one generation of random mating.

Table 1 Hardy–Weinberg equilibrium

		Male gametes	
		A_1 p_1	A_2 p_2
Female gametes	A_1 p_1	A_1A_1 p_1^2	A_1A_2 p_1p_2
	A_2 p_2	A_1A_2 p_1p_2	A_2A_2 p_2^2

The expected genotypic proportions are given for a biallelic locus with alleles A_1 and A_2, with respective allelic frequencies p_1 and p_2.

Departures from HW proportions

Deviations from HW equilibrium can be caused by any departure from the assumptions mentioned above, the most important being the departure from random mating. For instance, if there is positive assortative mating (mating of like with like) for some phenotypic character (e.g. color of skin), more homozygotes than expected under HW will be observed, for loci associated with the character. This generates a heterozygote deficit (or homozygote excess). The opposite effect will be observed for 'negative' assortative mating. Inbreeding from mating between related individuals is also possible in large populations and can cause departures from HW equilibrium. (***See*** Inbreeding; Nonrandom Mating.)

Nonrandom mating can be formalized by defining an index, F (Wright, 1951), which quantifies departures from HW proportions: $F=(H_{\mathrm{obs}}-H_{\mathrm{exp}})/H_{\mathrm{exp}}$, where H_{obs} is the observed proportion of heterozygotes and H_{exp} the proportions expected under

HW equilibrium. We can rewrite this as $H_{\text{obs}} = H_{\text{exp}}(1 - F)$. For a biallelic locus, the HW principle can be generalized to give the proportions of genotypes A_1A_1, A_2A_2 and A_1A_2: $p_1^2 + Fp_1p_2$, $p_2^2 + Fp_1p_2$ and $2p_1p_2(1 - F)$. When $F = 0$ we find HW proportions. For loci with more than two alleles, one F can be defined for each pair of alleles or by considering one allele against all others. (***See*** Kinship and Inbreeding.)

In theory, allele frequency differences between males and females can generate departures from HW proportions. However, after one generation of random mating, any difference in allele frequencies disappears (Hedrick, 2000).

Sex-linked genes

For genes located on the human X chromosome, genotypic frequencies can be different between sexes, because females carry two X chromosomes while males carry only one. Thus, genotypic frequencies can be in HW proportions for females only (since males are haploid for the X chromosome), but they will depend on allele frequencies in both males and females (Hedrick, 2000). It can be shown that, even if sexes have different initial allelic frequencies, a few generations of random mating will bring the frequencies in both sexes to the same value (Hedrick, 2000). (***See*** Chromosome Analysis and Identification; Sex Chromosomes.)

Wahlund effect

The Wahlund effect is the deficit of heterozygotes observed in a 'sample' due to the existence of unknown population subdivision. It is caused by the variance in allele frequency among subpopulations forming the sample.

A graphical representation of the HW equilibrium and of the Wahlund effect can be obtained by noting that, for a biallelic locus, the frequency of heterozygotes is a simple function of p_1: $2p_1p_2 = 2p_1(1 - p_1) = 2p_1 - 2p_1^2$ (**Figure 1**). The solid curve in **Figure 1** represents the expected proportion of heterozygotes in a randomly mating population for any value of p_1. Any point below or above the curve represents a departure from HW proportions and corresponds to either a deficit or an excess of heterozygotes respectively. For a locus with n alleles, different curves will be obtained for different heterozygotes A_iA_j, as the sum of allele frequencies will change depending on i and j.

Consider now the case where a population P is in fact subdivided in two subpopulations, P_1 and P_2 both at HW equilibrium for different allele frequencies. **Figure 1** shows that any sample taken from a mixture of P_1 and P_2, and thus any sample from P, will have a genetic composition graphically represented by the straight line between P_1 and P_2. The observed frequency of heterozygotes in the stratified sample from P will always be smaller than that expected under HW equilibrium, given its allelic frequencies.

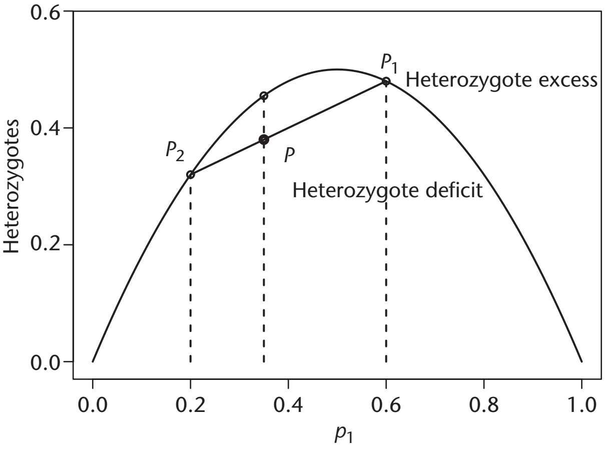

Figure 1 Hardy–Weinberg (HW) equilibrium and Wahlund effect. The solid curve represents the proportion of heterozygotes in a population at HW equilibrium for a biallelic locus, where p_1 is the allelic frequency of one of the two alleles. The population P is subdivided in two subpopulations or strata, P_1 and P_2, which are both at HW equilibrium for different allelic frequencies ($p_1 = 0.2$ and 0.6, respectively). P_1 and P_2 are both on the solid curve, whereas P is necessarily below, indicating that the Wahlund effect leads to a heterozygote deficit.

The Wahlund effect or the cryptic stratification of populations is particularly important because it generates heterozygote deficits that may be mistaken for inbreeding or assortative mating. Remember that one generation of random mating between individuals belonging to different strata is enough to establish the HW equilibrium at each locus. This means that admixture events followed by random mating cannot be detected with single-locus frequency data through the Wahlund effect and require multilocus data. Indeed, population structure, whether due to present-day structure (stratification, see below) or to ancient events (admixture), can create allelic associations between different loci, with implications for association studies. (***See*** Population Differentiation: Measures.)

Association Studies

Linkage disequilibrium and allelic associations

Consider two loci A and B, with alleles $A_1, \ldots, A_n$ and $B_1, \ldots, B_m$ respectively. If the two loci were independent, the frequency of particular gametes, say $p_{A_7B_2}$ for the pair A_7B_2 (called a haplotype), would simply be the product of their respective frequencies, $p_{A_7}p_{B_3}$. In other words, the genotype at one locus would provide no information about the genotype at the other. When two alleles are not statistically independent they are said to be positively (or negatively) associated if $p_{A_iB_j}$ is larger (or smaller) than $p_{A_i}p_{B_j}$. A natural measure of this association is therefore the difference $D = p_{A_iB_j} - p_{A_i}p_{B_j}$, the linkage disequilibrium (LD, also called gametic disequilibrium). When $D = 0$, the two loci are said to be at linkage equilibrium.

LD can be generated by a number of factors: physical linkage, selection on some haplotypes, founder effects, and admixture of populations which may themselves be at linkage equilibrium (Hedrick, 2000).

In a large, randomly mating population, LD between alleles at neutral loci is assumed to be mainly the result of physical linkage. If we call r the recombination rate between the two loci, the frequency $p_{A_iB_j}(t)$ of a particular haplotype in generation t is a function of r and $p_{A_iB_j}(t-1)$:

$$p_{A_iB_j}(t) = (1-r)p_{A_iB_j}(t-1) + rp_{A_i}p_{B_j} \tag{1}$$

where $(1-r)p_{A_iB_j}(t-1)$ represents the proportion of haplotypes from the previous generation without recombination, while $rp_{A_iB_j}p_{B_j}$ represents those generated by recombination (neglecting multiple recombination events). We can rewrite eqn (1) as

$$p_{A_iB_j}(t) - p_{A_i}p_{B_j} = (1-r)(p_{A_iB_j}(t-1) - p_{A_i}p_{B_j}) \tag{2}$$

or

$$D_t = (1-r)D_{t-1} = (1-r)^t D_0 \tag{3}$$

where D_0 is the LD at generation zero.

Note that, in a randomly mating population, D_0 might be due to a recent admixture or a founder event that would not have been detected by the analysis of HW proportions at individual loci. Equation (3) shows that LD will decay with time as a function of the recombination rate. If loci are not physically linked, then $r = 0.5$ and LD is halved every generation. Thus, LD caused by admixture or founder effects will decay much more rapidly for unlinked than for tightly linked loci ($r \sim 0$). The amount of LD still visible today in the human genome therefore depends on physical linkage and on the demographic history of populations. (***See*** Linkage Disequilibrium; Population History and Linkage Disequilibrium.)

Association studies and the problem of population stratification

Association studies aim at identifying genes involved in hereditary diseases or drug response. The basic principle is to compare allelic frequencies at marker loci between a set of affected individuals who are unrelated (case) and a set of unaffected individuals (control) in a random homogeneous population sample (Sham, 1998). A statistical association between a particular disease phenotype and genotype at a marker locus is usually considered as evidence of physical linkage between the marker locus (whose location is known) and a disease locus (whose location is unknown). Markers have to be polymorphic so that only one or some alleles may be associated with the allele(s) causing the disease. They also have to be in sufficiently large numbers so as to allow the fine mapping of a previously located candidate region for a disease gene. (***See*** Psychopharmacogenetics.)

Thus, association studies focus on ‘population samples’, while classic linkage analyses use ‘family pedigrees’ for comparing the segregation of markers and disease phenotypes. However, both types of study can be seen as the two ends of a continuum of methods for finding disease genes in human populations (Sham, 1998). Despite the success of linkage studies in locating genes for diseases with simple determinism, association studies have been advocated as a method of choice for mapping complex diseases and for the discovery of new genes from genome-wide scans using an increasing number of available markers such as single nucleotide polymorphisms (SNPs) (Risch and Merikangas, 1996; Reich *et al.*, 2001). One advantage is that association studies use LD in populations and thus take advantage of all the recombination events between a marker and a disease locus since the original disease mutation, instead of just events in the past few generations in a family. (***See*** Genetic Linkage Mapping; Linkage and Association Studies; Single Nucleotide Polymorphism (SNP); Single-base Mutation.)

The weakness of association studies is that their outcome can be affected by all evolutionary factors or historical events that can create LD. In particular, stratification has been seen as a major drawback, since it can generate LD between any markers that are not physically linked to a disease locus, but are at a high frequency in the subpopulation where the disease is most common. These ‘spurious associations’ therefore convey erroneous information on the location of disease loci. (***See*** Linkage and Association Studies: Replication.)

This has led to the development of methods using family data, such as the haplotype relative risk (HRR) method and the transmission disequilibrium test (TDT). These tests require the use of affected individuals and of their parents (a practical problem in some cases) to test the patterns of allele transmission (Sham, 1998). Importantly, they have been shown to be robust to variation in the population history including subdivision and recent admixture. (***See*** Relatives-based Tests of Association.)

However, these family-based tests can be difficult to implement, especially for late age-of-onset conditions. Moreover, in well-defined populations without stratification, simulations show that the TDT is less powerful than association studies using marker-based permutations tests (Long and Langley, 1999). The identification of population stratification without having to type new family members is therefore important. A simple test consists of looking for a Wahlund effect at an individual locus; more subtle approaches are possible (Sham, 1998).

Detection of LD with a number of markers located on different chromosomes (physically unlinked) is a good indicator of spurious associations for diseases caused by one gene with major effects (Sham, 1998). This idea has led to the development of methods to detect and/or correct for stratification in association studies using a set of loci that are known to be unlinked (Pritchard and Rosenberg, 1999; Reich and Goldstein, 2001). These approaches look promising, since they avoid the use of family data and the need to sample new individuals, and only require the typing of a few tens of unlinked markers in the same population samples.

Conclusions

The genetics of human populations is a complex matter governed by a number of parameters. We have seen that population stratification is one important demographic factor that can both create the Wahlund effect and inflate LD in association studies, etc. Empirical findings show that other practical problems can make association mapping less efficient. The study of Emahazion *et al.* (2001) shows that the use of a large number of markers within one population sample can greatly increase the number of significant but spurious associations due to chance alone (type-I errors). The interaction between many potential genes involved in the expression of the disease and between genetic and environmental effects can also generate errors or decrease efficiency of gene detection (Long *et al.*, 1997; Emahazion *et al.*, 2001). (***See*** Complex Multifactorial Genetic Diseases; Gene–Environment Interaction.)

With the development of large-scale association studies using SNPs, there has been a renewal of interest in understanding better the relative importance of the different factors that can affect the extent and the patterns of variation of LD in different human populations. Initial simulation and empirical studies have suggested that LD might only extend to a few kilobases (kb) and that, given the size of our genome, mapping would require more than 500 000 markers (Kruglyak, 1999). However, recent surveys on LD in the human genome estimate that 'LD blocks' might extend over far greater regions of the genome (30–300 kb; Abecasis *et al.*, 2001) with significant variation between populations (40–60 kb in North European Americans but much less in Nigerians; Reich *et al.*, 2001). These results stress the importance of population genetic studies to uncover the effects of past demographic history, population structure and growth rate for the estimation of recombination rates and LD patterns. A number of studies also show that LD varies across the genome. In particular, the role of crossover, and gene conversion and inversion in shaping LD may depend on the genetic distance between markers. (***See*** DNA Recombination; Genome Organization of Vertebrates; Genome Size.)

Future studies will help us improve our understanding of the extent and distribution of LD across populations (including population isolates; Service *et al.*, 2001) and genomic regions. The detailed analysis of SNPs' variability and informativeness is also likely to become of prime interest for the estimation of differential selection effects along the chromosome and for the fine mapping and positional cloning of susceptibility genes involved in complex diseases and drug responses. (***See*** Bioethics: Utilitarianism; Gene Mapping and Positional Cloning; Single Nucleotide Polymorphisms (SNPs): Identification and Scoring.)

See also

Genetic Load
Genetic Maps: Integration
Population Genetics: Historical Aspects

References

Abecasis GR, Noguchi E, Heinzmann A, *et al.* (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *American Journal of Human Genetics* **68**: 191–197.

Emahazion T, Feuk L, Jobs M, *et al.* (2001) SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. *Trends in Genetics* **17**(7): 407–413.

Hedrick PW (2000) *Genetics of Populations*, 2nd edn. Boston, MA: Jones and Bartlett Publishers.

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

Long AD, Grote MN and Langley CH (1997) Genetic analysis of complex diseases. *Science* **275**: 1328.

Long AD and Langley CH (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research* **9**: 720–731.

Pritchard JK and Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* **65**: 220–228.

Reich DE, Cargill M, Bolk S, *et al.* (2001) Linkage disequilibrium in the human genome. *Nature* **411**: 199–204.

Reich DE and Goldstein DB (2001) Detecting association in a case–control study while correcting for population stratification. *Genetic Epidemiology* **20**: 4–16.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* **273**: 1516–1517.

Service SK, Ophoff RA and Freimer NB (2001) The genome-wide distribution of background linkage disequilibrium in a population isolate. *Human Molecular Genetics* **10**: 545–551.

Sham P (1998) Statistics in human genetics. *Arnold Applications of Statistics*. New York, NY: Oxford University Press.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics* **15**: 323–354.

Further Reading

Crow JF and Kimura M (1970) *An Introduction to Population Genetics Theory*. New York, NY: Harper and Row.

Goldstein DB and Weale ME (2001) Population genomics: linkage disequilibrium holds the key. *Current Biology* **11**: 576–579.

Hartl DL and Clark AG (1989) *Principles of Population Genetics*, 2nd edn. Sunderland, MA: Sinauer Associates.

Liu BH (1998) *Statistical Genomics. Linkage, Mapping, and QTL Analyses.* Boca Raton, FL: CRC Press.

Nei M and Li W-H (1973) Linkage disequilibrium in subdivided populations. *Genetics* **75**: 213–219.

Pritchard JK and Przeworski M (2001) Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* **69**: 1–14.

Slatkin M (1994) Linkage disequilibrium in growing and stable populations. *Genetics* **137**: 331–336.

Spielman RS, McGinnis RE and Ewens WJ (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**: 506–516.

Stephens JC, Briscoe D and O'Brien SJ (1994). Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *American Journal of Human Genetics* **55**: 809–824.

Strachan T and Read AP (1996) *Human Molecular Genetics.* Oxford, UK: BIOS Scientific Publishers.

Weir BS (1996) *Genetic Data Analysis,* vol. II. Sunderland, MA: Sinauer Associates.