Population Genetic Data Analysis

Summer Institute in Statistical Genetics
Institute for Molecular Bioscience
University of Queensland

February 9-10, 2017

Bruce Weir: bsweir@uw.edu
University of Washington

Contents

PROBABILITY THEORY

Probability

Probability provides the language of data analysis.

Equiprobable outcomes definition:
Probability of event E is number of outcomes favorable to E divided by the total number of outcomes. e.g. Probability of a head $= 1/2$.

Long-run frequency definition:
If event E occurs n times in N identical experiments, the probability of E is the limit of n/N as N goes to infinity.

Subjective probability:
Probability is a measure of belief.

First Law of Probability

Law says that probability can take values only in the range zero to one and that an event which is certain has probability one.

$$\begin{cases} 0 \leq \Pr(E) \leq 1 \\ \\ \Pr(E|E) = 1 \text{ for any } E \end{cases}$$

i.e. If event E is true, then it has a probability of 1. For example:

$$\Pr(\text{Seed is Round}|\text{Seed is Round}) \;=\; 1$$

Second Law of Probability

If G and H are mutually exclusive events, then:

$$\Pr(G \text{ or } H) = \Pr(G) + \Pr(H)$$

For example,

$$\Pr(\text{Round or Wrinkled}) = \Pr(\text{Round}) + \Pr(\text{Wrinkled})$$

More generally, if $E_i, i = 1, \ldots r$, are mutually exclusive then

$$\begin{aligned} \Pr(E_1 \text{ or } \ldots \text{ or } E_r) &= \Pr(E_1) + \ldots + \Pr(E_r) \\ &= \sum_i \Pr(E_i) \end{aligned}$$

Complementary Probability

If $\Pr(E)$ is the probability that E is true then $\Pr(\bar{E})$ denotes the probability that E is false. Because these two events are mutually exclusive

$$\Pr(E \text{ or } \bar{E}) \;=\; \Pr(E) + \Pr(\bar{E})$$

and they are also exhaustive in that between them they cover all possibilities – one or other of them must be true. So,

$$\begin{aligned} \Pr(E) + \Pr(\bar{E}) &= 1 \\ \Pr(\bar{E}) &= 1 - \Pr(E) \end{aligned}$$

The probability that E is false is one minus the probability it is true.

Third Law of Probability

For any two events, G and H, the third law can be written:

$$\Pr(G \text{ and } H) \;=\; \Pr(G)\Pr(H|G)$$

There is no reason why G should precede H and the law can also be written:

$$\Pr(G \text{ and } H) \;=\; \Pr(H)\Pr(G|H)$$

For example

$$\begin{aligned}\Pr(\text{Seed is round \& is type AA}) \;&=\; \Pr(\text{Seed is round}|\text{Seed is type AA}) \\ &\quad \times \Pr(\text{Seed is type AA}) \\ &=\; 1 \times p_A^2\end{aligned}$$

Independent Events

If the information that H is true does nothing to change uncertainty about G, then

$$\Pr(G|H) \;=\; \Pr(G)$$

and

$$\Pr(H \text{ and } G) \;=\; \Pr(H)\Pr(G)$$

Events G, H are independent.

Law of Total Probability

If G, H are two mutually exclusive and exhaustive events (so that $H = \bar{G} = \text{not} - G$), then for any other event E, the law of total probability states that

$$\Pr(E) \;=\; \Pr(E|G)\Pr(G) + \Pr(E|H)\Pr(H)$$

This generalizes to any set of mutually exclusive and exhaustive events $\{S_i\}$:

$$\Pr(E) \;=\; \sum_i \Pr(E|S_i)\Pr(S_i)$$

For example

$$\begin{aligned}\Pr(\text{Seed is round}) \;&=\; \Pr(\text{Round}|\text{Type AA})\Pr(\text{Type AA}) \\ &\quad + \Pr(\text{Round}|\text{Type Aa})\Pr(\text{Type Aa}) \\ &\quad + \Pr(\text{Round}|\text{Type aa})\Pr(\text{Type aa}) \\ &=\; 1 \times p_A^2 + 1 \times 2p_A p_a + 0 \times p_a^2 = p_A(1 + p_A)\end{aligned}$$

Bayes' Theorem

Bayes' theorem relates $\Pr(G|H)$ to $\Pr(H|G)$:

$$\begin{aligned}\Pr(G|H) &= \frac{\Pr(GH)}{\Pr(H)}, \text{ from third law} \\ &= \frac{\Pr(H|G)\Pr(G)}{\Pr(H)}, \text{ from third law}\end{aligned}$$

If $\{G_i\}$ are exhaustive and mutually exclusive, Bayes' theorem can be written as

$$\Pr(G_i|H) = \frac{\Pr(H|G_i)\Pr(G_i)}{\sum_i \Pr(H|G_i)\Pr(G_i)}$$

Mendel's Data

Model: seed shape governed by gene **A** with alleles A, a:

Genotype	Phenotype
AA	Round
Aa	Round
aa	Wrinkled

Cross two inbred lines: AA and aa. All offspring (F_1 generation) are Aa, and so have round seeds.

F_2 generation

Self an F_1 plant: each allele it transmits is equally likely to be A or a, and alleles are independent, so for F_2 generation:

$$\begin{aligned}
\Pr(AA) &= \Pr(A)\Pr(A) = 0.25 \\
\Pr(Aa) &= \Pr(A)\Pr(a) + \Pr(a)\Pr(A) = 0.5 \\
\Pr(aa) &= \Pr(a)\Pr(a) = 0.25
\end{aligned}$$

Probability that an F_2 seed (observed on F_1 parental plant) is round:

$$\begin{aligned}
\Pr(\text{Round}) &= \Pr(\text{Round}|AA)\Pr(AA) \\
&\quad + \Pr(\text{Round}|Aa)\Pr(Aa) \\
&\quad + \Pr(\text{Round}|aa)\Pr(aa) \\
&= 1 \times 0.25 + 1 \times 0.5 + 0 \times 0.25 \\
&= 0.75
\end{aligned}$$

F_2 generation

What are the proportions of AA and Aa among F_2 plants with round seeds? From Bayes' Theorem:

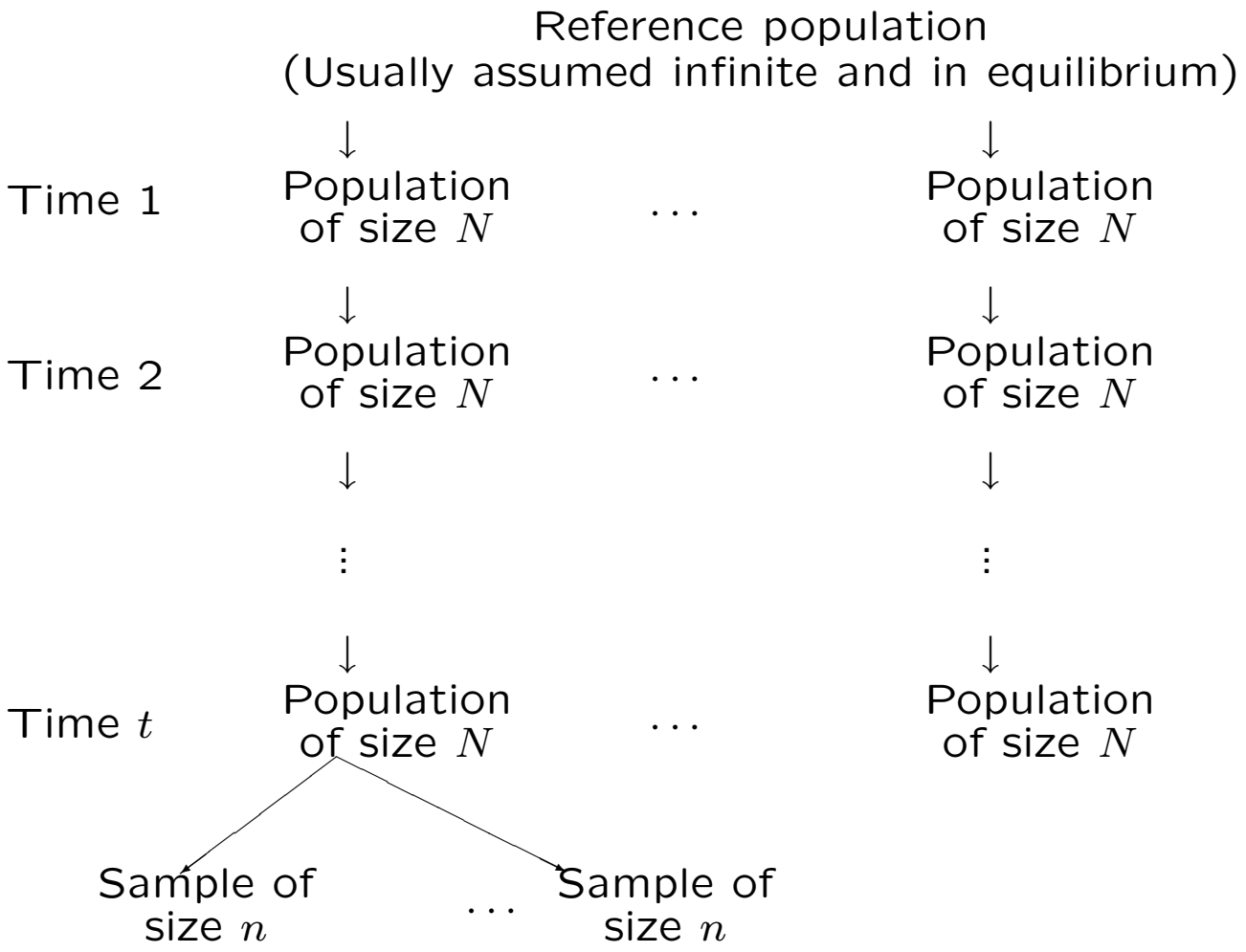
$$\begin{aligned}
\Pr(F_2 = AA | F_2 \text{ Round}) &= \frac{\Pr(F_2 \text{ Round} | AA) \Pr(F_2\ AA)}{\Pr(F_2 \text{ round})} \\
&= \frac{1 \times \frac{1}{4}}{\frac{3}{4}} \\
&= \frac{1}{3}
\end{aligned}$$

Sampling

Statistical sampling: The variation among repeated samples from the same population is analogous to "fixed" sampling. Inferences can be made about that particular population.

Genetic sampling: The variation among replicate (conceptual) populations is analogous to "random" sampling. Inferences are made to all populations with the same history.

Classical Model



Coalescent Theory

An alternative framework works with genealogical history of a sample of alleles. There is a tree linking all alleles in a current sample to the "most recent common ancestral allele." Allelic variation due to mutations since that ancestral allele.

The coalescent approach requires mutation and may be more appropriate for long-term evolution and analyses involving more than one species. The classical approach allows mutation but does not require it: within one species variation among populations may be due primarily to drift.

ALLELE FREQUENCIES

Properties of Estimators

Consistency	Increasing accuracy as sample size increases
Unbiasedness	Expected value is the parameter
Efficiency	Smallest variance
Sufficiency	Contains all the information in the data about parameter

Binomial Distribution

Most population genetic data consists of numbers of observations in some categories. The values and frequencies of these counts form a *distribution*.

Toss a coin n times, and note the number of heads. There are $(n+1)$ outcomes, and the number of times each outcome is observed in many sets of n tosses gives the sampling distribution. Or: sample n alleles from a population and observe x copies of type A.

Binomial distribution

If every toss has the same chance p of giving a head:

Probability of x heads in a row is

$$p \times p \times \ldots \times p \;=\; p^x$$

Probability of $n - x$ tails in a row is

$$(1-p) \times (1-p) \times \ldots \times (1-p) \;=\; (1-p)^{n-x}$$

The number of ways of ordering x heads and $n - x$ tails among n outcomes is $n!/[x!(n-x)!]$.

The binomial probability of x successes in n trials is

$$\Pr(x|p) \;=\; \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Binomial Likelihood

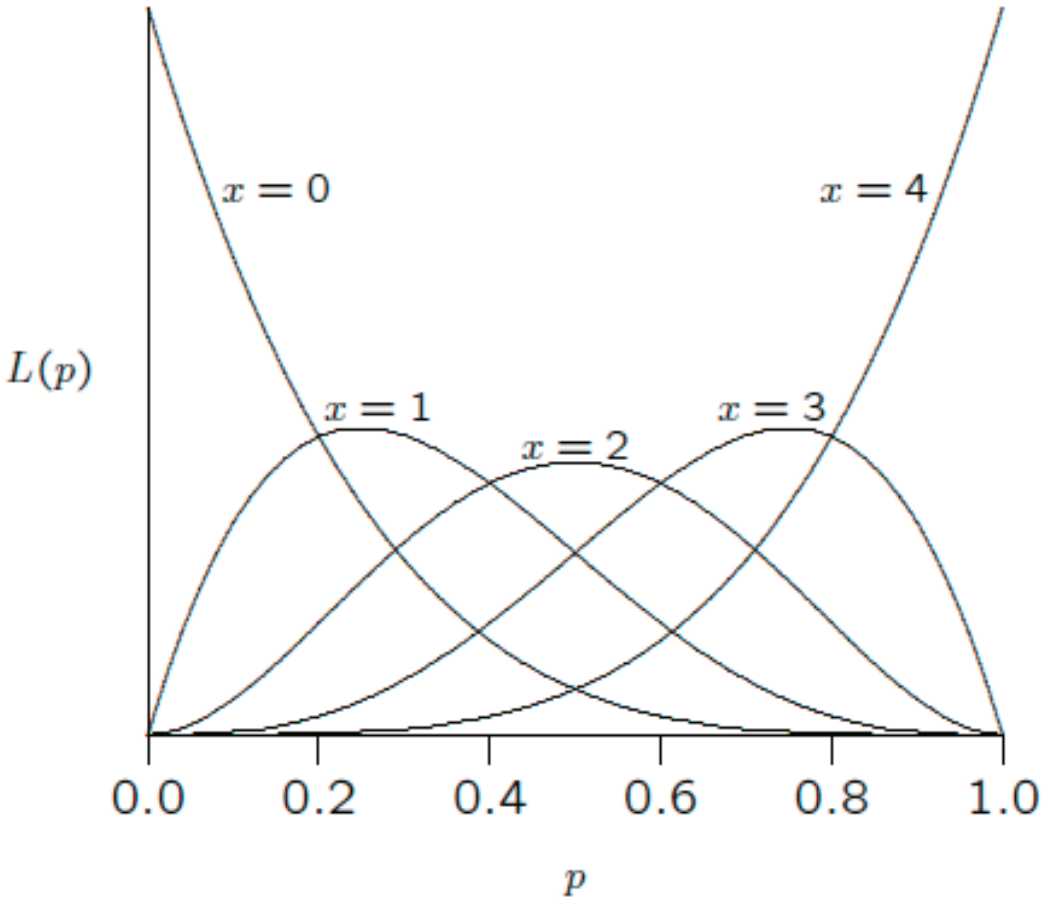
The quantity $\Pr(x|p)$ is the *probability of the data*, x successes in n trials, when each trial has probability p of success.

The same quantity, written as $L(p|x)$, is the *likelihood of the parameter*, p, when the value x has been observed. The terms that do not involve p are not needed, so

$$L(p|x) \quad \propto \quad p^x(1-p)^{(n-x)}$$

Each value of x gives a different likelihood curve, and each curve points to a p value with maximum likelihood. This leads to *maximum likelihood estimation*.

Likelihood $L(p|x, n = 4)$



Binomial Mean

If there are n trials, each of which has probability p of giving a success, the *mean* or the *expected number* of successes is np.

The *sample proportion* of successes is

$$\tilde{p} \;=\; \frac{x}{n}$$

(This is also the maximum likelihood estimate of p.)

The expected, or *mean*, value of $\tilde{p}$ is p.

$$\mathcal{E}(\tilde{p}) \;=\; p$$

Binomial Variance

The expected value of the squared difference between the number of successes and its mean, $(x - np)^2$, is $np(1-p)$. This is the *variance* of the number of successes in n trials, and indicates the spread of the distribution.

The variance of the sample proportion $\tilde{p}$ is

$$\text{Var}(\tilde{p}) = \frac{p(1-p)}{n}$$

Normal Approximation

Provided np is not too small (e.g. not less than 5), the binomial distribution can be approximated by the normal distribution with the same mean and variance. In particular:

$$\tilde{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

To use the normal distribution in practice, change to the *standard normal* variable z with a mean of 0, and a variance of 1:

$$z = \frac{\tilde{p} - p}{\sqrt{p(1-p)/n}}$$

For a standard normal, 95% of the values lie between ± 1.96. The normal approximation to the binomial therefore implies that 95% of the values of $\tilde{p}$ lie in the range

$$p \pm 1.96\sqrt{p(1-p)/n}$$

Confidence Intervals

A 95% confidence interval is a variable quantity. It has end-points which vary with the sample. Expect that 95% of samples will lead to an interval that includes the unknown true value p_c.

The standard normal variable z has 95% of its values between -1.96 and $+1.96$. This suggests that a 95% confidence interval for the binomial parameter p is

$$\tilde{p} \ \pm \ 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$$

Confidence Intervals

For samples of size 10, the 11 possible confidence intervals are:

$\tilde{p}_C$	Confidence Interval	
0.0	0.0 ± 0.00	$0.00, 0.00$
0.1	$0.1 \pm 2\sqrt{0.009}$	$0.00, 0.29$
0.2	$0.2 \pm 2\sqrt{0.016}$	$0.00, 0.45$
0.3	$0.3 \pm 2\sqrt{0.021}$	$0.02, 0.58$
0.4	$0.4 \pm 2\sqrt{0.024}$	$0.10, 0.70$
0.5	$0.5 \pm 2\sqrt{0.025}$	$0.19, 0.81$
0.6	$0.6 \pm 2\sqrt{0.024}$	$0.30, 0.90$
0.7	$0.7 \pm 2\sqrt{0.021}$	$0.42, 0.98$
0.8	$0.8 \pm 2\sqrt{0.016}$	$0.55, 1.00$
0.9	$0.9 \pm 2\sqrt{0.009}$	$0.71, 1.00$
1.0	1.0 ± 0.00	$1.00, 1.00$

Can modify interval a little by extending it by the "continuity correction" $\pm 1/2n$ in each direction.

Confidence Intervals

To be 95% sure that the estimate is no more than 0.01 from the true value, $1.96\sqrt{p(1-p)/n}$ should be less than 0.01. The widest confidence interval is when $p = 0.5$, and then need

$$0.01 \geq 1.96\sqrt{0.5 \times 0.5/n}$$

which means that $n \geq 10,000$. For a width of 0.03 instead of 0.01, $n \approx 1,000$.

If the true value of p was about 0.05, however,

$$\begin{aligned} 0.01 &\geq 2\sqrt{0.05 \times 0.95/n} \\ n &\geq 1,900 \approx 2,000 \end{aligned}$$

Exact Confidence Intervals: One-sided

The normal-based confidence intervals are constructed to be symmetric about the sample value, unless the interval goes outside the interval from 0 to 1. They are therefore less satisfactory the closer the true value is to 0 or 1.

More accurate confidence limits follow from the binomial distribution exactly. For events with low probabilities p, how large could p be for there to be at least a 5% chance of seeing no more than x (i.e. $0, 1, 2, \ldots x$) occurrences of that event among n events. If this upper bound is p_U,

$$\sum_{k=0}^{x} \Pr(k) \geq 0.05$$

$$\sum_{k=0}^{x} \binom{n}{k} p_U^k (1-p_U)^{n-k} \geq 0.05$$

If $x = 0$, then $(1-p_U)^n \geq 0.05$ or $p_U \leq 1 - 0.05^{1/n}$ and this is 0.0295 if $n = 100$. More generally $p_U \approx 3/n$ when $x = 0$.

Exact Confidence Intervals: Two-sided

Now want to know how large p could be for there to be at least a 2.5% chance of seeing no more than x (i.e. $0, 1, 2 \ldots x$) occurrences, and in knowing how small p could be for there to be at least a 2.5% chance of seeing at least x (i.e. $x, x{+}1, x{+}2, \ldots n$) occurrences then we need

$$\sum_{k=0}^{x} \binom{n}{k} p_U^k (1-p_U)^{n-k} \geq 0.025$$

$$\sum_{k=x}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} \geq 0.025$$

The second of these equations may be written as

$$\sum_{k=0}^{x-1} \binom{n}{k} p_L^k (1-p_L)^{n-k} \leq 0.975$$

If $x = n$, then $p_L^n \leq 0.975$ or $p_L \leq 0.975^{1/n}$ and this is 0.9997 if $n = 100$. Interval is not symmetric if $p \neq 0.5$.

Bootstrapping

An alternative method for constructing confidence intervals uses *numerical resampling*. A set of samples is drawn, with replacement, from the original sample to mimic the variation among samples from the original population. Each new sample is the same size as the original sample, and is called a *bootstrap sample*.

The middle 95% of the sample values $\tilde{p}$ from a large number of bootstrap samples provides a 95% confidence interval.

Multinomial Distribution

Toss two coins n times. For each double toss, the probabilities of the three outcomes are:

2 heads	$p_{HH} = 1/4$
1 head, 1 tail	$p_{HT} = 1/2$
2 tails	$p_{TT} = 1/4$

The probability of x lots of 2 heads is $(p_{HH})^x$, etc.

The numbers of ways of ordering x, y, z occurrences of the three outcomes is $n!/[x!y!z!]$ where $n = x + y + z$.

The multinomial probability for x of HH, and y of HT or TH and z of TT in n trials is:

$$\Pr(x, y, z) = \frac{n!}{x!y!z!}(p_{HH})^x(p_{HT})^y(p_{TT})^z$$

Multinomial Variances and Covariances

If $\{p_i\}$ are the probabilities for a series of categories, the sample proportions $\tilde{p}_i$ from a sample of n observations have these properties:

$$\begin{aligned} \mathcal{E}(\tilde{p}_i) &= p_i \\ \text{Var}(\tilde{p}_i) &= \frac{1}{n} p_i(1-p_i) \\ \text{Cov}(\tilde{p}_i, \tilde{p}_j) &= -\frac{1}{n} p_i p_j, \quad i \neq j \end{aligned}$$

The covariance is defined as $\mathcal{E}[(\tilde{p}_i - p_i)(\tilde{p}_j - p_j)]$.

For the sample counts:

$$\begin{aligned} \mathcal{E}(n_i) &= np_i \\ \text{Var}(n_i) &= np_i(1-p_i \\ \text{Cov}(n_i, n_j) &= -np_i p_j, \quad i \neq j \end{aligned}$$

Allele Frequency Sampling Distribution

If a locus has alleles A and a, in a sample of size n the allele counts are sums of genotype counts:

$$\begin{aligned} n &= n_{AA} + n_{Aa} + n_{aa} \\ n_A &= 2n_{AA} + n_{Aa} \\ n_a &= 2n_{aa} + n_{Aa} \\ 2n &= n_A + n_a \end{aligned}$$

Genotype counts in a random sample are multinomially distributed. What about allele counts? Approach this question by calculating variance of n_A.

Within-population Variance

$$\begin{aligned}
\text{Var}(n_A) &= \text{Var}(2n_{AA} + n_{Aa}) \\
&= \text{Var}(2n_{AA}) + 2\text{Cov}(2n_{AA}, n_{Aa}) + \text{Var}(n_{Aa}) \\
&= 2np_A(1-p_A) + 2n(P_{AA} - p_A^2)
\end{aligned}$$

This is not the same as the binomial variance $2np_A(1-p_A)$ unless $P_{AA} = p_A^2$. In general, the allele frequency distribution is not binomial.

The variance of the sample allele frequency $\tilde{p}_A = n_A/(2n)$ can be written as

$$\text{Var}(\tilde{p}_A) = \frac{p_A(1-p_A)}{2n} + \frac{P_{AA} - p_A^2}{2n}$$

Within-population Variance

It is convenient to reparameterize genotype frequencies with the (within-population) *inbreeding coefficient* f:

$$\begin{aligned} P_{AA} &= p_A^2 + f p_A(1-p_A) \\ P_{Aa} &= 2p_A p_a - 2f p_A p_a \\ P_{aa} &= p_a^2 + f p_a(1-p_a) \end{aligned}$$

Then the variance can be written as

$$\mathrm{Var}(\tilde{p}_A) = \frac{p_A(1-p_A)(1+f)}{2n}$$

This variance is different from the binomial variance of $p_A(1-p_A)/2n$.

Bounds on f

Since

$$\begin{aligned} p_A \geq P_{AA} &= p_A^2 + f p_A(1-p_A) \geq 0 \\ p_a \geq P_{aa} &= p_a^2 + f p_a(1-p_a) \geq 0 \end{aligned}$$

there are bounds on f:

$$\begin{aligned} -p_A/(1-p_A) \leq &\ f \ \leq 1 \\ -p_a/(1-p_a) \leq &\ f \ \leq 1 \end{aligned}$$

or

$$\max\left(-\frac{p_A}{p_a}, -\frac{p_a}{p_A}\right) \leq \ f \ \leq 1$$

This range of values is [-1,1] when $p_A = p_a$.

Indicator Variables

A very convenient way to derive many statistical genetic results is to define an indicator variable x_{ij} for allele j in individual i:

$$x_{ij} = \begin{cases} 1 & \text{if allele is } A \\ 0 & \text{if allele is not } A \end{cases}$$

Then

$$\begin{aligned} \mathcal{E}(x_{ij}) &= p_A \\ \mathcal{E}(x_{ij}^2) &= p_A \\ \mathcal{E}(x_{ij}x_{ij'}) &= P_{AA} \end{aligned}$$

If there is random sampling, individuals are independent, and

$$\mathcal{E}(x_{ij}x_{i'j'}) = \mathcal{E}(x_{ij})\mathcal{E}(x_{i'j'}) = p_A^2$$

Intraclass Correlation

The inbreeding coefficient is the correlation of the indicator variables for the two alleles at a locus carried by an individual. This is because:

$$\begin{aligned} \mathsf{Var}(x_{ij}) &= \mathcal{E}(x_{ij}^2) - [\mathcal{E}(x_{ij})]^2 \\ &= p_A(1-p_A) \\ &= \mathsf{Var}(x_{ij'}), \;\; j \neq j' \end{aligned}$$

and

$$\begin{aligned} \mathsf{Cov}(x_{ij}, x_{ij'}) &= \mathcal{E}(x_{ij}x_{ij'}) - [\mathcal{E}(x_{ij})][\mathcal{E}(x_{ij'})], \;\; j \neq j' \\ &= P_{AA} - p_A^2 \\ &= f p_A(1-p_A) \end{aligned}$$

so

$$\mathsf{Corr}(x_{ij}, x_{ij'}) = \frac{\mathsf{Cov}(x_{ij}, x_{ij'})}{\sqrt{\mathsf{Var}(x_{ij})\mathsf{Var}(x_{ij'})}} = f$$

Maximum Likelihood Estimation: Binomial

For binomial sample of size n, the likelihood of p_A for n_A alleles of type A is

$$L(p_A|n_A) \;=\; C(p_A)^{n_A}(1-p_A)^{n-n_A}$$

and is maximized when

$$\frac{\partial L(p_A|n_A)}{\partial p_A} = 0 \quad \text{or when} \quad \frac{\partial \ln L(p_A|n_a)}{\partial p_A} = 0$$

Now

$$\ln L(p_A|n_A) \;=\; \ln C + n_A \ln(p_A) + (n-n_A)\ln(1-p_A)$$

so

$$\frac{\partial \ln L(p_A|n_A)}{\partial p_A} \;=\; \frac{n_A}{p_A} - \frac{n-n_A}{1-p_A}$$

and this is zero when $p_A = \widehat{p}_A = n_A/n$.

Maximum Likelihood Estimation: Multinomial

If $\{n_i\}$ are multinomial with parameters n and $\{Q_i\}$, then the MLE's of Q_i are n_i/n. This will always hold for genotype proportions, but not always for allele proportions.

For two alleles, the MLE's for genotype proportions are:

$$\begin{aligned} \hat{P}_{AA} &= n_{AA}/n \\ \hat{P}_{Aa} &= n_{Aa}/n \\ \hat{P}_{aa} &= n_{aa}/n \end{aligned}$$

Does this lead to estimates of allele proportions and the within-population inbreeding coefficient?

$$\begin{aligned} P_{AA} &= p_A^2 + f p_A(1-p_A) \\ P_{Aa} &= 2p_A(1-p_A) - 2f p_A(1-p_A) \\ P_{aa} &= (1-p_A)^2 + f p_A(1-p_A) \end{aligned}$$

Maximum Likelihood Estimation

The likelihood function for p_A, f is

$$L(p_A, f) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}[p_A^2 + p_A(1-p_A)f]^{n_{AA}}$$

$$\times[2p_A(1-p_A)f]^{n_{Aa}}[(1-p_A)^2 + p_A(1-p_A)f]^{n_{aa}}$$

and it is difficult to find, analytically, the values of p_A and f that maximize this function or its logarithm.

There is an alternative way of finding maximum likelihood estimates in this case: equating the observed and expected values of the genotype frequencies.

Bailey's Method

Because the number of parameters (2) equals the number of degrees of freedom in this case, we can just equate observed and expected (using the estimates of p_A and f) genotype proportions

$$\begin{aligned} n_{AA}/n &= \hat{p}_A^2 + \hat{f}\hat{p}_A(1-\hat{p}_A) \\ n_{Aa}/n &= 2\hat{p}_A(1-\hat{p}_A) - 2\hat{f}\hat{p}_A(1-\hat{p}_A) \\ n_{aa}/n &= (1-\hat{p}_A)^2 + \hat{f}\hat{p}_A(1-\hat{p}_A) \end{aligned}$$

Solving these equations (e.g. by adding the first equation to half the second equation) for $\hat{p}_A$ and $\hat{f}$:

$$\hat{p}_A = \frac{2n_{AA}+n_{Aa}}{2n} = \tilde{p}_A$$

$$\hat{f} = \frac{4n_{AA}n_{aa}-n_{Aa}^2}{(2n_{AA}+n_{Aa})(2n_{aa}+n_{Aa})} = 1 - \frac{\tilde{P}_{Aa}}{2\tilde{p}_A\tilde{p}_a}$$

Three-allele Case

With three alleles, there are six genotypes and 5 df. To use Bailey's method, would need five parameters: 2 allele frequencies and 3 inbreeding coefficients:

$$\begin{aligned}
P_{11} &= p_1^2 + f_{12}p_1p_2 + f_{13}p_1p_3 \\
P_{12} &= 2p_1p_2 - 2f_{12}p_1p_2 \\
P_{22} &= p_2^2 + f_{12}p_1p_2 + f_{23}p_2p_3 \\
P_{13} &= 2p_1p_3 - 2f_{13}p_1p_3 \\
P_{23} &= 2p_2p_3 - 2f_{23}p_2p_3 \\
P_{33} &= p_3^2 + f_{13}p_1p_3 + f_{23}p_2p_3
\end{aligned}$$

We would generally prefer to have only one inbreeding coefficient f. It is a difficult numerical problem to find the MLE for f.

Method of Moments

An alternative to maximum likelihood estimation is the method of moments (MoM) where observed values of statistics are set equal to their expected values. In general, this does not lead to unique estimates or to estimates with variances as small as those for maximum likelihood. (Bailey's method is for the special case where the MLEs are also MoM estimates.)

Method of Moments

For the inbreeding coefficient at loci with m alleles, two different MoM estimates are

$$\begin{aligned}
\hat{f}_W &= \frac{\sum_{u=1}^{m}(\tilde{P}_{uu}-\tilde{p}_u^2)+\frac{1}{2n}\sum_{u=1}^{m}(\tilde{p}_u-\tilde{P}_{uu})}{\sum_{u=1}^{m}\tilde{p}_u(1-\tilde{p}_u)-\frac{1}{2n}\sum_{u=1}^{m}(\tilde{p}_u-\tilde{P}_{uu})} \\
&\approx \frac{\sum_{u=1}^{m}(\tilde{P}_{uu}-\tilde{p}_u^2)}{\sum_{u=1}^{m}\tilde{p}_u(1-\tilde{p}_u)} \\
\hat{f}_H &= \frac{1}{m-1}\sum_{u=1}^{m}\left(\frac{\tilde{P}_{uu}-\tilde{p}_u^2}{\tilde{p}_u}\right)
\end{aligned}$$

For loci with two alleles, $m = 2$, the two moment estimates are equal to each other and to the maximum likelihood estimate:

$$\hat{f}_W = \hat{f}_H = 1 - \frac{\tilde{P}_{Aa}}{2\tilde{p}_A\tilde{p}_a}$$

Expectations of Moment Estimates

The expected values of the estimated inbreeding coefficients can be found by using the results

$$\begin{aligned}
\mathcal{E}(\tilde{P}_{uu}) &= P_{uu} = p_u^2 + f p_u(1-p_u) \\
\mathcal{E}(\tilde{p}_u) &= p_u \\
\mathcal{E}(\tilde{p}_u^2) &= p_u^2 + \frac{1}{2n} p_u(1-p_u)(1+f)
\end{aligned}$$

Then, approximating the expectation of a ratio by the ratio of expectations:

$$\begin{aligned}
\mathcal{E}(\hat{f}_W) &\approx \frac{f\frac{n-1}{n}(1-\sum_{u=1}^m p_u^2)}{\frac{n-1}{n}(1-\sum_{u=1}^m p_u^2)} = f \\
\mathcal{E}(\hat{f}_H) &\approx \frac{1}{m-1}\sum_{u=1}^m \left(\frac{p_u(1-p_u)(f-\frac{1+f}{2n})}{p_u}\right) \\
&= f - \frac{1+f}{2n} \approx f
\end{aligned}$$

MLE for Recessive Alleles

Suppose allele a is recessive to allele A. If there is Hardy-Weinberg equilibrium, the likelihood for the two phenotypes is

$$\begin{aligned} L(p_a) &= (1-p_a^2)^{n-n_{aa}}(p_a^2)^{n_{aa}} \\ \ln(L(p_a) &= (n-n_{aa})\ln(1-p_a^2) + 2n_{aa}\ln(p_a) \end{aligned}$$

where there are n_{aa} individuals of type aa and $n-n_{aa}$ of type A. Differentiating wrt p_a:

$$\frac{\partial \ln L(p_a)}{\partial p_a} = -\frac{2p_a(n-n_{aa})}{1-p_a^2} + \frac{2n_{aa}}{p_a}$$

Setting this to zero leads to an equation that can be solved explicitly: $p_a^2 = n_{aa}/n$. No need for iteration.

EM Algorithm for Recessive Alleles

An alternative way of finding maximum likelihood estimates when there are "missing data" involves *Estimation* of the missing data and then *Maximization* of the likelihood. For a locus with allele A dominant to a the missing information is the frequencies $(1-p_a)^2$ of AA, and $2p_a(1-p_a)$ of Aa genotypes. Only the joint frequency $(1-p_a^2)$ of $AA+Aa$ can be observed.

Estimate the missing genotype counts (assuming independence of alleles):

$$n_{AA} = \frac{(1-p_a)^2}{1-p_a^2}(n-n_{aa}) = \frac{(1-p_a)(n-n_{aa})}{(1+p_a)}$$

$$n_{Aa} = \frac{2p_a(1-p_a)}{1-p_a^2}(n-n_{aa}) = \frac{2p_a(n-n_{aa})}{(1+p_a)}$$

EM Algorithm for Recessive Alleles

Maximize the likelihood (using Bailey's method):

$$\begin{aligned}
\widehat{p}_a &= \frac{n_{Aa} + 2n_{aa}}{2n} \\
&= \frac{1}{2n}\left(\frac{2p_a(n - n_{aa})}{(1 + p_a)} + 2n_{aa}\right) \\
&= \frac{2(np_a + n_{aa})}{2n(1 + p_a)}
\end{aligned}$$

An initial estimate p_a is put into the right hand side to give an updated estimated $\widehat{p}_a$ on the left hand side. This is then put back into the right hand side to give an iterative equation for p_a.

This procedure also has explicit solution $\widehat{p}_a = \sqrt{(n_{aa}/n)}$.

EM Algorithm for Two Loci

For two loci with two alleles each, the ten two-locus frequencies are:

Genotype	Actual	Expected	Genotype	Actual	Expected
AB/AB	P_{AB}^{AB}	p_{AB}^2	AB/Ab	P_{Ab}^{AB}	$2p_{AB}p_{Ab}$
AB/aB	P_{aB}^{AB}	$2p_{AB}p_{aB}$	AB/ab	P_{ab}^{AB}	$2p_{AB}p_{ab}$
Ab/Ab	P_{Ab}^{Ab}	p_{Ab}^2	Ab/aB	P_{aB}^{Ab}	$2p_{Ab}p_{aB}$
Ab/ab	P_{ab}^{Ab}	$2p_{Ab}p_{ab}$	aB/aB	P_{aB}^{aB}	p_{aB}^2
aB/ab	P_{ab}^{aB}	$2p_{aB}p_{ab}$	ab/ab	P_{ab}^{ab}	p_{ab}^2

EM Algorithm for Two Loci

Gamete frequencies are marginal sums:

$$
\begin{aligned}
p_{AB} &= P_{AB}^{AB} + \frac{1}{2}(P_{Ab}^{AB} + P_{aB}^{AB} + P_{ab}^{AB}) \\
p_{Ab} &= P_{Ab}^{Ab} + \frac{1}{2}(P_{AB}^{Ab} + P_{ab}^{Ab} + P_{aB}^{Ab}) \\
p_{aB} &= P_{aB}^{aB} + \frac{1}{2}(P_{AB}^{aB} + P_{ab}^{aB} + P_{Ab}^{aB}) \\
p_{ab} &= P_{ab}^{ab} + \frac{1}{2}(P_{Ab}^{ab} + P_{aB}^{ab} + P_{AB}^{ab})
\end{aligned}
$$

Can arrange gamete frequencies as two-way table to show that only one of them is unknown when the allele frequencies are known:

p_{AB}	p_{Ab}	p_A
p_{aB}	p_{ab}	p_a
p_B	p_b	1

EM Algorithm for Two Loci

The two double heterozygote frequencies P_{ab}^{AB}, P_{aB}^{Ab} are "missing data."

Assume initial value of p_{AB} and *Estimate* the missing counts:

$$\begin{aligned} n_{ab}^{AB} &= \frac{p_{AB}p_{ab}}{p_{AB}p_{ab}+p_{Ab}p_{aB}} n_{AaBb} \\ n_{aB}^{Ab} &= \frac{p_{Ab}p_{aB}}{p_{AB}p_{ab}+p_{Ab}p_{aB}} n_{AaBb} \end{aligned}$$

and then *Maximize* the likelihood by setting

$$p_{AB} = \frac{1}{2n}\left(2n_{AB}^{AB}+n_{Ab}^{AB}+n_{aB}^{AB}+n_{ab}^{AB}\right)$$

Example

As an example, consider the data

	BB	Bb	bb	Total
AA	$n_{AABB} = 5$	$n_{AABb} = 3$	$n_{AAbb} = 2$	$n_{AA} = 10$
Aa	$n_{AaBB} = 3$	$n_{AaBb} = 2$	$n_{Aabb} = 0$	$n_{Aa} = 5$
aa	$n_{aaBB} = 0$	$n_{aaBb} = 0$	$n_{aabb} = 0$	$n_{aa} = 0$
Total	$n_{BB} = 8$	$n_{Bb} = 5$	$n_{bb} = 2$	$n = 15$

There is one unknown gamete count $x = n_{AB} = 2np_{AB}$ for AB:

	B	b	Total
A	$n_{AB} = x$	$n_{Ab} = 25 - x$	$n_A = 25$
a	$n_{aB} = 21 - x$	$n_{ab} = x - 16$	$n_a = 5$
Total	$n_B = 21$	$n_b = 9$	$2n = 30$

$$21 \geq x \geq 16$$

Example

EM iterative equation:

$$
\begin{aligned}
x' &= 2n_{AABB} + n_{AABb} + n_{AaBB} + n_{AB/ab} \\
&= 2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{2p_{AB}p_{ab}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} n_{AaBb} \\
&= 10 + 3 + 3 + 2 \times \frac{2x(x-16)}{2x(x-16) + 2(25-x)(21-x)} \\
&= 16 + \frac{x(x-16)}{x(x-16) + (25-x)(21-x)}
\end{aligned}
$$

Note that, if $x = 16$, then $x' = 16$!

Example

A good starting value would assume independence of A and B alleles: $x = 2n * p_A * p_B = (25 \times 21/30) = 17.5$.

Successive iterates are:

Iterate	x value
0	17.5000
1	17.0000
2	16.6939
3	16.4893
4	16.3473
5	16.2472
...	...

The solution is actually $x = 16$. This particular example does not have convergence to the MLE for some starting values for x.

ALLELIC ASSOCIATION

Hardy-Weinberg Law

For a random mating population, expect that genotype frequencies are products of allele frequencies.

For a locus with two alleles, A, a:

$$\begin{aligned} P_{AA} &= (p_A)^2 \\ P_{Aa} &= 2p_A p_a \\ P_{aa} &= (p_a)^2 \end{aligned}$$

These are also the results of setting the inbreeding coefficient f to zero.

For a locus with several alleles A_i:

$$\begin{aligned} P_{A_iA_i} &= (p_{A_i})^2 \\ P_{A_iA_j} &= 2p_{A_i}p_{A_j} \end{aligned}$$

Inference about HWE

Departures from HWE can be described by the within-population inbreeding coefficient f. This has an MLE that can be written as

$$\widehat{f} \;=\; \frac{4n_{AA}n_{aa} - n_{Aa}^2}{(2n_{AA}+n_{Aa})(2n_{aa}+n_{Aa})}$$

and we can use "Delta method" to find

$$\begin{aligned} \mathcal{E}(\widehat{f}) &= f \\ \mathrm{Var}(\widehat{f}) &\approx \frac{1}{2np_Ap_a}(1-f)[2p_Ap_a(1-f)(1-2f)+f(2-f)] \end{aligned}$$

If $\widehat{f}$ is assumed to be normally distributed then, $(\widehat{f}-f)/\sqrt{\mathrm{Var}(\widehat{f})} \sim N(0,1)$. When H_0 is true, the square of this quantity has a chi-square distribution.

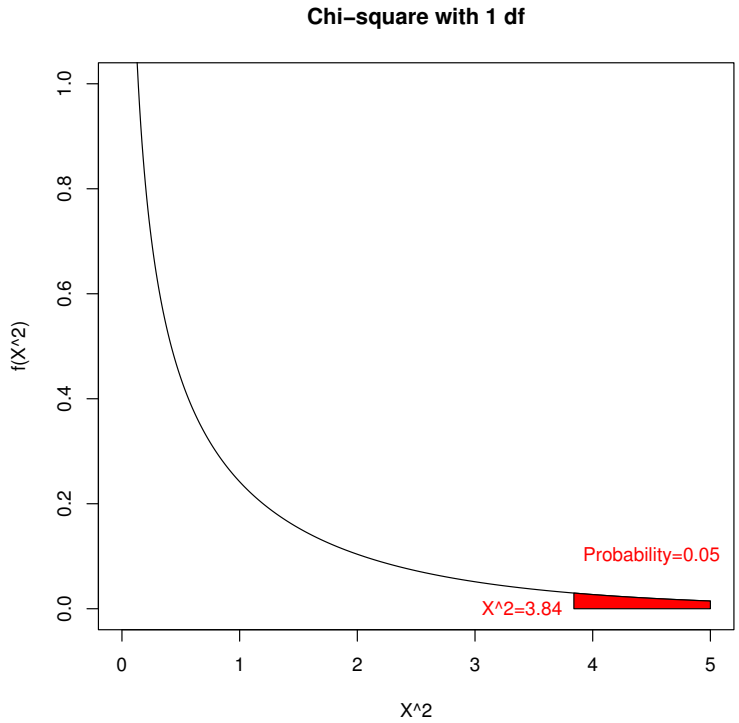
Inference about HWE

Since $\text{Var}(\widehat{f}) = 1/n$ when $f = 0$:

$$\begin{aligned} X^2 &= \left(\frac{\widehat{f} - f}{\sqrt{\text{Var}(\widehat{f})}} \right)^2 \\ &= \frac{\widehat{f}^2}{1/n} \\ &= n\widehat{f}^2 \end{aligned}$$

is appropriate for testing $H_0 : f = 0$. When H_0 is true, $X^2 \sim \chi^2_{(1)}$. Reject HWE if $X^2 > 3.84$.

Significance level of HWE test



The area under the chi-square curve to the right of $X^2 = 3.84$ is the probability of rejecting HWE when HWE is true. This is the significance level of the test.

Goodness-of-fit Test

An alternative, but equivalent, test is the goodness-of-fit test.

Genotype	Observed	Expected	$\frac{(\text{Obs.}-\text{Exp.})^2}{\text{Exp.}}$
AA	n_{AA}	$n\tilde{p}_A^2$	$n\tilde{p}_a^2\hat{f}^2$
Aa	n_{Aa}	$2n\tilde{p}_A\tilde{p}_a$	$2n\tilde{p}_A\tilde{p}_a\hat{f}^2$
aa	n_{aa}	$n\tilde{p}_a^2$	$n\tilde{p}_A^2\hat{f}^2$

The test statistic is

$$X^2 = \sum \frac{(\text{Obs.} - \text{Exp})^2}{\text{Exp.}} = n\hat{f}^2$$

Goodness-of-fit Test

Does a sample of 6 AA, 3 Aa, 1 aa support Hardy-Weinberg?

First need to estimate allele frequencies:

$$\tilde{p}_A = \tilde{P}_{AA} + \frac{1}{2}\tilde{P}_{Aa} = 0.75$$

$$\tilde{p}_a = \tilde{P}_{aa} + \frac{1}{2}\tilde{P}_{Aa} = 0.25$$

Then form "expected" counts:

$$\begin{aligned} n_{AA} &= n(\tilde{p}_A)^2 = 5.625 \\ n_{Aa} &= 2n\tilde{p}_A\tilde{p}_a = 3.750 \\ n_{aa} &= n(\tilde{p}_a)^2 = 0.625 \end{aligned}$$

Goodness-of-fit Test

Perform the chi-square test:

Genotype	Observed	Expected	$(\text{Obs.} - \text{Exp.})^2/\text{Exp.}$
AA	6	5.625	0.025
Aa	3	3.750	0.150
aa	1	0.625	0.225
Total	10	10	0.400

Note that $\widehat{f} = 1 - 0.3/(2 \times 0.75 \times 0.25) = 0.2$ and $X^2 = n\widehat{f}^2$.

Sample size determination

Although Fisher's exact test (below) is generally preferred for small samples, the normal or chi-square test has the advantage of simplifying power calculations.

Assuming that $\widehat{f}$ is normally distributed, form the test statistic

$$z = \frac{\widehat{f} - f}{\sqrt{\text{Var}(\widehat{f})}}$$

Under the null hypothesis $H_0 : f = 0$ this is $z_0 = \sqrt{n}\widehat{f}$. For a two-sided test, reject at the α% level if $z_0 \leq z_{\alpha/2}$ or $z_0 \geq z_{1-\alpha/2} = -z_{\alpha/2}$. For a 5% test, reject if $z_0 \leq -1.96$ or $z_0 \geq 1.96$.

Sample size determination

If the hypothesis is false, the normal test statistic is

$$z = \frac{\hat{f} - f}{\sqrt{\text{Var}(\hat{f})}} \approx \sqrt{n}(\hat{f} - f) = z_0 - \sqrt{n}f$$

(using the null-hypothesis value of the variance in the denominator). Suppose $\hat{f} > 0$ so rejection occurs when $z_0 \geq -z_{\alpha/2}$. With this rejection region, the probability of rejecting is $\geq (1 - \beta)$ if the rejection region amounts to $z = z_0 - \sqrt{n}f \geq z_\beta$. i.e.

$$\begin{aligned} -z_{\alpha/2} - \sqrt{n}f &= z_\beta \\ nf^2 &= (z_{\alpha/2} + z_\beta)^2 \end{aligned}$$

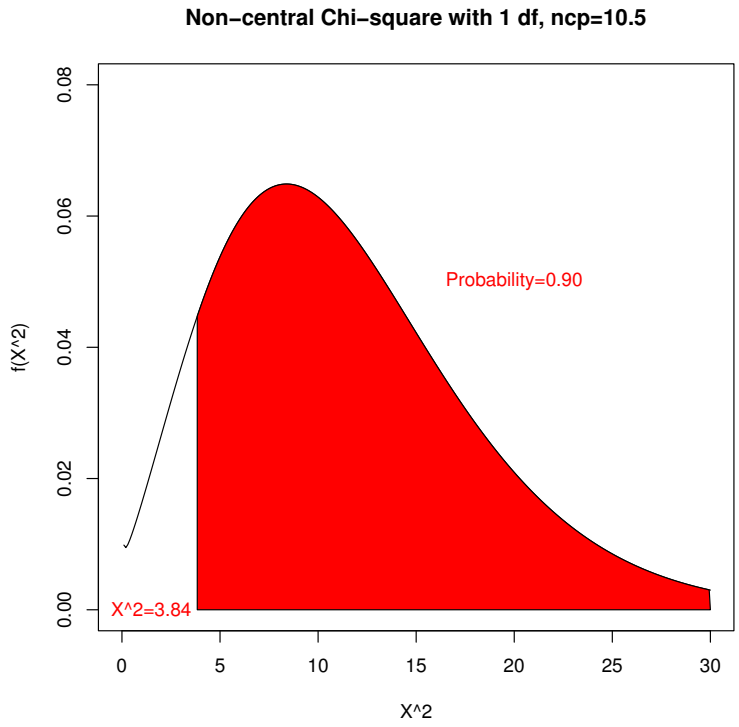
For 5% significance level $-z_{\alpha/2} = 1.96$, and for 90% power $z_\beta = -1.28$ so we need $nf^2 \geq (-1.96 - 1.28)^2 = 10.5$. i.e. n has to be over 100,000 when $f = 0.01$.

Sample size determination

More directly, when the Hardy-Weinberg hypothesis is not true, the test statistic $n\hat{f}^2$ has a non-central chi-square distribution with one degree of freedom (df) and non-centrality parameter $\lambda = nf^2$. To reach 90% power with a 5% significance level, for example, it is necessary that $\lambda \geq 10.5$.

In this one-df case, the non-centrality value follows from percentiles of the standard normal distribution. If z_x is the xth percentile of the standard normal, than for significance level α and power $1 - \beta$, $\lambda = (z_{\alpha/2} + z_\beta)^2$.

Power of HWE test



The area under the non-central chi-square curve to the right of $X^2 = 3.84$ is the probability of rejecting HWE when HWE is false. This is the power of the test. In this plot, the non-centrality parameter is $\lambda = 10.5$.

Significance Levels and p-values

The *significance level* α of a test is the probability of a false rejection. It is specified by the user, and along with the null hypothesis, it determines the rejection region. The specified, or "nominal" value may not be achieved for an actual test.

Once the test has been conducted on a data set, the probability of the observed test statistic, *or a more extreme value*, if the null hypothesis is true is the *p-value*. The chi-square and normal tests shown above give approximate p-values because they use a continuous distribution for discrete data.

An alternative class of tests, "exact tests," use a discrete distribution for discrete data and provide accurate p-values. It may be difficult to construct an exact test with a particular nominal significance level.

Exact HWE Test

The preferred test for HWE is an exact one. The test rests on the assumption that individuals are sampled randomly from a population so that genotype counts have a multinomial distribution:

$$\Pr(n_{AA}, n_{Aa}, n_{aa}) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(P_{AA})^{n_{AA}}(P_{Aa})^{n_{Aa}}(P_{aa})^{n_{aa}}$$

This equation is always true, and when there is HWE ($P_{AA} = p_A^2$ etc.) there is the additional result that the allele counts have a binomial distribution:

$$\Pr(n_A, n_a) = \frac{(2n)!}{n_A!n_a!}(p_A)^{n_A}(p_a)^{n_a}$$

Exact HWE Test

Putting these together gives the conditional probability

$$\begin{aligned}
\Pr(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a, \text{HWE}) &= \frac{\Pr(n_{AA}, n_{Aa}, n_{aa}, n_A, n_a | \text{HWE})}{\Pr(n_A, n_a | \text{HWE})} \\
&= \frac{\frac{n!}{n_{AA}! n_{Aa}! n_{aa}!} (p_A^2)^{n_{AA}} (2 p_A p_a)^{n_{Aa}} (p_a^2)^{n_{aa}}}{\frac{(2n)!}{n_A! n_a!} (p_A)^{n_A} (p_a)^{n_a}} \\
&= \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!} \frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}
\end{aligned}$$

Reject the Hardy-Weinberg hypothesis if this quantity, the probability of the genotypic array conditional on the allelic array under HWE, is among the smallest of its possible values.

Exact HWE Test

For convenience, write the probability of the genotypic array, conditional on the allelic array and HWE, as $\Pr(n_{Aa}|n, n_A)$. Reject the HWE hypothesis for a data set if this value is among the smallest probabilities.

As an example, consider $(n_{AA} = 1, n_{Aa} = 0, n_{aa} = 49)$. The allele counts are $(n_A = 2, n_a = 98)$ and there are only two possible genotype arrays with those allele counts:

AA	Aa	aa	$\Pr(n_{Aa}\|n, n_A)$
1	0	49	$\frac{50!}{1!0!49!}\frac{2^0 2!98!}{100!} = \frac{1}{99}$
0	2	48	$\frac{50!}{0!2!48!}\frac{2^2 2!98!}{100!} = \frac{98}{99}$

Exact HWE Test

The probability of the data ($n_{AA} = 1, n_{Aa} = 0, n_{aa} = 49$), conditional on the allele frequencies and on HWE, is $1/99 = 0.01$ which is not nearly as small as the value suggested by the chi-square statistic of 50. (Because $\tilde{P}_{Aa} = 0, \hat{f} = 1, X^2 = n$.)

Traditionally, the p-value is the (conditional) probability of the data plus the probabilities of all the less-probable datasets. The probabilities are all calculated assuming HWE is true. More recently (Graffelman and Moreno, Statistical Applications in Genetics and Molecular Biology 2013; 12:433-448) it has been shown that the test has a significance value closer to the nominal value if the p-value is half the probability of the data plus the probabilities of all datasets that are less probable under the null hypothesis. For the previous slide then, the p-value is 1/198=0.005. A 5% exact significance level is not achievable.

Mid p-value

For this example:

AA	Aa	aa	$\Pr(n_{Aa}\|n, n_A)$	$p-$value	mid $p-$value
1	0	49	$\frac{50!}{1!0!49!}\frac{2^0 2!98!}{100!} = \frac{1}{99}$	0.0101	0.0051
0	2	48	$\frac{50!}{0!2!48!}\frac{2^2 2!98!}{100!} = \frac{98}{99}$	1.0000	0.5051

The average mid-p value is exactly 0.5, as for a uniform distribution on [0,1].

Mid-p values

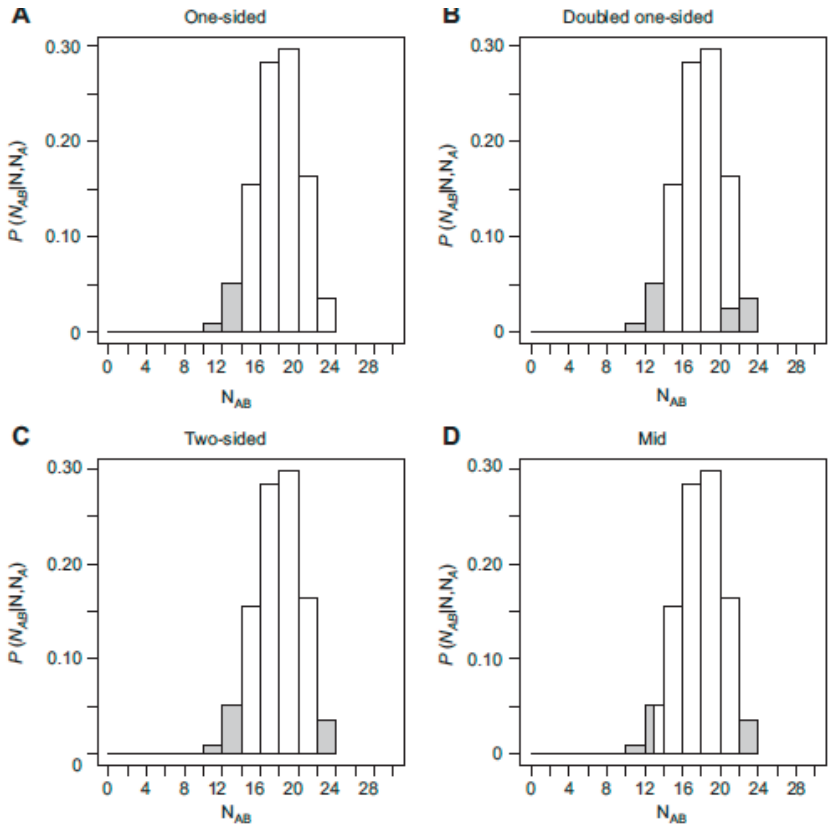


Figure 1 Computation of the *p*-value in an exact test for HWP, for a sample of 50 individuals with a minor allele count of 23, for which 13 heterozygotes were observed. (A) One-sided *p*-value in a test for heterozygote dearth. (B) *p*-value obtained by doubling the one-sided tail. (C) Standard two-sided *p*-value, (D) Mid *p*-value based on half the probability of the observed sample.

Example

For a sample of size $n = 100$ with minor allele count of 14, there are 8 sets of possible genotype counts:

			Exact		Chi-square	
n_{AA}	n_{Aa}	n_{aa}	Prob.	Mid p-value	X^2	p-value
93	0	7	0.0000	0.0000*	100.00	0.0000*
92	2	6	0.0000	0.0000*	71.64	0.0000*
91	4	5	0.0000	0.0000*	47.99	0.0000*
90	6	4	0.0002	0.0001*	29.07	0.0000*
89	8	3	0.0051	0.0028*	14.87	0.0001*
88	10	2	0.0602	0.0354*	5.38	0.0204*
87	12	1	0.3209	0.2260	0.61	0.4348
86	14	0	0.6136	0.6932	0.57	0.4503

So, for a nominal 5% significance level, the actual significance level is 0.0204 for a chi-square test that rejects when $n_{Aa} \leq 10$ and is 0.0354 for an exact test that also rejects when $n_{Aa} \leq 10$.

Effect of Minor Allele Frequency

The minor allele frequency (MAF) in the previous example was $14/200 = 0.07$. How does the exact test behave with other MAF values?

In particular, what is the size of the actual significance level when the nominal value is $\alpha = 0.05$? In other words, we decide to reject HWE for any sample with a p-value of 0.05 or less and choose the rejection region accordingly, what are the probabilities of rejecting? We would hope that the empirical significance level would be close to the nominal value, but we find that it may not be.

$n_a = 16$ minor alleles

When the minor allele frequency is 0.08, for a nominal 5% significance level, the actual significance level is 0.0070 for an exact test that rejects when $n_{Aa} \leq 10$.

n_{AA}	n_{Aa}	n_{aa}	Pr($n_{Aa}|n_a$)	Mid $p-$value
92	0	8	.0000	.0000
91	2	7	.0000	.0000
90	4	6	.0000	.0000
89	6	5	.0000	.0000
88	8	4	.0008	.0004
87	10	3	.0123	.0070
86	12	2	.0974	.0618
85	14	1	.3681	.2946
84	16	0	.5215	.7392

$n_a = 15$ minor alleles

When the minor allele frequency is 0.075, for a nominal 5% significance level, the actual significance level is 0.0474 for an exact test that rejects when $n_{Aa} \leq 11$.

n_{AA}	n_{Aa}	n_{aa}	Pr($n_{Aa}|n_a$)	Mid $p-$value
92	1	7	.0000	.0000
91	3	6	.0000	.0000
90	5	5	.0000	.0000
89	7	4	.0004	.0002
88	9	3	.0081	.0045
87	11	2	.0776	.0474
86	13	1	.3464	.2594
85	15	0	.5675	.7163

$n_a = 13$ minor alleles

When the minor allele frequency is 0.065, for a nominal 5% significance level, the actual significance level is 0.0483 for an exact test that rejects when $n_{Aa} \leq 9$.

n_{AA}	n_{Aa}	n_{aa}	Pr($n_{Aa}|n_a$)	Mid $p-$value
93	1	6	.0000	.0000
92	3	5	.0000	.0000
91	5	4	.0001	.0000
90	7	3	.0030	.0016
89	9	2	.0452	.0257
88	11	1	.2923	.1945
87	13	0	.6595	.6704

$n_a = 12$ minor alleles

When the minor allele frequency is 0.06, for a nominal 5% significance level, the actual significance level is 0.0344 for an exact test that rejects when $n_{Aa} \leq 8$.

n_{AA}	n_{Aa}	n_{aa}	Pr($n_{Aa}|n_a$)	Mid $p-$value
94	0	6	.0000	.0000
93	2	5	.0000	.0000
92	4	4	.0000	.0000
91	6	3	.0017	.0009
90	8	2	.0327	.0181
89	10	1	.2612	.1650
88	12	0	.7045	.6479

Graffelman and Moreno, 2013

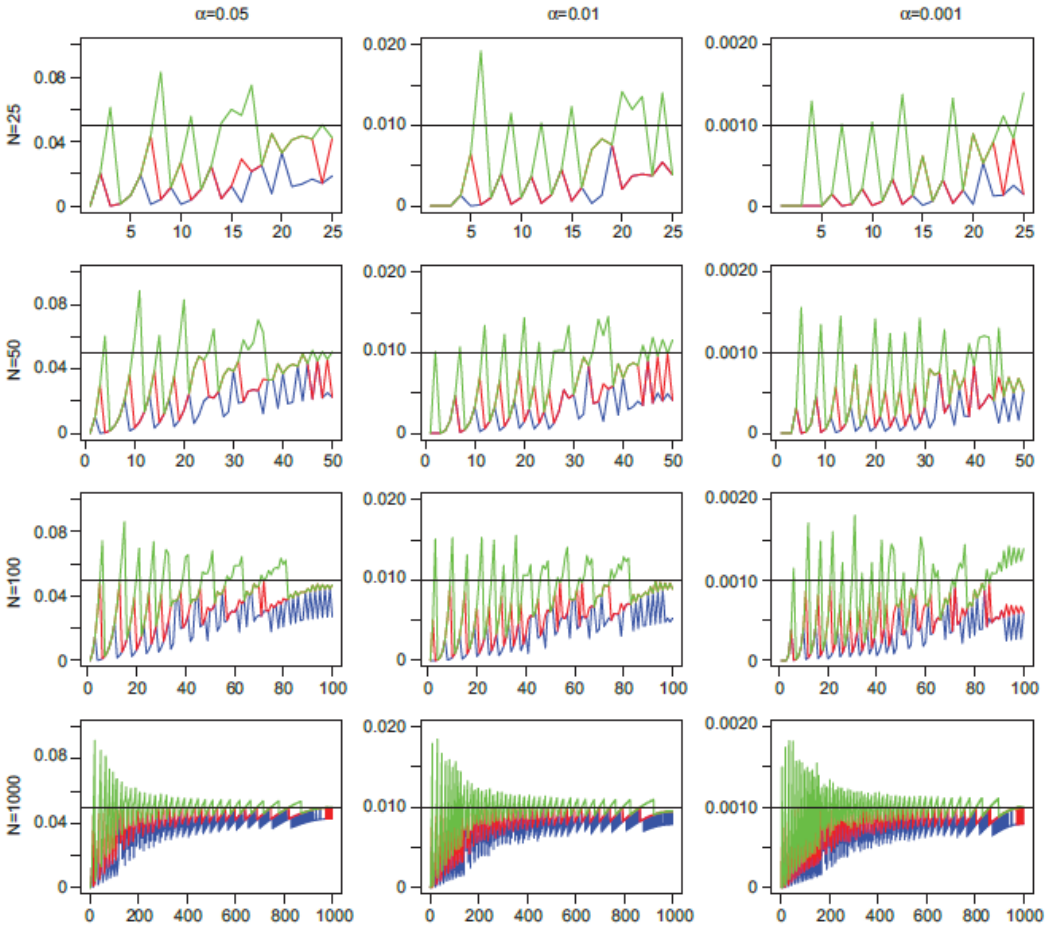


Figure 2 Type I error rate against minor allele count for different sample sizes (25, 50, 100 and 1000) and significance levels (0.05, 0.01, and 0.001) for exact tests with standard two-sided (red), doubled one-sided (blue) and mid *p*-values (green).

Power of Exact Test

If there is not HWE:

$$
\begin{aligned}
\Pr(n_{Aa}|n_A, n_a) &= \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(P_{AA})^{n_{AA}}(P_{Aa})^{n_{Aa}}(P_{aa})^{n_{aa}} \\
&= \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(P_{AA})^{\frac{n_A - n_{Aa}}{2}}(P_{Aa})^{n_{Aa}}(P_{aa})^{\frac{n_a - n_{Aa}}{2}} \\
&= \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(\sqrt{P_{AA}})^{n_A}(\sqrt{P_{aa}})^{n_a}\left(\frac{P_{Aa}}{\sqrt{P_{AA}P_{aa}}}\right)^{n_{Aa}} \\
&= \frac{C\psi^{n_{Aa}}}{n_{AA}!n_{Aa}!n_{aa}!}
\end{aligned}
$$

where $\psi = P_{Aa}/(\sqrt{P_{AA}P_{aa}})$ measures the departure from HWE. The constant C makes the probabilities sum to one over all possible n_{Aa} values: $C = 1/[\sum_{n_{Aa}} \psi^{n_{Aa}}/(n_{AA}!n_{Aa}!n_{aa}!)]$.

Power of Exact Test

Once the rejection region has been determined, the power of the test (the probability of rejecting) can be found by adding these probabilities for all sets of genotype counts in the region. HWE corresponds to $\psi = 2$. What is the power to detect HWE when $\psi = 1$, the sample size is $n = 10$ and the sample allele counts are $n_A = 15, n_a = 5$. Note that $C = 1/[1/(5!5!0!) + 1/(6!3!1!) + 1/(7!1!2!)]$.

			$\Pr(n_{Aa}\|n_A, n)$	
n_{AA}	n_{Aa}	n_{aa}	$\psi = 2$	$\psi = 1$
5	5	0	0.520	0.262
6	3	1	0.433	0.364
7	1	2	0.047	0.374

The $\psi = 2$ column (i.e. HWE) shows that the rejection region is $n_{Aa} = 1$. The $\psi = 1$ column (not HWE) shows that the power (the probability $n_{Aa} = 1$ when $\psi = 1$) is 37.4%.

Power when $n_a = 16$

The rejection region of $n_{Aa} \leq 10$ is determined from null hypothesis, and the power is determined from the multinomial distribution. Rejection is unlikely if $f < 0$.

		$\Pr(n_{Aa} \mid n_a = 16, n = 100)$						
	ψ	.250	.500	1.000	2.000	4.000	8.000	16.000
n_{Aa}	f	.631	.398	.157	.000	$-.062$	$-.081$	$-.085$
0		.0042	.0000	.0000	.0000	.0000	.0000	.0000
2		.0956	.0026	.0000	.0000	.0000	.0000	.0000
4		.3172	.0349	.0003	.0000	.0000	.0000	.0000
6		.3568	.1569	.0056	.0000	.0000	.0000	.0000
8		.1772	.3116	.0441	.0008	.0000	.0000	.0000
10		.0433	.3047	.1725	.0123	.0003	.0000	.0000
12		.0054	.1506	.3411	.0974	.0098	.0007	.0000
14		.0003	.0356	.3223	.3681	.1485	.0422	.0109
16		.0000	.0032	.1142	.5214	.8414	.9571	.9890
	Power	.9943	.8107	.2225	.0131	.0003	.0000	.0000

Graffelman and Moreno, 2013

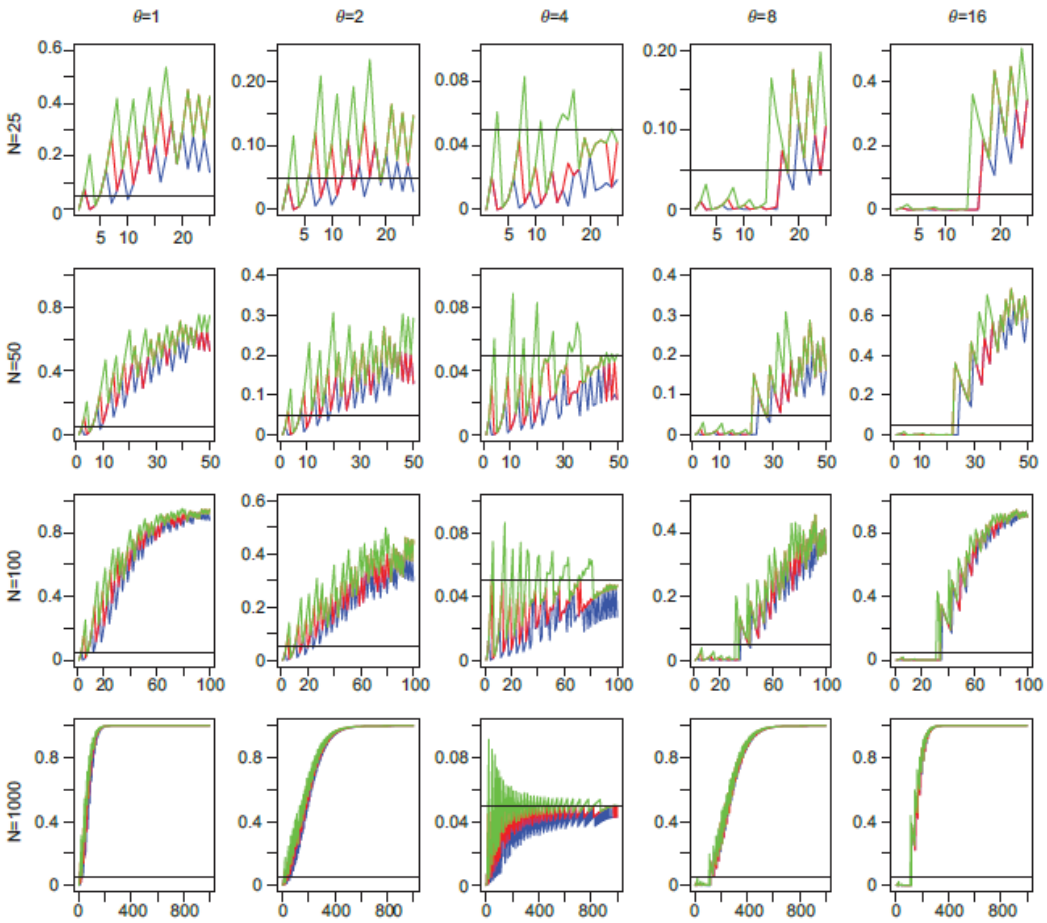


Figure 3 Power of HWP exact tests against minor allele count for different sample sizes (25, 50, 100 and 1000) and degree of disequilibrium (1, 2, 4, 8 and 16). Standard two-sided (red), doubled one-sided (blue) and mid *p*-values (green).

θ in this plot is ψ^2.

Multiple Testing

When multiple tests are performed, each at significance level α, a proportion α of the tests are expected to cause rejection even if all the hypotheses are true.

Bonferroni correction makes the overall (experimentwise) significance level equal to α by adjusting the level for each individual test to α'. If α is the probability that at least one of the L tests causes rejection, it is also 1 minus the probability that none of the tests causes rejection:

$$\begin{aligned} \alpha &= 1-(1-\alpha')^L \\ &\approx L\alpha' \end{aligned}$$

provided the L tests are independent.

If $L = 15$, need $\alpha' = 0.0033$ in order for $\alpha = 0.05$.

Combining p-values

There is also the issue that if the same hypothesis is tested L times and just fails to cause rejection each time, there is some overall evidence against the hypothesis.

Suppose that tests have been conducted for each of L hypotheses $H_i, i = 1, 2, \ldots, L$. For each test the p-value p_i is calculated: if H_i is true, this is the probability of observing a test statistic as extreme as or more extreme than the observed value in the direction of rejection.

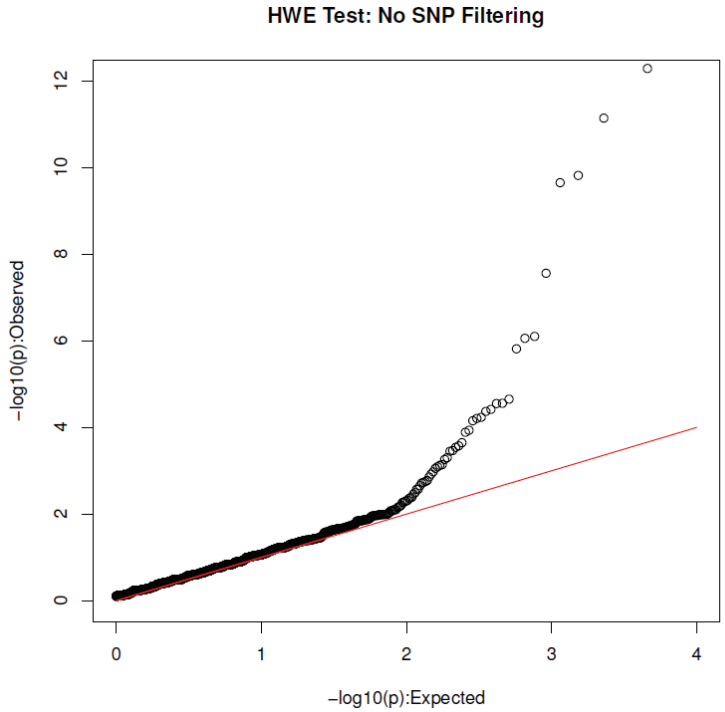
Methods for combining p-values rest on p having a uniform distribution when the hypothesis is true.

QQ-Plots

A convenient approach to considering multiple-testing issues is to use QQ-plots. If all the hypotheses being tested are true then the resulting p-values are uniformly distributed between 0 and 1.

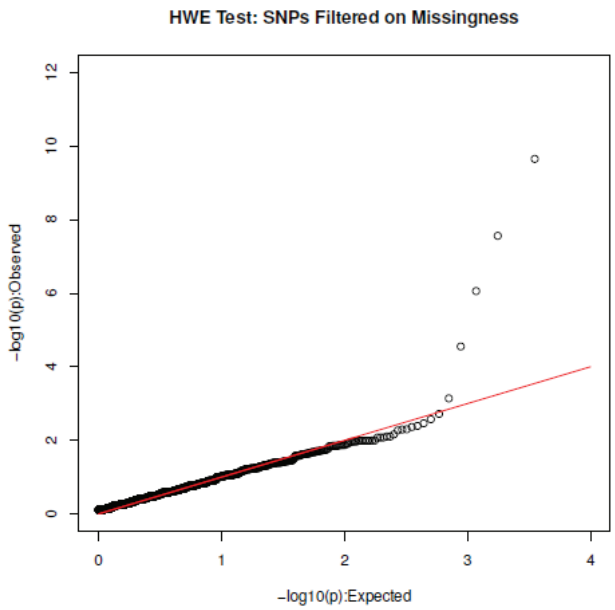
For a set of n tests, we would expect to see p values at $1/(n+1), 2/(n+1), \ldots, n/(n+1)$. We plot the observed p-values against these expected values: the smallest against $1/(n+1)$ and the largest against $n/(n+1)$. It is more convenient to transform to $-\log_{10}(p)$ to accentuate the extremely small p values. The point at which the observed values start departing from the expected values is an indication of "significant" values in a way that takes into account the number of tests.

QQ-Plots



The results for 9208 SNPs on human chromosome 1 for the AMD controls. Bonferroni would suggest rejecting HWE when $p \leq 0.05/9205 = 5.4 \times 10^{-6}$ or $-\log_{10}(p) \geq 5.3$.

QQ-Plots



The same set of results as on the previous slide except now that any SNP with any missing data was excluded. Now 7446 SNPs and Bonferroni would reject if $-\log_{10}(p) \geq 5.2$. All five outliers had zero counts for the minor allele homozygote and at least 32 heterozygotes in a sample of size 50.

HWE Testing for X-linked Markers

Usual procedure is to ignore males and test for HWE with male data only.

If there is HWE, the allele frequencies at X-linked markers are the same in males and females.

Different allele frequencies for males and females suggest a departure from HWE.

HWE Testing for X-linked Markers

Sample sizes: n_m males, n_f females, $n = n_m + n_f$.

Male allele counts: m_A, m_B.

Female genotype counts: f_{AA}, f_{AB}, f_{BB}.

Total sample allele counts: n_A, n_B.

Probability of data, under HWE for females and equal male and female allele frequencies:

$$\Pr(m_A, f_{AB} | n, n_A, n_m) = \frac{n_A! n_B! n_m! n_f! 2^{f_{AB}}}{m_A! m_B! f_{AA}! f_{AB}! f_{BB}!}$$

An exact test for the joint hypotheses of female HWE and equal male and female allele frequencies was constructed from this probability, as described in version 1.5.6 of the *HardyWeinberg* package.

Graffelman and Weir, 2016

Table 2 All possible samples for a set of 20 individuals (10 males and 10 females) with a total of 6 A alleles

	m_A	m_B	f_{AA}	f_{AB}	f_{BB}	*Prob*
1	0	10	3	0	7	0.0002
2	0	10	2	2	6	0.0085
3	0	10	1	4	5	0.0340
4	0	10	0	6	4	0.0226
5	1	9	2	1	7	0.0121
6	1	9	1	3	6	0.1132
7	1	9	0	5	5	0.1358
8	2	8	2	0	8	0.0034
9	2	8	1	2	7	0.1091
10	2	8	0	4	6	0.2546
11	3	7	1	1	8	0.0364
12	3	7	0	3	7	0.1940
13	4	6	1	0	9	0.0035
14	4	6	0	2	8	0.0637
15	5	5	0	1	9	0.0085
16	6	4	0	0	10	0.0004

KING-robust Kinship Estimates

The influence of all members of a sample is eliminated by the KING-robust* estimates for pairs of individuals:

$$\widehat{\theta}^{K}_{jj'} \;=\; \frac{N(AB,AB)_{jj'} - 2N(AA,BB)_{jj'}}{N(AB)_{j} + N(AB)_{j'}}$$

where the N's are the numbers of loci with the indicated genotypes for individuals j and individual pairs j, j'. The expected values of these estimates are

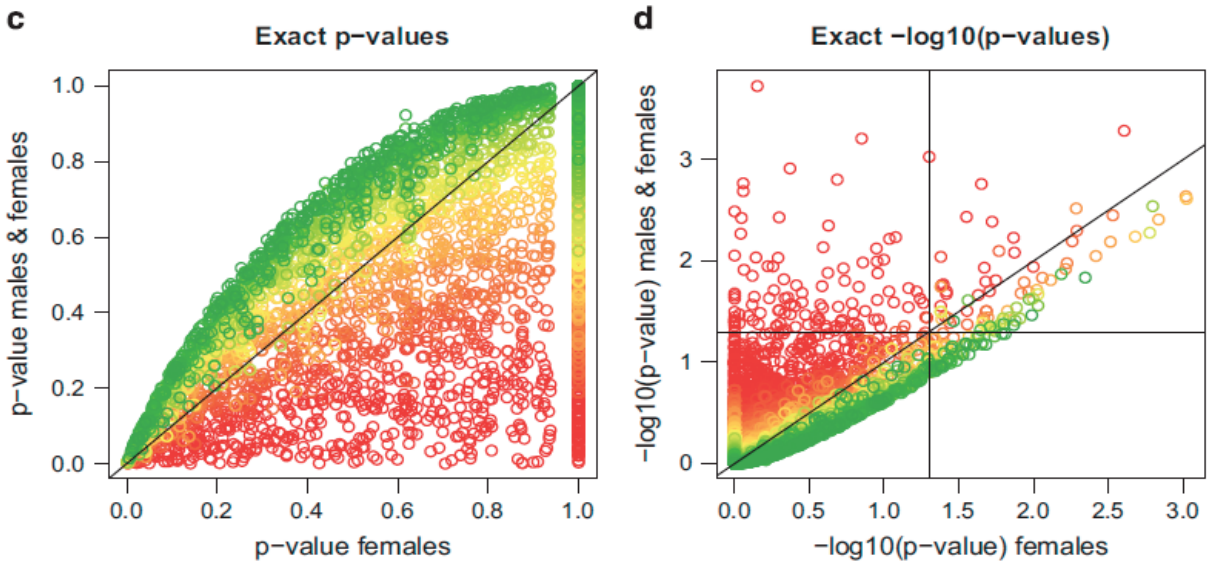
$$\mathcal{E}(\widehat{\theta}^{K}_{jj'}) \;=\; \frac{\theta_{jj'} - \frac{1}{2}(F_j + F_{j'})}{1 - \frac{1}{2}(F_j + F_{j'})}$$

These estimates for a pair of individuals are relative to the average inbreeding coefficients of those individuals.

[* Manichaikul et al., Bioinformatics 26:2867-2873, 2010.]

Heredity 116:558–568.

Graffelman and Weir, 2016



Scatter plots of P-values in original and -log10 scale for exact test for HWE using females only and using both males and females for 4158 SNPs at the X chromosome of the venous thrombosis database. The horizontal and vertical black lines correspond to a significance level of 5%. Points colored according to their significance level in Fisher's test for equality of allele frequencies (range 0-1 from red to green).

Linkage Disequilibrium

This term reserved for association between pairs of alleles – one at each of two loci.

When gametic data are available, could refer to gametic disequilibrium.

When genotypic data are available, but gametes can be inferred, can make inferences about gametic and non-gametic pairs of alleles.

When genotypic data are available, but gametes cannot be inferred, can work with composite measures of disequilibrium.

Linkage Disequilibrium

For alleles A and B are two loci, the usual measure of linkage disequilibrium is

$$D_{AB} = P_{AB} - p_A p_B$$

Whether or not this is zero does not provide a direct statement about linkage between the two loci. For example, consider marker YFM and disease DTD:

		A	N	Total
YFM	+	1	24	25
	−	0	75	75
Total		1	99	100

$$D_{A+} = \frac{1}{100} - \frac{1}{100}\frac{25}{100} = 0.0075, \text{ (maximum possible value)}$$

Gametic Linkage Disequilibrium

For loci **A, B** define indicator variables x, y that take the value 1 for allele A, B and 0 for any other alleles. If gametes within individuals are indexed by j, $j = 1, 2$ then for expectations over samples from the same population

$$\begin{aligned}
\mathcal{E}(x_j) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_j^2) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j^2) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_1 x_2) = P_{AA} \;\; &, \;\; \mathcal{E}(y_1 y_2) = P_{BB} \\
\mathcal{E}(x_1 y_1) = P_{AB} \;\; &, \;\; \mathcal{E}(x_2 y_2) = P_{AB}
\end{aligned}$$

The variances of x_j, y_j are $p_A(1-p_A), p_B(1-p_B)$ for $j = 1, 2$ and the covariance and correlation coefficients for x and y are

$$\begin{aligned}
\mathrm{Cov}(x_1, y_1) = \mathrm{Cov}(x_2, y_2) &= P_{AB} - p_A p_B = D_{AB} \\
\mathrm{Corr}(x_1, y_1) = \mathrm{Corr}(x_2, y_2) &= D_{AB} / \sqrt{[p_A(1-p_A) p_B(1-p_B)]} = \rho_{AB}
\end{aligned}$$

Estimation of LD

With random sampling of gametes, gamete counts have a multinomial distribution:

$$\begin{aligned}\Pr(n_{AB}, n_{Ab}, n_{aB}, n_{ab}) &= \frac{n!(P_{AB})^{n_{AB}}(P_{Ab})^{n_{Ab}}(P_{aB})^{n_{aB}}(P_{ab})^{n_{ab}}}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!} \\ &= \frac{n!(p_Ap_B + D_{AB})^{n_{AB}}(p_Ap_b - D_{AB})^{n_{Ab}}}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!} \\ &\quad \times (p_ap_B - D_{AB})^{n_{aB}}(p_ap_b + D_{AB})^{n_{ab}}\end{aligned}$$

and this provides the maximum likelihood estimates of D_{AB} and ρ_{AB}:

$$\begin{aligned}\hat{D}_{AB} &= \frac{n_{AB}}{n} - \frac{n_{AB} + n_{Ab}}{n} \times \frac{n_{AB} + n_{aB}}{n} = \tilde{P}_{AB} - \tilde{p}_A\tilde{p}_B \\ \hat{\rho}_{AB} = r_{AB} &= \frac{\hat{D}_{AB}}{\sqrt{\tilde{p}_A\tilde{p}_a\tilde{p}_B\tilde{p}_b}}\end{aligned}$$

Testing LD

Write MLE of D_{AB} as

$$\hat{D}_{AB} = \frac{n_{AB}n_{ab} - n_{Ab}n_{aB}}{(n_{AB}+n_{Ab})(n_{aB}+n_{ab})(n_{AB}+n_{aB})(n_{Ab}+n_{ab})}$$

and use "Delta method" to find

$$\begin{aligned}\text{Var}(\hat{D}_{AB}) &\approx \frac{1}{n}[p_A(1-p_A)p_B(1-p_B) \\ &\quad + (1-2p_A)(1-2p_B)D_{AB} - D_{AB}^2]\end{aligned}$$

When $D_{AB} = 0$, $\text{Var}(\hat{D}_{AB}) = p_A(1-p_A)p_B(1-p_B)/n$.

If $\hat{D}_{AB}$ is assumed to be normally distributed then

$$X_{AB}^2 = \frac{\hat{D}_{AB}^2}{\text{Var}(\hat{D}_{AB})} = n\hat{\rho}_{AB}^2 = nr_{AB}^2$$

is appropriate for testing $H_0 : D_{AB} = 0$. When H_0 is true, $X_{AB}^2 \sim \chi_{(1)}^2$. Note the analogy to the test statistic for Hardy-Weinberg equilibrium: $X^2 = n\hat{f}^2$.

Goodness-of-fit Test

The test statistic for the 2×2 table

n_{AB}	n_{Ab}	n_A
n_{aB}	n_{ab}	n_a
n_B	n_b	n

has the value

$$X^2 = \frac{n(n_{AB}n_{ab} - n_{Ab}n_{aB})^2}{n_A n_a n_B n_b}$$

For DTD/YFM example, $X^2 = 3.03$. This is not statistically significant, even though disequilibrium was maximal.

Composite Disequilibrium

When genotypes are scored, it is often not possible to distinguish between the two double heterozygotes AB/ab and Ab/aB, so that gametic frequencies cannot be inferred.

Under the assumption of random mating, in which genotypic frequencies are assumed to be the products of gametic frequencies, it is possible to estimate gametic frequencies with the EM algorithm. To avoid making the random-mating assumption, however, it is possible to work with a set of composite disequilibrium coefficients.

Composite Disequilibrium

Although the separate digenic frequencies p_{AB} (one gamete) and $p_{A,B}$ (two gametes) cannot be observed, their sum can be since

$$\begin{aligned} p_{AB} &= P_{AB}^{AB} + \frac{1}{2}P_{Ab}^{AB} + \frac{1}{2}P_{aB}^{AB} + \frac{1}{2}P_{ab}^{AB} \\ p_{A,B} &= P_{AB}^{AB} + \frac{1}{2}P_{Ab}^{AB} + \frac{1}{2}P_{aB}^{AB} + \frac{1}{2}P_{aB}^{Ab} \\ p_{AB} + p_{A,B} &= 2P_{AB}^{AB} + P_{Ab}^{AB} + P_{aB}^{AB} + \frac{P_{ab}^{AB} + P_{aB}^{Ab}}{2} \end{aligned}$$

Digenic disequilibrium is measured with a composite measure Δ_{AB} defined as

$$\begin{aligned} \Delta_{AB} &= p_{AB} + p_{A,B} - 2p_A p_B \\ &= D_{AB} + D_{A,B} \end{aligned}$$

which is the sum of the gametic ($D_{AB} = p_{AB} - p_A p_B$) and nongametic ($D_{A,B} = p_{A,B} - p_A p_B$) coefficients.

Composite Disequilibrium

If the counts of the nine genotypic classes are

	BB	Bb	bb
AA	n_1	n_2	n_3
Aa	n_4	n_5	n_6
aa	n_7	n_8	n_9

the count for pairs of alleles in an individual being A and B, whether received from the same or different parents, is

$$n_{AB} = 2n_1 + n_2 + n_4 + \frac{1}{2}n_5$$

and the MLE for Δ is

$$\hat{\Delta}_{AB} = \frac{1}{n}n_{AB} - 2\tilde{p}_A\tilde{p}_B$$

Composite Linkage Disequilibrium

For loci **A, B** define indicator variables x, y that take the value 1 for allele A, B and 0 for any other alleles. If gametes within individuals are indexed by j, $j = 1, 2$ then for expectations over samples from the same population

$$\begin{aligned}
\mathcal{E}(x_j) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_j^2) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_1 x_2) = P_{AA} \;\; &, \;\; \mathcal{E}(y_1 y_2) = P_{BB} \\
\mathcal{E}(x_1 y_1) = P_{AB} \;\; &, \;\; \mathcal{E}(x_2 y_2) = P_{AB} \\
\mathcal{E}(x_1 y_2) = P_{A,B} \;\; &, \;\; \mathcal{E}(x_2 y_1) = P_{A,B}
\end{aligned}$$

Write

$$\begin{aligned}
D_A = P_{AA} - p_A^2 \;\; &, \;\; D_B = P_{BB} - p_B^2 \\
D_{AB} = P_{AB} - p_A p_B \;\; &, \;\; D_{A,B} = P_{A,B} - p_A p_B \\
\Delta_{AB} = D_{AB} + D_{A,B} &
\end{aligned}$$

Composite Linkage Disequilibrium

Now set $X = x_1 + x_2, Y = y_1 + y_2$ to get

$$\begin{aligned}
\mathcal{E}(X) = 2p_A \quad &, \quad \mathcal{E}(Y) = 2p_B \\
\mathcal{E}(X^2) = 2(p_A + P_{AA}) \quad &, \quad \mathcal{E}(Y^2) = 2(p_B + P_{BB}) \\
\mathrm{Var}(X) = 2p_A(1-p_A)(1+f_A) \quad &, \quad \mathrm{Var}(Y) = 2p_B(1-p_B)(1+f_B)
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{E}(XY) &= 2(P_{AB} + P_{A,B}) \\
\mathrm{Cov}(X,Y) &= 2(P_{AB} - p_A p_B) + 2(P_{A,B} - p_A p_B) \\
&= 2(D_{AB} + D_{A,B}) = 2\Delta_{AB} \\
\mathrm{Corr}(X,Y) &= \frac{\Delta_{AB}}{\sqrt{p_A(1-p_A)(1+f_A)p_B(1-p_B)(1+f_B)}}
\end{aligned}$$

Composite Linkage Disequilibrium

$$\hat{\Delta}_{AB} = n_{AB}/n - 2\tilde{p}_A\tilde{p}_B$$

where

$$n_{AB} = 2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{1}{2}n_{AaBb}$$

This does not require phased data.

By analogy to the gametic linkage disequilibrium result, a test statistic for $\Delta_{AB} = 0$ is

$$X^2_{AB} = \frac{n\hat{\Delta}^2_{AB}}{\tilde{p}_A(1-\tilde{p}_A)(1+\hat{f}_A)\tilde{p}_B(1-\tilde{p}_B)(1+\hat{f}_B)}$$

This is assumed to be approximately $\chi^2_{(1)}$ under the null hypothesis.

Example

For the data on slide 74:

	BB	Bb	bb	Total
AA	$n_{AABB} = 5$	$n_{AABb} = 3$	$n_{AAbb} = 2$	$n_{AA} = 10$
Aa	$n_{AaBB} = 3$	$n_{AaBb} = 2$	$n_{Aabb} = 0$	$n_{Aa} = 5$
aa	$n_{aaBB} = 0$	$n_{aaBb} = 0$	$n_{aabb} = 0$	$n_{aa} = 0$
Total	$n_{BB} = 8$	$n_{Bb} = 5$	$n_{bb} = 2$	$n = 15$

$$
\begin{aligned}
n_{AB} &= 2 \times 5 + 3 + 3 + \frac{1}{2}(2) = 17 \\
n_A &= 25, \;\; \tilde{p}_A = 5/6 \\
n_B &= 21, \;\; \tilde{p}_B = 7/10
\end{aligned}
$$

Example

The estimated composite disequilibrium coefficient is

$$\hat{\Delta}_{AB} = \frac{17}{15} - 2\frac{25}{30}\frac{21}{30} = -\frac{1}{30} = -0.033$$

Previous work on EM algorithm estimated p_{AB} as 16/30 so

$$\hat{D}_{AB} = \frac{16}{30} - \frac{25}{30}\frac{21}{30} = -\frac{1}{20} = -0.050$$

LD vs Composite LD Estimates

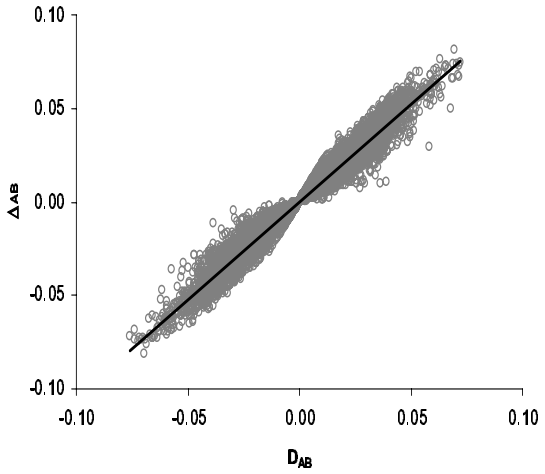


Figure 6: Linkage disequilibrium estimates

A comparison of gametic LD estimates from the EM algorithm assuming HWE vs composite LD with no HWE assumption.

Reality Check: HWE for AMD Data

SNP		AA	AB	BB
rs10492941				
	Case	0	17	79
	Control	1	5	44
rs380390				
	Case	50	35	11
	Control	6	25	19

HWE for AMD Data

$$\widehat{f} \;=\; 1 - \frac{\tilde{P}_{AB}}{2\tilde{p}_A\tilde{p}_B}$$

rs10492941 Control

$n =$
$n_A = (2n_{AA} + n_{AB})/(2n) =$
$n_B = (2n_{BB} + n_{AB})/(2n) =$
$\widehat{f} = 1 - (2nn_{AB})/(n_A n_B) =$
$X^2 = n\widehat{f}^2 =$

rs380390 Conrol

$n =$
$n_A = (2n_{AA} + n_{AB})/(2n) =$
$n_B = (2n_{BB} + n_{AB})/(2n) =$
$\widehat{f} = 1 - (2nn_{AB})/(n_A n_B) =$
$X^2 = n\widehat{f}^2 =$

POPULATION STRUCTURE

Population Data

Individuals from several populations are scored at a series of marker loci. At each locus, an individual has two alleles, one from each parent, and these can be identified. For example, at locus D3S1358:

Allele	AFC	NSC	QLC	SAC	TAC	VIA	WAB
11	.000	.001	.002	.001	.000	.000	.000
12	.004	.003	.001	.001	.000	.000	.010
13	.008	.003	.002	.002	.000	.000	.001
14	.123	.098	.159	.125	.152	.008	.075
15	.261	.264	.365	.252	.244	.385	.353
16	.250	.270	.250	.265	.241	.277	.242
17	.187	.198	.123	.202	.197	.246	.190
18	.154	.152	.091	.144	.157	.077	.122
19	.012	.011	.006	.007	.010	.008	.007
20	.002	.000	.000	.000	.000	.000	.000

Questions of Interest

- How much genetic variation is there? (animal conservation)
- How much migration (gene flow) is there between populations? (molecular ecology)
- How does the genetic structure of populations affect tests for linkage between genetic markers and human disease genes? (human genetics)
- How should the evidence of matching marker profiles be quantified? (forensic science)
- What is the evolutionary history of the populations sampled? (evolutionary genetics)

Statistical Analysis

Possible to approach these data from purely statistical viewpoint.

Could test for differences in allele frequencies among populations.

Could use various multivariate techniques to cluster populations.

These analyses may not answer the biological questions.

The Genetic Problem

- How do we describe the genetic similarity between populations or between individuals?

- Do our methods need changing now that we have detailed sequence data?

- What do we do with estimated genetic similarities?

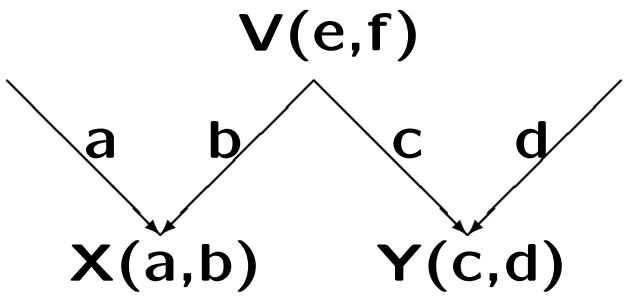
Identity by descent

We describe relatedness and population structure with the concept of identity by descent (ibd): two alleles at a locus are ibd if they have both descended from the same allele at some time in the past.

We write θ for the probability of two alleles being ibd.

Half siblings

For example, for half siblings X, Y with a common parent V:



The alleles b, c received by X, Y have a 50% chance of both being a copy of an allele, e or f, carried by parent V and so being ibd.

Coancestry of Half sibs

The coancestry coefficient of two individuals is the probability a random allele from one is ibd to a random allele from the other.

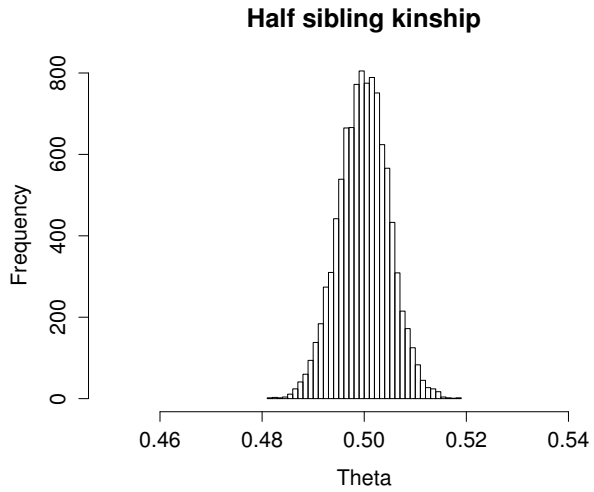
A random allele from X is a or b, each with probability of 0.5, and random allele from Y is c or d, each with probability of 0.5. Only one pair, b, c can be ibd and that event also has probability 0.5. Setting these out in a 2×2 table:

			Y	
			0.5	0.5
		Pr(ibd)	c	d
X	0.5	a	0	0
	0.5	b	0.5	0

The coancestry of X, Y is $\theta_{XY} = 0.5 \times 0.5 \times 0.5 = 0.125$.

Actual vs Predicted Coancestry

For any particular gene, the two alleles received by half siblings from their common parent either are, or are not, ibd. The "actual" kinship $\ddot{\theta}$ is either 1 or 0, even though the predicted value is $\theta = 0.5$. This plot shows the actual kinship coefficients, averaged over 10,000 loci:



Path Counting

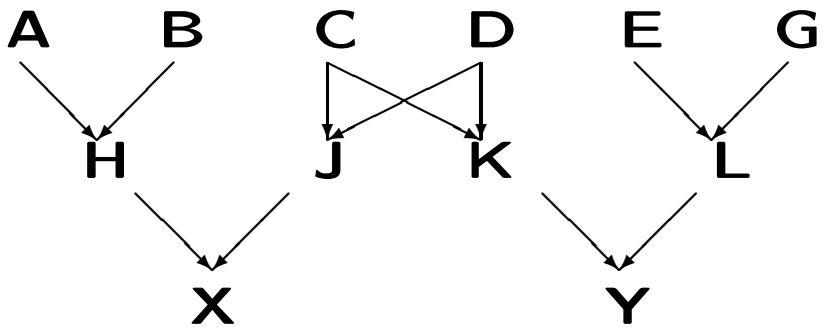
The inbreeding coefficient F of an individual is the probability that its two alleles are ibd: i.e. that an allele chosen randomly from one parent is ibd to an allele chosen randomly from the other parent.

If the parents X, Y of an individual I have ancestor A in common, and if there are n individuals (including X, Y, I) in the path linking the parents through A, then the inbreeding coefficient of I, or the coancestry of X and Y, is

$$F_I = \theta_{XY} = \left(\frac{1}{2}\right)^n (1 + F_A)$$

If there are several ancestors, this expression is summed over all the ancestors.

First cousins



The common ancestors of cousins X and Y are C and D. The paths linking X, Y to their common ancestors are $XJCKY$ and $XJDKY$ and these each have $n = 5$ individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^5 = \frac{1}{16}$$

Actual kinships will vary around this expected value.

"Relative to"

"There is no absolute measure of ibd: ibd is always relative to some reference population." [Thompson, Genetics, 2013.]

Imagine a large group of first cousins, who pair randomly and have children. Are the children inbred?

Within that group, the parents are not related and because they pair randomly the children are not inbred.

From the perspective of an observer from outside the group, however, the parents are related with kinships of $\theta = 1/16$ and their children are inbred with $F = 1/16$. These values are relative to those in the rest of the population.

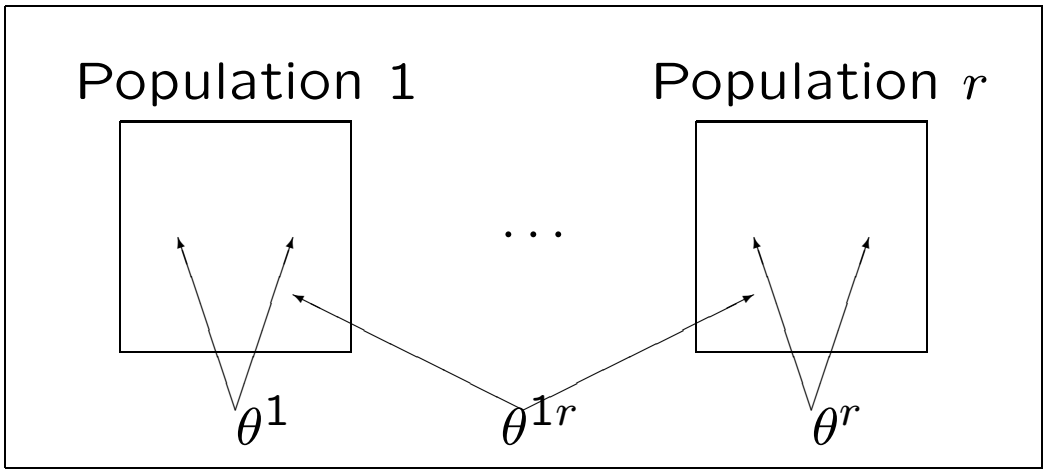
Population Structure

The coancestry coefficient for two individuals is the probability that two alleles, one taken randomly from each individual, are ibd.

We can extend this idea to the coancestry for a population: the probability that two alleles taken randomly from the population are ibd. If the population is in Hardy-Weinberg equilibrium, we can draw alleles randomly without regard to which individuals in which they are carried.

A further extension is when two alleles are drawn randomly, one from one population and one from another population.

Several Populations



θ's are statements about pairs of alleles: the probabilities the pairs are identical by descent.

θ^W is the average of the within-population coancestries θ^i.

θ^B is the average of the population-pair coancestries $\theta^{ii'}, i \neq i'$.

Population Structure

The usual measure of genetic population structure is written as F_{ST}. We can define this quantity for a set of populations as

$$F_{ST} = \frac{\theta^W - \theta^B}{1 - \theta^B}$$

and we might prefer to write it as β^W to avoid confusion with the usual definitions that refer to allele frequencies.

We also have such a quantity for each population:

$$\beta^i = \frac{\theta^i - \theta^B}{1 - \theta^B}$$

Genetic Drift

If a population has a constant size N diploid individuals, and there are no evolutionary forces such as mutation or migration, then with completely random mating two alleles in a generation have a probability $1/2N$ of coming from the same allele in the previous generation and so are ibd. With probability $(1-1/2N)$ they come from different alleles in the previous generation and are ibd with probability equal to the coancestry in that generation:

$$\theta(t+1) \;=\; \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\theta(t)$$

If the founding population has coancestry $\theta(0)$, then t generations later

$$\theta(t) \;=\; 1 - [1 - \theta(0)]\left(\frac{2N-1}{2N}\right)^t$$

and this tends to 1 as t becomes large.

Genetic Drift and Mutation

If the loss of genetic variation is opposed by mutation introducing variation, then θ has a value between 0 and 1. If every mutation gives a new allelic type ("infinite alleles mutation") then the transition equation is changed to

$$\theta(t+1) \;=\; (1-\mu)^2\left[\frac{1}{2N}+\left(1-\frac{1}{2N}\right)\theta(t)\right]$$

and the equilibrium coancestry is

$$\theta \;=\; \frac{1}{1+4N\mu}$$

Genetic Drift, Mutation and Migration

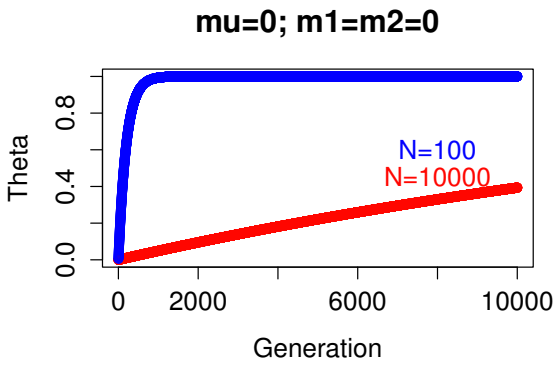
Now suppose there are two populations, each undergoing drift and mutation but with migration between them. If a fraction m_i of the alleles forming a generation in population i come from the other population in the previous generation, then

$$\begin{aligned}
\theta^1(t+1) &= (1-\mu)^2\left[(1-m_1)^2\phi^1(t) + 2m_1(1-m_1)\theta^{12}(t) + m_1^2\phi^2(t)\right] \\
\theta^2(t+1) &= (1-\mu)^2\left[(1-m_2)^2\phi^2(t) + 2m_2(1-m_2)\theta^{12}(t) + m_2^2\phi^1(t)\right] \\
\theta^{12}(t+1) &= (1-\mu)^2\left[(1-m_1)m_2\phi^1(t) + [(1-m_1)(1-m_2)+m_1m_2]\theta^{12}(t)\right. \\
&\quad \left. + m_1(1-m_2)\phi^2(t)\right]
\end{aligned}$$

where $\phi^i(t) = 1/(2N_i) + (2N_i - 1)\theta^i(t)/(2N_i),\ i = 1, 2.$

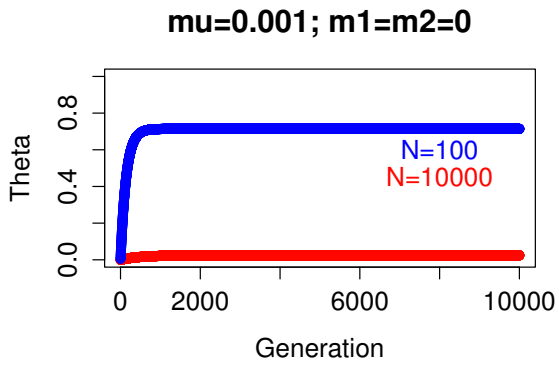
A consequence of these equations is that $\theta^1(t) + \theta^2(t) \geq 2\theta^{12}(t)$, or that $\theta^W \geq \theta^B$ and so $\beta^W = F_{ST}$ is positive. However, it is not necessary that each of θ^1, θ^2 exceeds θ^{12}.

Drift Only



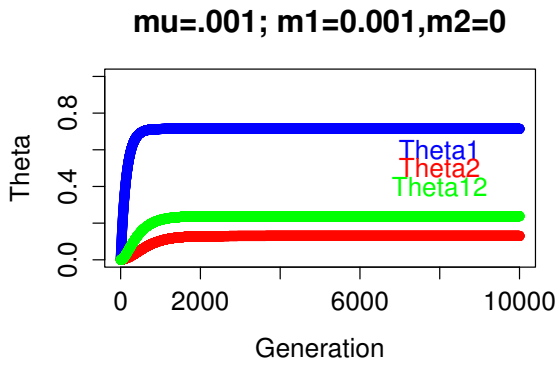
Drift Only

$\beta_1, \beta_2 > 0$



Drift and Mutation

$\beta_1, \beta_2 > 0$



Drift, Mutation and Migration

$\beta_1 > 0, \beta_2 < 0$

Allele Frequencies

If a sample of n_i alleles from population i has n_{iu} copies of allele u, sample allele frequencies are $\tilde{p}_{iu} = n_{iu}/n_i$ and

$$n_{iu} \sim \text{Binomial}\,(n_i, \ddot{p}_{iu})$$

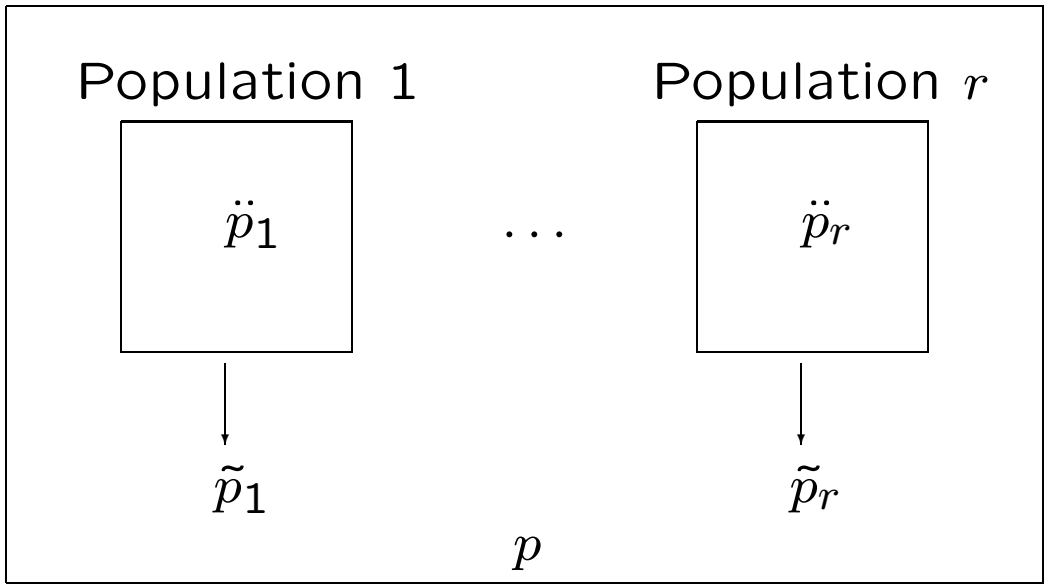
The binomial distribution describes variation among replicate samples from the same population. "Statistical sampling."

A wide class of evolutionary models leads to the beta distribution describing variation of actual allele frequencies in a population among replicates of the evolutionary process from a founding population:

$$\ddot{p}_{iu} \sim \text{Beta}\left(\frac{(1-\theta_i)p_u}{\theta_i}, \frac{(1-\theta_i)(1-p_u)}{\theta_i}\right)$$

"Genetic sampling."

Several Populations



Total Mean and Variance

Taking expectations over replicate samples from a population and over replicates of the population:

$$\begin{aligned} \mathcal{E}(\tilde{p}_i) &= p \\ \mathrm{Var}(\tilde{p}_i) &= p(1-p)\left(\theta_i + \frac{1-\theta_i}{n_i}\right) \end{aligned}$$

Allelic Matching

The θ's refer to identity by descent (ibd), whereas what we see is identity in state (ibs).

We work with the proportions $\tilde{M}$ of pairs of alleles, within or between individuals, or within or between populations, that match (i.e. are ibs).

We have a genetic model that says the expected value of an $\tilde{M}$ is $\theta + (1 - \theta)H$ where θ refers to the same set of alleles as does $\tilde{M}$ and H is the sum of squares of allele frequencies in the whole population from which the observations are drawn.

Matching Proportions for Individuals

Two distinct alleles from individual j in population i
$\tilde{M}^i_j = \frac{1}{2}\sum_u X^i_{ju}(X^i_{ju}-1),\ \ \mathcal{E}(\tilde{M}^i_j) = H + (1-H)F^i_j$

Average over individuals in population i
$\tilde{M}^i_W = \frac{1}{n_i}\sum_{j=1}^{n_i} \tilde{M}^i_j,\ \ \mathcal{E}(\tilde{M}^i_W) = H + (1-H)F^i_W$

One allele from each of individuals j, j' in population $i^\dagger$
$\tilde{M}^i_{jj'} = \frac{1}{4}\sum_u X^i_{ju}X^i_{j'u},\ \ \mathcal{E}(\tilde{M}^i_{jj'}) = H + (1-H)\theta^i_{jj'}$

Average over pairs of individuals in population i
$\tilde{M}^i_B = \frac{1}{n_i(n_i-1)}\sum_{j=1}^{n_i}\sum_{j'=1,j'\neq j}^{n_i} \tilde{M}^i_{jj'},\ \ \mathcal{E}(\tilde{M}^i_B) = H + (1-H)\theta^i_B$

Matching Proportions for Populations

Two distinct alleles from population i

$\tilde{M}^i = \frac{1}{2n_i(2n_i-1)}\left(2\sum_{j=1}^{n_i}\tilde{M}^i_j + 4\sum_{j=1}^{n_i}\sum_{j'=1,j'\neq j}^{n_i}\tilde{M}^i_{jj'}\right)$

$\mathcal{E}(\tilde{M}^i) = H + (1-H)\left(\frac{F^i_W}{2n_i-1} + \frac{(2n_i-2)\theta^i_B}{2n_i-1}\right)$

or $\tilde{M}^i = \frac{2n_i}{2n_i-1}\sum_u \tilde{p}^2_{iu} - \frac{1}{2n_i-1}$, $\mathcal{E}(\tilde{M}^i) = H + (1-H)\theta^i$

Average over populations

$\tilde{M}^W = \frac{1}{r}\sum_{i=1}^r \tilde{M}^i$, $\mathcal{E}(\tilde{M}^W = H + (1-H)\theta^W$

One allele from each of populations i, i'

$\tilde{M}^{ii'} = \sum_u \tilde{p}_{iu}\tilde{p}_{i'u}$, $\mathcal{E}(\tilde{M}^{ii'} = H + (1-H)\theta^{ii'}$

Average over pairs of populations

$\tilde{M}^B = \frac{1}{r(r-1)}\sum_{i=1}^r\sum_{i'=1,i'\neq i}^r \tilde{M}^{ii'}, \mathcal{E}(\tilde{M}^B) = H + (1-H)\theta^B$

Matching Proportions for Individuals

For individuals i with genotypes AA, AB, BB the allelic matching proportions $\tilde{M}_i$ are $1, 0, 1$ respectively.

For pairs of individuals i, i', each with genotypes AA, AB, BB, the matching proportions $\tilde{M}_{ii'}$ are

		i'		
	$\tilde{M}_{ii'}$	AA	AB	BB
	AA	1	0.5	0
i	AB	0.5	0.5	0.5
	BB	0	0.5	1

Notation

For SNPs:

$$\begin{aligned} \tilde{M}^i_j &= (X^i_j - 1)^2 \\ \tilde{M}^i_{jj'} &= \frac{1}{2}[1 + (X^i_j - 1)(X^i_{j'} - 1)] \end{aligned}$$

Averages:

$$\begin{aligned} F^i_W &= \frac{1}{n_i} \sum_{j=1}^{n_i} F^i_j \\ \theta^i_B &= \frac{1}{n_i(n_i-1)} \sum_{j=1}^{n_i} \sum_{j'=1, j' \neq j}^{n_i} \theta^i_{jj'} \\ \theta^W &= \frac{1}{r} \sum_{i=1}^{r} \theta^i \\ \theta^B &= \frac{1}{r(r-1)} \sum_{i=1}^{r} \sum_{i'=1, i' \neq i}^{2} \theta^{ii'} \end{aligned}$$

Estimates for Individuals

Two distinct alleles from individual j in population i

$$\widehat{\beta}^i_j = \frac{\tilde{M}^i_j - \tilde{M}^i_B}{1-\tilde{M}^i_B},\ \mathcal{E}(\widehat{\beta}^i_j) = \beta^i_j = \frac{F^i_j - \theta^i_B}{1-\theta^i_B}$$

Average over individuals in population i

$$\widehat{\beta}^i_W = \frac{\tilde{M}^i_W - \tilde{M}^i_B}{1-\tilde{M}^i_B},\ \mathcal{E}(\widehat{\beta}^i_W) = F^i_{IS} = \beta^i_W = \frac{F^i_W - \theta^i_B}{1-\theta^i_B}$$

One allele from each of individuals j, j' in population i

$$\widehat{\beta}^i_{jj'} = \frac{\tilde{M}^i_{jj'} - \tilde{M}^i_B}{1-\tilde{M}^i_B},\ \mathcal{E}(\widehat{\beta}^i_{jj'}) = \beta^i_{jj'} = \frac{\theta^i_{jj'} - \theta^i_B}{1-\theta^i_B}$$

Average over pairs of individuals in population i

$$\widehat{\beta}^i_B = \frac{\tilde{M}^i_B - \tilde{M}^i_B}{1-\tilde{M}^i_B},\ \mathcal{E}(\widehat{\beta}^i_B) = \beta^i_B = 0$$

Estimates for Populations

Two distinct alleles from population i

$\hat{\beta}^i = \frac{\tilde{M}^i - \tilde{M}^B}{1-\tilde{M}^B}, \;\; \mathcal{E}(\hat{\beta}^i) = \beta^i = \frac{\theta^i - \theta^B}{1-\theta^B}$

Average over populations

$\hat{\beta}^W = \frac{\tilde{M}^W - \tilde{M}^B}{1-\tilde{M}^B}, \;\; \mathcal{E}(\hat{\beta}^W) = F_{ST} = \beta^W = \frac{\theta^W - \theta^B}{1-\theta^B}$

One allele from each of populations i, i'

$\hat{\beta}^{ii'} = \frac{\tilde{M}^{ii'} - \tilde{M}^B}{1-\tilde{M}^B}, \;\; \mathcal{E}(\hat{\beta}^{ii'}) = \beta^{ii'} = \frac{\theta^{ii'} - \theta^B}{1-\theta^B}$

Average over populations

$\hat{\beta}^B = \frac{\tilde{M}^B - \tilde{M}^B}{1-\tilde{M}^B}, \;\; \mathcal{E}(\hat{\beta}^B) = \beta^B = 0$

Multiple Loci

The unweighted estimators for locus l are of the form

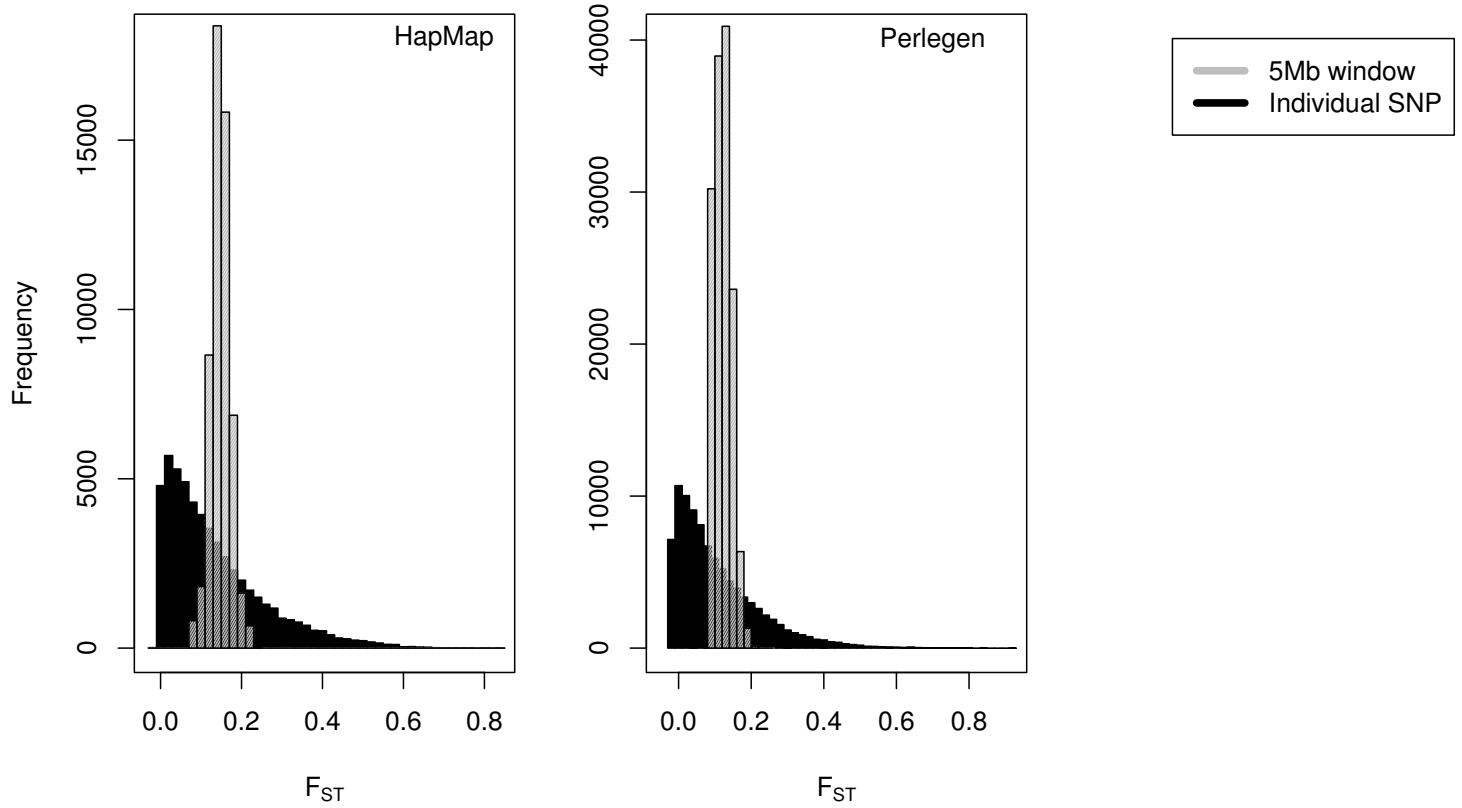
$$\text{Estimator}_l \;=\; \frac{\tilde{M}_l^x - \tilde{M}_l^B}{1 - \tilde{M}_l^B}, \;\; x = i, W, ij$$

Here x can refer to population i, the average W over populations, or the pair of populations i, j. With several loci, these can be extended to

$$\text{Estimator} \;=\; \frac{\sum_l (\tilde{M}_l^x - \tilde{M}_l^B)}{\sum_l (1 - \tilde{M}_l^B)} \;\; x = i, ij, W$$

and these estimate $(\theta^x - \theta^B)/(1 - \theta^B)$ if each locus has the same value of the θ's. Otherwise they estimate a weighted average of the different θ values, where the weights are functions of the allele frequencies at the loci in the sum.

Effect of Number of Loci



Weir and Cockerham 1984 Estimator

The WC84 estimate assumed all θ^i were the same and all $\theta^{ii'}$ were zero. If those assumptions are relaxed, the WC84 estimator has expectation

$$\mathcal{E}(\hat{\theta}_{\mathrm{WC}}) = \frac{\theta^{W*} - \theta^{B*} + Q}{1 - \theta^{B*} + Q}$$

where

$$\theta^{W*} = \frac{\sum_i n_i^c \theta^i}{\sum_i n_i^c} \quad , \quad \theta^{B*} = \frac{\sum_{i \neq i'} n_i n_{i'} \theta^{ii'}}{\sum_{i \neq i'} n_i n_{i'}}$$

$$n_i^c = n_i - \frac{n_i^2}{\sum_i n_i} \quad , \quad n_c = \frac{1}{r-1} \sum_i n_i^c$$

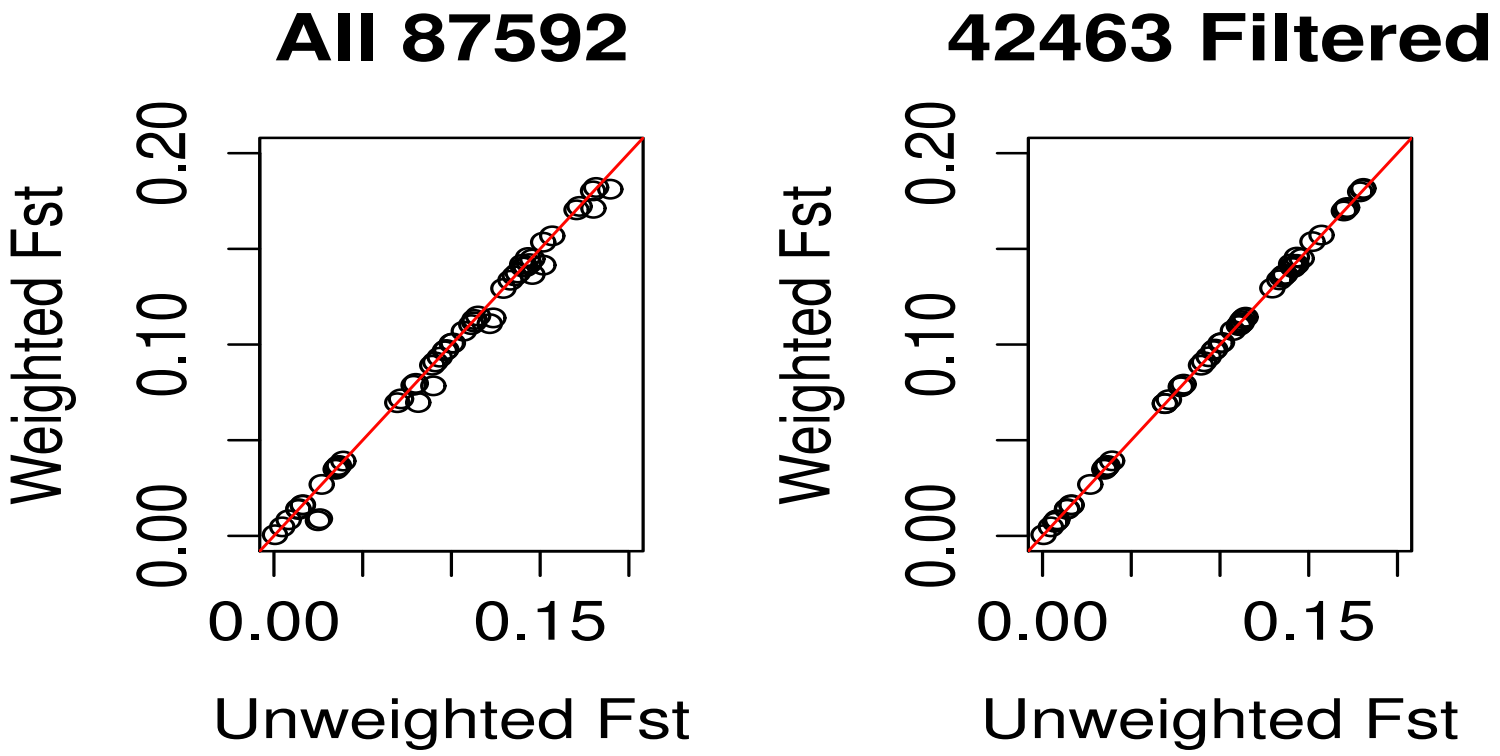
$$Q = \frac{1}{(r-1)n_c} \sum_i \left(\frac{n_i}{\bar{n}} - 1 \right) \theta_i$$

If the WC84 model holds ($\theta^{ii} = \theta$), **or if** $n_i = n$, **or if** n_c is large, then $Q = 0$ and $\mathcal{E}(\hat{\theta}_{WC}) = (\theta^W - \theta^B)/(1 - \theta^B)$.

HapMap III SNP Data

Code	Population Description	Sample size
ASW	African ancestry in Southwest USA	142
CEU	Utah residents with Northern and Western European ancestry from CEPH collection	324
CHB	Han Chinese in Beijing, China	160
CHD	Chinese in Metropolitan Denver, Colorado	140
GIH	Gujarati Indians in Houston, Texas	166
JPT	Japanese in Tokyo, Japan	168
LWK	Luhya in Webuye, Kenya	166
MXL	Mexican ancestry in Los Angeles, California	142
MKK	Maasai in Kinyawa, Kenya	342
TSI	Toscani in Italia	154
YRI	Yoruba in Ibadan, Nigeria	326

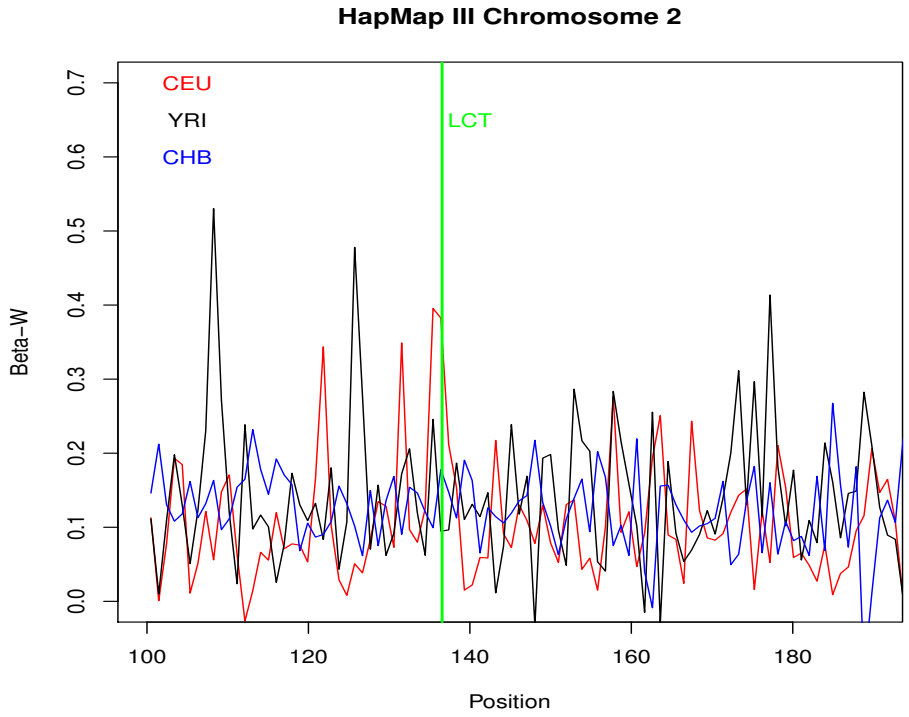
Weighted vs Unweighted Estimators $\widehat{\beta}_{ii'}$



Chr 1 SNPs. Weighted: WC84; Unweighted: New β estimates. Right hand plot ignores SNPs with less than 5 observed copies in a population.

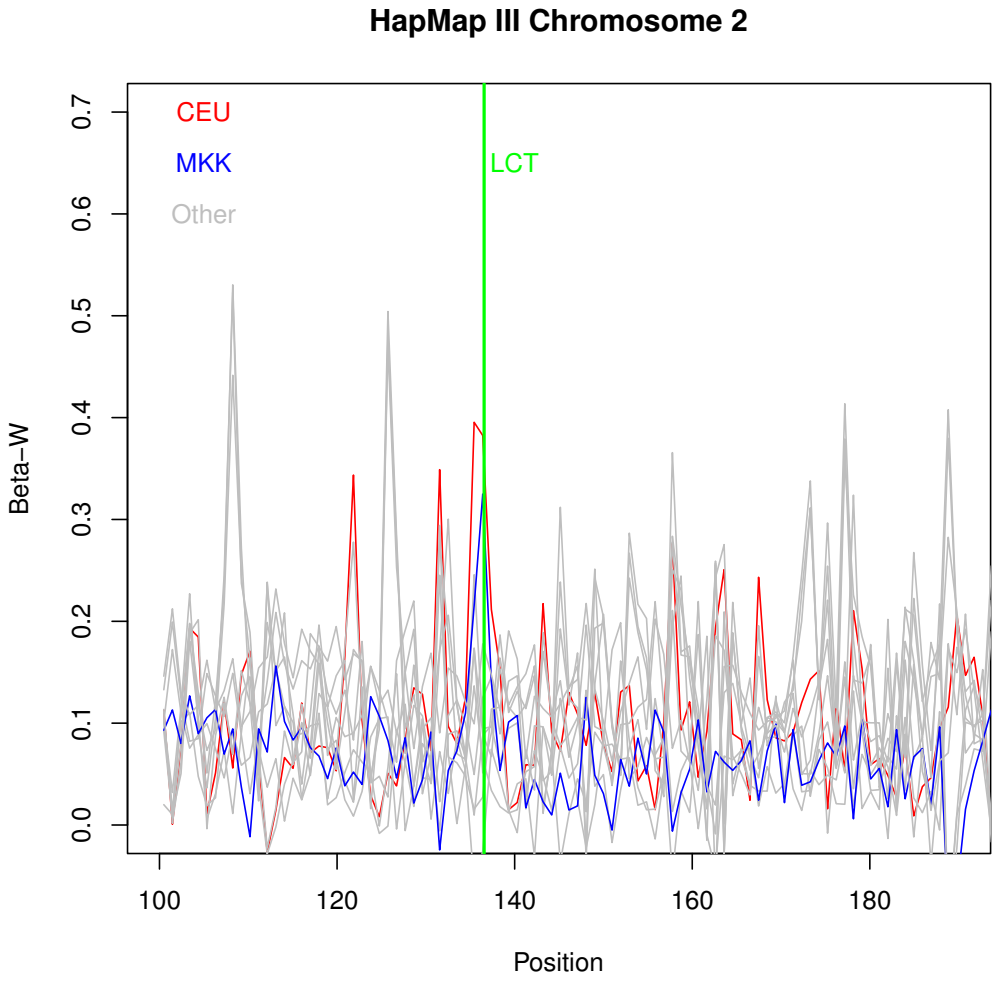
Multiple Populations: Detecting Selection

$\widehat{\beta}^i$ in LCT Region: $i = 1, 2, 3$.



These plots are for $\widehat{\beta}_i$'s, averaged over several SNPs. Little signal in the average $\widehat{\beta}_W$.

$\widehat{\beta}^i$ in LCT Region: $i = 1, 2, \ldots 11$.



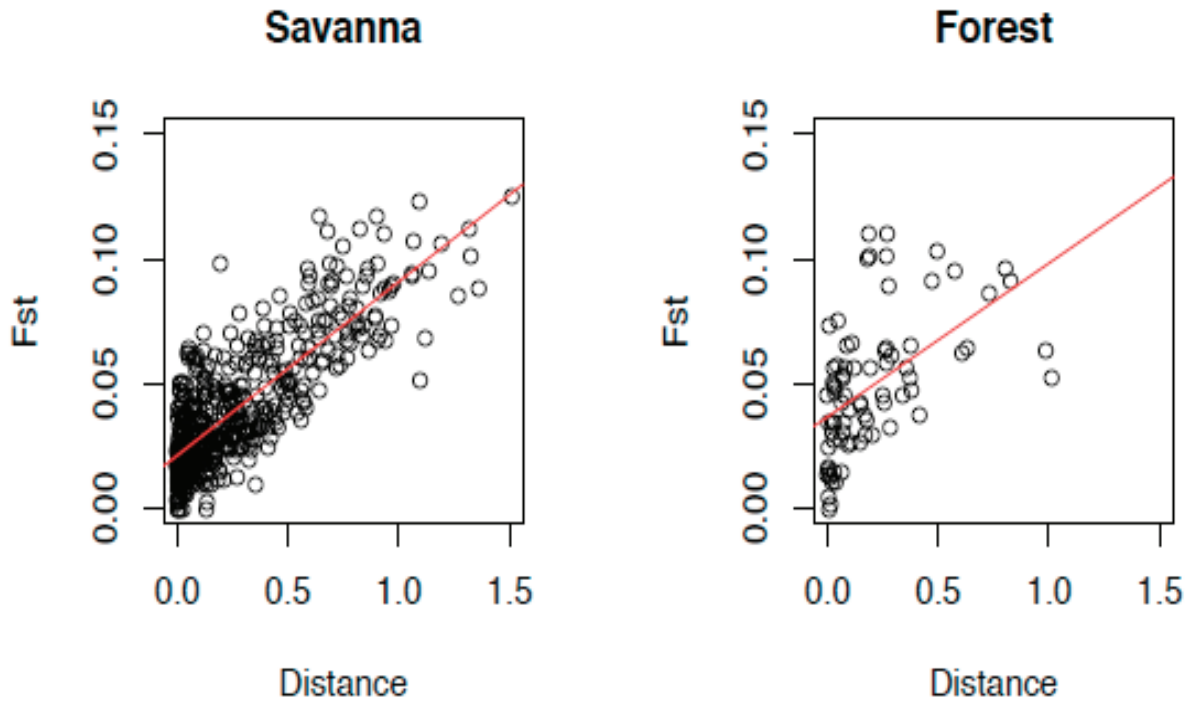
MKK Population

"The Maasai are a pastoral people in Kenya and Tanzania, whose traditional diet of milk, $\cdots$ is rich in lactose $\cdots$. In spite of this, they have low levels of blood cholesterol, and seldom suffer from gallstones or cardiac diseases.

Analysis of HapMap 3 data using Fixation Index (Fst) identified genomic regions and single nucleotide polymorphisms (SNPs) as strong candidates for recent selection for lactase persistence $\cdots$ from the Maasai population in Kinyawa, Kenya (MKK). The strongest signal identified by all three metrics was a 1.7 Mb region on Chr2q21. This region contains the gene LCT (Lactase) involved in lactase persistence."

[Wagh et al., PLoS One 7: e44751, 2012.]

Two Populations: $\widehat{\beta}_W$ as Geographic Distance



[Wasser et al., Science 349:84–87, 2015.]

Two Populations: General

For loci, such as SNPs, with two alleles, the β estimates for two populations can be written as

$$\begin{aligned}
\widehat{\beta}^1 &= \frac{(\tilde{p}_1 - \tilde{p}_2)(2\tilde{p}_1 - 1)}{\tilde{p}_1(1-\tilde{p}_2) + \tilde{p}_2(1-\tilde{p}_1)} \\
\widehat{\beta}^2 &= \frac{(\tilde{p}_2 - \tilde{p}_1)(2\tilde{p}_2 - 1)}{\tilde{p}_1(1-\tilde{p}_2) + \tilde{p}_2(1-\tilde{p}_1)} \\
\widehat{\beta}^W &= \frac{(\tilde{p}_1 - \tilde{p}_2)^2}{\tilde{p}_1(1-\tilde{p}_2) + \tilde{p}_2(1-\tilde{p}_1)}
\end{aligned}$$

Each estimate reflects difference of the two sample allele frequencies. Either $\widehat{\beta}^1$ or $\widehat{\beta}^2$ can be negative, but $\widehat{\beta}_W$ is positive.

The two populations could be cases and controls for a disease.

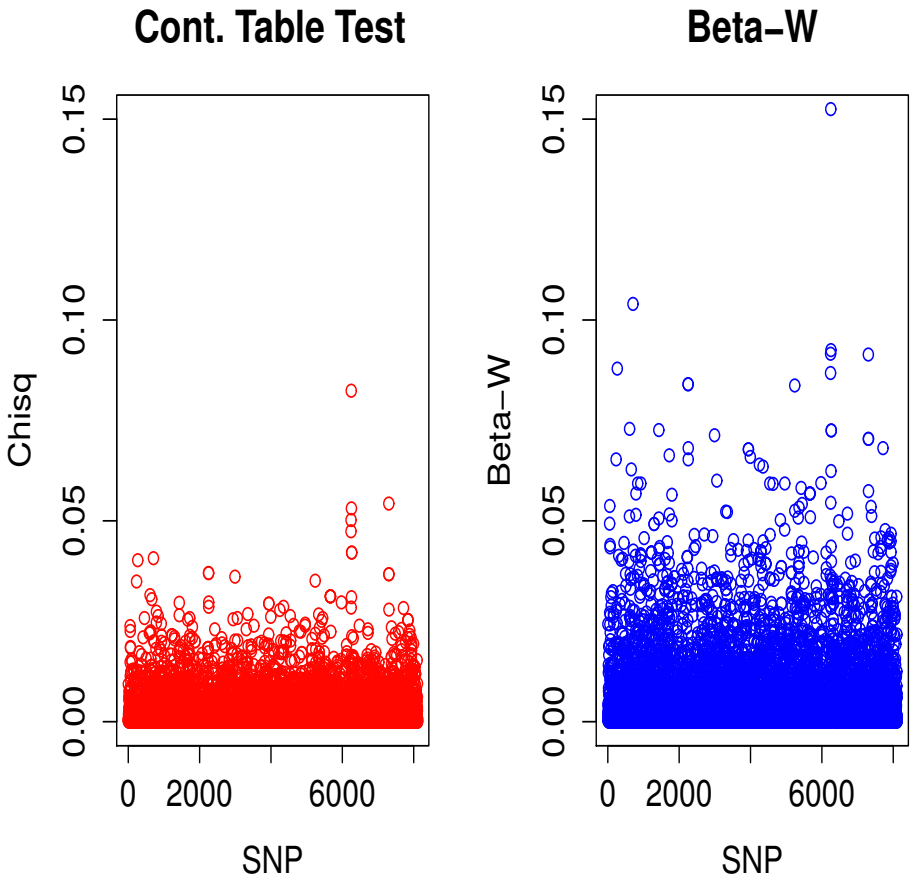
Two Populations: Case-Control Test

A 2×2 contingency table statistic for marker allele counts in cases and controls can be used as an allelic case-control test statistic for marker-trait association. This quantity is

$$X^2 = \frac{n_1 n_2 (\tilde{p}_1 - \tilde{p}_2)^2}{(n_1 + n_2)^2 \bar{p}(1 - \bar{p})}$$

where $\bar{p} = (n_1 \tilde{p}_1 + n_2 \tilde{p}_2)/(n_1 + n_2)$.

AMD Data

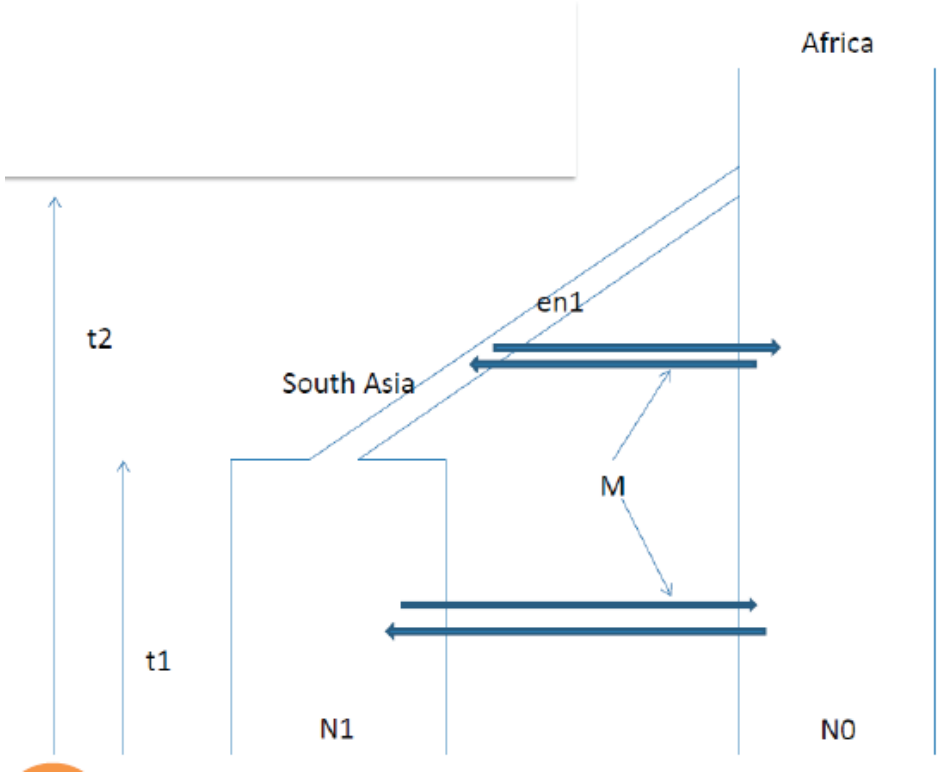

Cont. Table Test
Chisq
0.00
0.05
0.10
0.15
0
2000
6000
SNP
Beta–W
Beta–W
0.00
0.05
0.10
0.15
0
2000
6000
SNP

Private Alleles

Now suppose the reference allele is private to population 1: $\tilde{p}_1 = \epsilon$ and $\tilde{p}_i = 0, i \neq 1$:

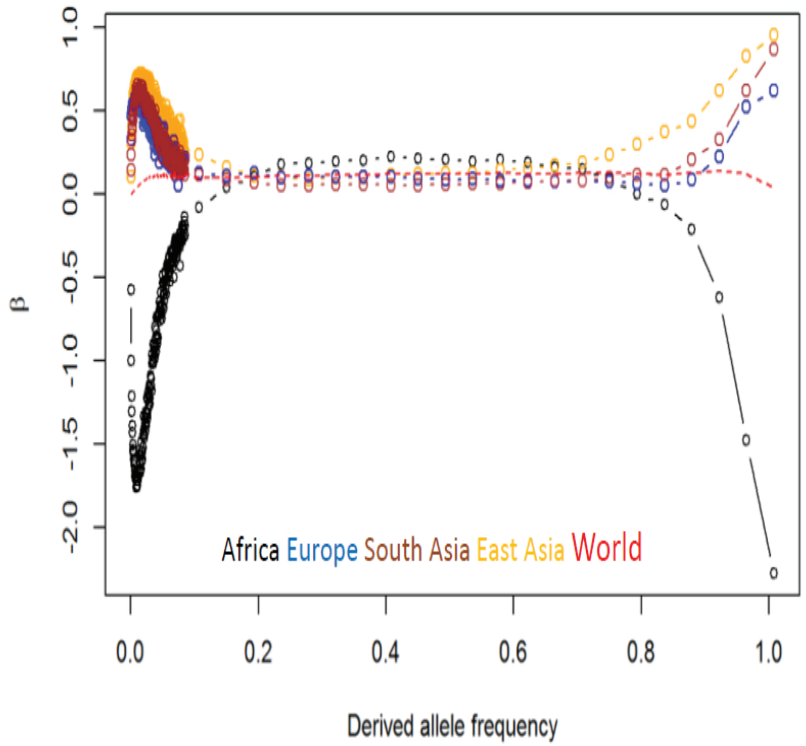
$$\begin{aligned} \widehat{\beta}^1 &= 1 - r(1-\epsilon) < 0 \\ \widehat{\beta}^i &= 1, i \neq 1 \\ \widehat{\beta}^W &= \epsilon > 0 \end{aligned}$$

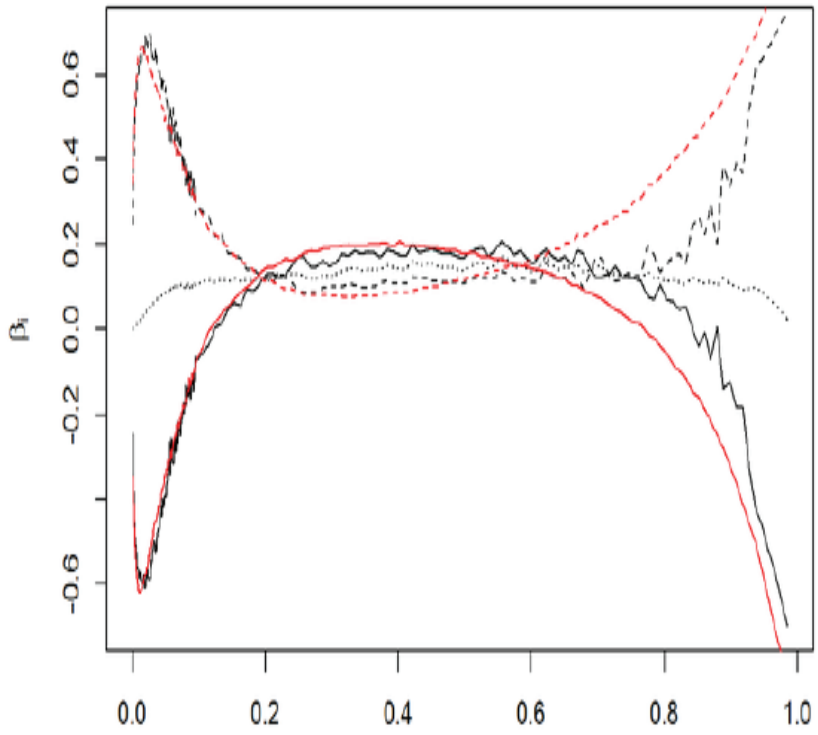
Human History: model



Used $N_0 = 10,000; N_1 = 9,000; eN1 = 1,800; t_1 = 4,600; t_2 = 45,000, M = 7$ in simulations.

Human History: data





Data on left, simulations on right.

Forensic Application

A key issue in forensic genetics is to determine the matching probability.

The probability an unknown person has a genotype, given that the suspect has been seen to have that type is

$$\Pr(AA|AA) = \frac{[3\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)}$$

$$\Pr(AB|AB) = \frac{23\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)}$$

Here θ refers to the population to which the unknown person and the suspect belong.

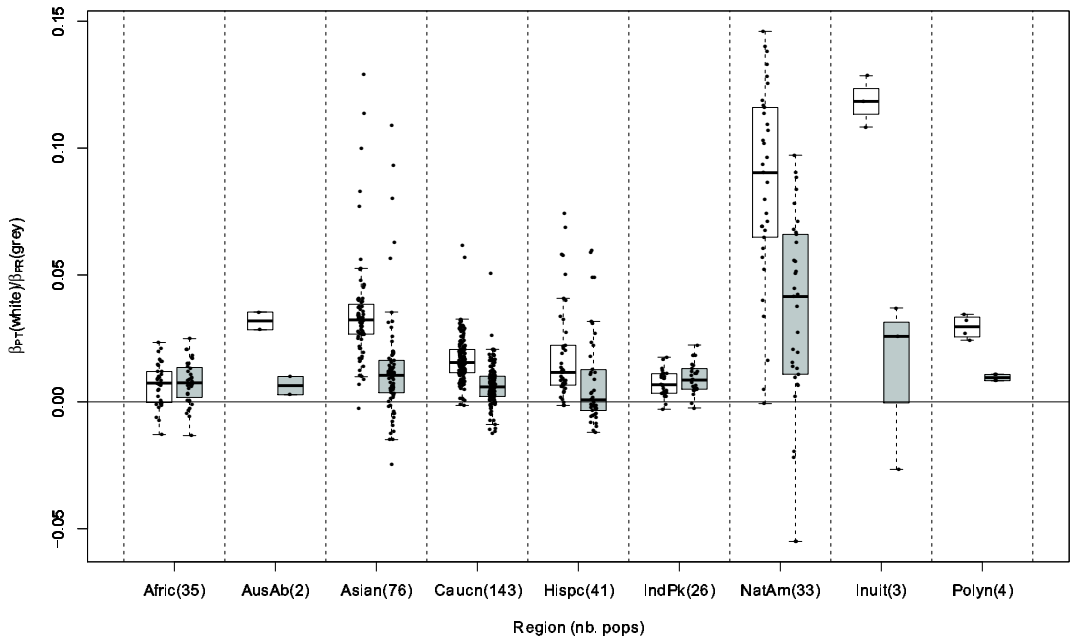
Effect of Reference Group of Populations

Buckleton et al. (2016) gave population-specific F_{ST} estimates for a set of 446 populations, using published data for 24 microsatellite loci collected for forensic purposes.

For a set of African populations, the average within-population matching proportion was $\tilde{M}^W = 0.1884$ and the average between-population-pair averages were $\tilde{M}^B = 0.1691$ within the African region and $\tilde{M}^B = 0.1726$ for all pairs of populations in the study. There is a larger F_{ST} for the set of African populations ($\hat{\beta}^W = 0.0082$) with Africa as a reference set than there is ($\hat{\beta}^W = 0.0020$) with the world as a reference set.

For a set of Inuit populations, the average within-population matching proportion was $\tilde{M}^W = 0.4379$ whereas the average between-population-pair matching proportions were $\tilde{M}^B = 0.1726$ for pairs within the Inuit group and $\tilde{M}^B = 0.0090$ for all pairs of populations in the study. There is a smaller F_{ST} ($\hat{\beta}^W = 0.0205$) with Inuit as a reference set than with the world as a reference set ($\hat{\beta}^W = 0.1057$).

Effect of Reference Group of Populations



Plots of $\widehat{\beta}_i$. For the white box plots, $\tilde{M}_B$ is for all pairs of populations. For the grey box plots, $\tilde{M}_B$ is for all pairs of populations in that group. [Buckleton et al. 2016. Forensic Science International:Genetics 23:91-100.]

Worldwide Autosomal-STR Survey

Buckleton et al, Forensic Sci Int, 2016 compiled a survey of 250 published papers showing allele frequencies at 24 forensic STR markers from 446 populations in 8 ancestral groups. Represents data from 494,473 individuals.

The ancestral groups were identified by a combination of clustering and geographic criteria.

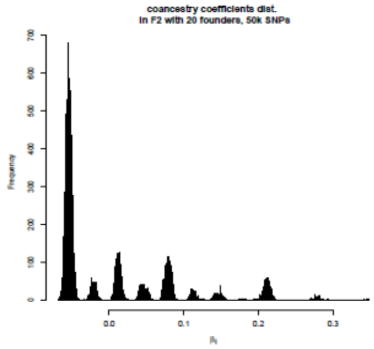
Moment estimates were obtained for each locus l in each population i from

$$\hat{\beta}_l^i = \frac{\tilde{M}_l^i - \tilde{M}_l^B}{1 - \tilde{M}_l^B}$$

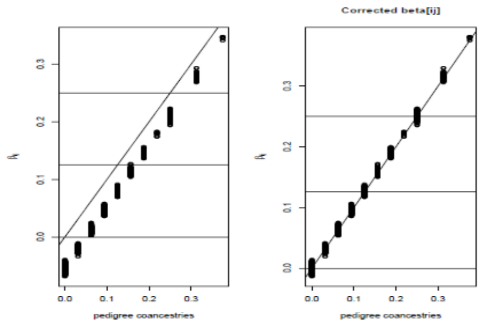
STR Survey: $\widehat{\beta}^W$ Values for Groups and Loci

	Geographic Region								
Locus	Africa	AusAb	Asian	Cauc	Hisp	IndPK	NatAm	Poly	Aver.
CSF1PO	0.003	0.002	0.008	0.008	0.002	0.007	0.055	0.026	0.011
D1S1656	0.000	0.000	0.000	0.002	0.003	0.000	0.000	0.000	0.011
D2S441	0.000	0.000	0.002	0.003	0.021	0.000	0.000	0.000	0.020
D2S1338	0.009	0.004	0.011	0.017	0.013	0.003	0.023	0.005	0.031
D3S1358	0.004	0.010	0.009	0.006	0.012	0.040	0.079	0.001	0.025
D5S818	0.002	0.013	0.009	0.008	0.014	0.018	0.044	0.007	0.029
D6S1043	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016
D7S820	0.004	0.021	0.010	0.007	0.007	0.046	0.030	0.005	0.026
D8S1179	0.003	0.007	0.012	0.006	0.002	0.031	0.020	0.008	0.019
D10S1248	0.000	0.000	0.000	0.002	0.004	0.000	0.000	0.000	0.007
D12S391	0.000	0.000	0.000	0.003	0.020	0.000	0.000	0.000	0.010
D13S317	0.015	0.016	0.013	0.008	0.014	0.025	0.050	0.014	0.038
D16S539	0.007	0.002	0.015	0.006	0.009	0.005	0.048	0.004	0.021
D18S51	0.011	0.012	0.014	0.006	0.004	0.010	0.033	0.003	0.018
D19S433	0.009	0.001	0.009	0.010	0.014	0.000	0.022	0.014	0.023
D21S11	0.014	0.012	0.013	0.007	0.006	0.023	0.067	0.018	0.021
D22S1045	0.000	0.000	0.007	0.001	0.000	0.000	0.000	0.000	0.015
FGA	0.002	0.009	0.012	0.004	0.007	0.016	0.021	0.006	0.013
PENTAD	0.008	0.000	0.012	0.012	0.002	0.017	0.000	0.000	0.022
PENTAE	0.002	0.000	0.017	0.006	0.003	0.012	0.000	0.000	0.020
SE33	0.000	0.000	0.012	0.001	0.000	0.000	0.000	0.000	0.004
TH01	0.022	0.001	0.022	0.016	0.018	0.014	0.071	0.017	0.071
TPOX	0.019	0.087	0.016	0.011	0.007	0.018	0.064	0.031	0.035
VWA	0.009	0.007	0.017	0.007	0.012	0.022	0.028	0.005	0.023
All Loci	0.006	0.014	0.010	0.007	0.008	0.018	0.043	0.011	0.022

Relatedness Estimates from Simulated Data



Change estimates to be relative to average of least related values.



Before (left) and after (right) change.

Standard Estimates

The quantities $\widehat{\theta}_{jj'}$ (e.g. Astle and Balding, 2009, Statistical Science 24:451-471)

$$\widehat{\theta}_{jj'} = \frac{(X_j - 2p)(X_{j'} - 2p)}{4p(1-p)}$$

are unbiased for $\theta_{jj'}$ if the reference allele frequencies p are known.

As the allele frequencies are not known, it is usual to replace p with $\tilde{p}$, the sample allele frequency for the set of individuals being studied. Then, for large sample sizes,

$$\mathcal{E}(\widehat{\theta}_{jj'}) = \frac{\theta_{jj'} - \psi_j - \psi_{j'} + \theta_B}{1 - \theta_B} \neq \theta_{jj'}$$

where $\psi_j = \sum_{j' \neq j} \theta_{jj'}/(r-1)$.

KING-robust Kinship Estimates

The influence of all members of a sample is eliminated by the KING-robust* estimates for pairs of individuals:

$$\widehat{\theta}^{K}_{jj'} \;=\; \frac{N(AB,AB)_{jj'} - 2N(AA,BB)_{jj'}}{N(AB)_{j} + N(AB)_{j'}}$$

where the N's are the numbers of loci with the indicated genotypes for individuals j and individual pairs j, j'. The expected values of these estimates are

$$\mathcal{E}(\widehat{\theta}^{K}_{jj'}) \;=\; \frac{\theta_{jj'} - \frac{1}{2}(F_j + F_{j'})}{1 - \frac{1}{2}(F_j + F_{j'})}$$

These estimates for a pair of individuals are relative to the average inbreeding coefficients of those individuals.

[* Manichaikul et al., Bioinformatics 26:2867-2873, 2010.]

Inbreeding Standard Estimates

For inbreeding, the quantities $\widehat{\theta}_{jj}$

$$\widehat{\theta}_{jj'} \;=\; \frac{(X_j - 2p)^2)}{4p(1-p)}$$

are unbiased for $\theta_{jj} = (1+F_j)/2$ if the reference allele frequencies p are known.

As the allele frequencies are not known, it is usual to replace p with $\tilde{p}$, the sample allele frequency for the set of individuals being studied. Then, for large sample sizes,

$$\mathcal{E}(\widehat{\theta}_{jj}) \;=\; \frac{\theta_{jj} - 2\psi_j + \theta_B}{1-\theta_B} \neq \theta_{jj'}$$

where $\psi_j = \sum_{j' \neq j} \theta_{jj'}/(r-1)$.

Inbreeding Standard Estimates

Then the standard estimator (e.g. Astle and Balding, 2009, Statistical Science 24:451-471) is formed by averaging over loci $l, l = 1, 2, \ldots L$:

$$\hat{F}_1 = \frac{1}{L}\sum_{l=1}^{L}\frac{(X_l - 2p_l)^2}{2p_l(1-p_l)} - 1$$

Another one (Yang et al, Nature Genetics 42:565-569, 2010) is

$$\begin{aligned}\hat{F}_2 &= \frac{1}{L}\sum_{l=1}^{L}\frac{X_l^2 - (1+2p_l)X_l + 2p_l^2}{2p_l(1-p_l)} \\ &= \frac{1}{L}\sum_{l=1}^{L}\frac{(X-2p_l)^2 - (1-2p_l)(X-2p_l)}{2p_l(1-p_l)} - 1\end{aligned}$$

If the p_l are known, both these are unbiased. The second one has a smaller variance.

Alternative MOMs for Individual Inbreeding Coefficients

The variances can be reduced by an alternative weighting over loci:

$$\widehat{F}_3 = \frac{\sum_{l=1}^{L}(X_l - 2p_l)^2}{\sum_{l=1}^{L} 2p_l(1-p_l)} - 1$$

$$\widehat{F}_4 = \frac{\sum_{l=1}^{L}[X_l^2 - (1+2p_l)X_l + 2p_l^2]}{\sum_{l=1}^{L} 2p_l(1-p_l)}$$

Toy Example

Suppose five loci have genotypes

$$MM, Mm, mm, Mm, MM$$

.

The preferred moment estimate $\hat{F}_3$ is

$$\hat{F}_3 = \frac{(2-2p_1)^2 + (1-2p_2)^2 + (2p_3)^2 + (1-2p_4)^2 + (2-2p_5)^2}{2[p_1(1-p_1) + p_2(1-p_2) + p_3(1-p_3) + p_4(1-p_4) + p_5(1-p_5)]} - 1$$

If all five M allele frequencies were 0.5,

$$\hat{F}_3 = \frac{1+0+1+0+1}{2[1/4+1/4+1/4+1/4+1/4]} - 1 = 0.2$$

MLE for Individual Inbreeding Coefficients

To avoid having to choose among MOMs can set up an MLE although there is more numerical work needed. An iterative method makes use of Bayes' theorem. If F represents the probability the individual in question has two IBD alleles at a locus, i.e. is inbred at that locus,

$$\begin{aligned}
\Pr(A_lA_l|\text{inbred}) = p_l \quad &, \quad \Pr(A_lA_l|\text{Not inbred}) = p_l^2 \\
\Pr(A_la_l|\text{inbred}) = 0 \quad &, \quad \Pr(A_la_l|\text{Not inbred}) = 2p_l(1-p_l) \\
\Pr(a_la_l|\text{inbred}) = 1-p_l \quad &, \quad \Pr(a_la_l|\text{Not inbred}) = (1-p_l)^2
\end{aligned}$$

From Bayes' theorem then

$$\begin{aligned}
\Pr(\text{inbred}|A_lA_l) &= \frac{\Pr(A_lA_l|\text{inbred})\Pr(\text{inbred})}{\Pr(A_lA_l)} = \frac{p_lF}{p_l^2 + Fp_l(1-p_l)} \\
\Pr(\text{inbred}|A_la_l) &= \frac{\Pr(A_la_l|\text{inbred})\Pr(\text{inbred})}{\Pr(A_la_l)} = 0 \\
\Pr(\text{inbred}|a_la_l) &= \frac{\Pr(a_la_l|\text{inbred})\Pr(\text{inbred})}{\Pr(a_la_l)} = \frac{(1-p_l)F}{(1-p_l)^2 + Fp_l(1-p_l)}
\end{aligned}$$

MLE for Individual Inbreeding Coefficients

This suggests an iterative scheme: assign an initial value to F, and then average the updated values over loci. If G_l is the genotype at locus l, the updated value F' is

$$F' = \frac{1}{L}\sum_{l=1}^{L} \Pr(\text{inbred}|G_l)$$

This value is then substituted into the right hand side and the process continues until convergence.

Toy Example

Suppose 5 loci have genotypes

$$MM, Mm, mm, Mm, MM$$

.

Then the updated estimate is

$$F' = \frac{1}{5}\left(\frac{F}{p_1 + F(1-p_1)} + 0 + \frac{F}{(1-p_3) + Fp_3} + 0 + \frac{F}{p_5 + (1-p_5)F}\right)$$

If all the $p_l = 0.5$,

$$F' = \frac{1}{5}\left(\frac{2F}{1+F} + 0 + \frac{2F}{1+F} + 0 + \frac{2F}{1+F}\right) = \frac{6F}{5(1+F)}$$

and this converges to $F = 0.2$.

Realistic Example

Using 1369 SNPs spread out along chromosome 22, Xiuwen Zheng found that

```
Minimum MAF   .000
 Average of ratios MOM mean and sd  .0484  .2851
 Ratio of averages MOM mean and sd  .0272  .0874
                   MLE mean and sd  .0381  .0576
 Minimum MAF   .050
 Average of ratios MOM mean and sd  .0359  .1516
 Ratio of averages MOM mean and sd  .0264  .0860
                   MLE mean and sd  .0380  .0576
 Minimum MAF   .100
 Average of ratios MOM mean and sd  .0358  .1585
 Ratio of averages MOM mean and sd  .0260  .0864
                   MLE mean and sd  .0382  .0577
```

Method of Moments for Relatedness Coefficients

PLINK (Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R,Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., & Sham, P.C. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81,559–575.) uses MOM to estimate three IBD coefficients k_0, k_1, k_2 for non-inbred relatives. Two individuals are scored as being in IBS states 0,1,2.

State : Genotypes	Probability
$2 : (MM, MM), (mm, mm), (Mm, Mm)$	$(p_M^4 + 4p_M^2 p_m^2 + p_m^4)k_0$ $+ k_1(p_M^3 + p_M p_m + p_m^3) + k_2$
$1 : (MM, Mm), (Mm, MM), (mm, Mm), (Mm, mm)$	$4p_M p_m(p_M^2 + p_m^2)k_0 + 2p_M p_m k_1$
$0 : (MM, mm), (mm, MM)$	$2p_M^2 p_m^2 k_0$

MOM Approach: k_0

Count the number of loci in IBS state $i; i = 0, 1, 2$. These numbers are N_0, N_1, N_2. The previous table gives the probabilities of IBS state i given IBD state j. From

$$\Pr(\text{IBS} = 0) \;=\; \Pr(\text{IBS} = 0|\text{IBD} = 0)\Pr(\text{IBD} = 0)$$

sum over loci l to get

$$N_0 \;=\; \sum_l 2p_l^2(1-p_l)^2 \Pr(\text{IBD} = 0)$$

This gives a moment estimate

$$\Pr(\text{IBD} = 0) \;=\; \frac{N_0}{\sum_l 2p_l^2(1-p_l)^2}$$

MOM Approach: k_1

From

$$\begin{aligned}\Pr(\text{IBS}=1) &= \Pr(\text{IBS}=1|\text{IBD}=0)\Pr(\text{IBD}=0) \\ &\quad + \Pr(\text{IBS}=1|\text{IBD}=1)\Pr(\text{IBD}=1)\end{aligned}$$

sum over loci to get

$$\begin{aligned}N_1 &= \Pr(\text{IBD}=0)\sum_l 4p_l(1-p_l)[p_l^2+(1-p_l)^2] \\ &\quad + \Pr(\text{IBD}=1)\sum_l 2p_l(1-p_l)\end{aligned}$$

but we already have an estimate of $\Pr(\text{IBD}=0)$. Therefore

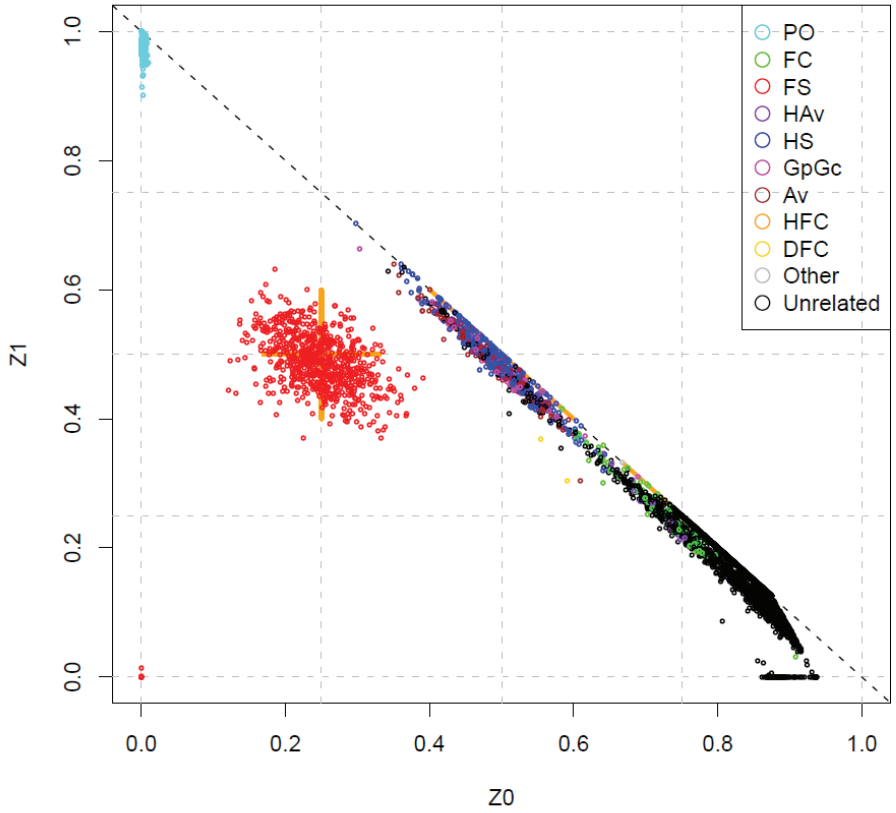
$$\Pr(\text{IBD}=1) = \frac{N_1 - \sum_l 4p_l(1-p_l)[p_l^2+(1-p_l)^2]\Pr(\text{IBD}=0)}{\sum_l 2p_l(1-p_l)}$$

MOM Approach: k_2

Having estimated k_0 and k_1, find $\widehat{k}_2$ as $1 - \widehat{k}_0 - \widehat{k}_1$.

Could then estimate θ as $\widehat{k}_2/2 + \widehat{k}_1/4$ or could go to a direct estimate.

PLINK Example



MLE for Relatedness Coefficients

For any SNP there are six distinct pairs of genotypes with probabilities depending on allele frequencies for that SNP and on a set of three k parameters that are assumed to be the same for all SNPs. If G is the observed pair of genotypes, we know the conditional probabilities $\Pr(G|D_i)$ where the D_i represent the identity states (with probabilities k_i).

G	$\Pr(G) = \sum_i \Pr(G\|D_i)k_i$
MM, MM	$k_2 p_M^2 + k_1 p_M^3 + k_0 p_M^4$
mm, mm	$k_2 p_m^2 + k_1 p_m^3 + k_0 p_m^4$
MM, mm	$2k_0 p_M^2 p_m^2$
MM, Mm	$2k_1 p_M^2 p_m + 4k_0 p_M^3 p_m$
mm, Mm	$2k_1 p_M p_m^2 + 4k_0 p_M p_m^3$
Mm, Mm	$2k_2 p_M p_m + k_1 p_M p_m + 4k_0 p_M^2 p_m^2$

MLE for Relatedness Coefficients

An iterative algorithm for estimating the k's from observed genotypes G_l at SNP l is based on Bayes' theorem for the probability of descent state $D_i, i = 0, 1, 2$:

$$\Pr(D_i|G_l) \;=\; \frac{\Pr(G_l|D_i)\Pr(D_i)}{\Pr(G_l)}$$

The procedure begins with initial estimates of the $k_i = \Pr(D_i)$'s.

The updated estimates are obtained by averaging over L loci:

$$k_i' \;=\; \frac{1}{L}\sum_{l=1}^{L}\left(\frac{\Pr(G_l|D_i)k_i}{\sum_j \Pr(G_l|D_j)k_j}\right), \;\; i = 0, 1, 2$$

These updated values are then substituted into the right hand side until they no longer change (or change by less than some specified small amount).

Toy Example

Suppose the 5-locus genotypes for U, V are:

$$MM, Mm, mm, Mm, MM$$

and

$$MM, mm, Mm, Mm, Mm$$

The updating equations are:

$$\begin{aligned} k_2' &= \frac{1}{5}\left(\frac{p_1^2 k_2}{p_1^2 k_2 + p_2^3 k_1 + p_1^4 k_0} + 0 + 0\right. \\ &\left. + \frac{2p_4(1-p_4)k_2}{2p_4(1-p_4)k_2 + p_4(1-p_4)k_1 + 4p_4^2(1-p_4)^2 k_0} + 0\right) \end{aligned}$$

Toy Example

$$
\begin{aligned}
k_1' \;=\; \frac{1}{5}\Bigg(& \frac{p_1^3 k_1}{p_1^2 k_2 + p_2^3 k_1 + p_1^4 k_0} + \frac{2p_2(1-p_2)^2 k_1}{2p_2(1-p_2)^2 k_1 + 4p_2(1-p_2)^3 k_0} \\
& + \frac{2p_3(1-p_3)^2 k_1}{2p_3(1-p_3)^2 k_1 + 4p_3(1-p_3)^3 k_0} \\
& + \frac{p_4(1-p_4)k_1}{2p_4(1-p_4)k_2 + p_4(1-p_4)k_1 + 4p_4^2(1-p_4)^2 k_0} \\
& + \frac{2p_5^2(1-p_5)k_1}{2p_5^2(1-p_5)k_1 + 4p_5^3(1-p_5)k_0} \Bigg)
\end{aligned}
$$

Toy Example

$$
\begin{aligned}
k_0' \;=\; \frac{1}{5}\Bigg(& \frac{p_1^4 k_0}{p_1^2 k_2 + p_2^3 k_1 + p_1^4 k_0} + \frac{4p_2(1-p_2)^3 k_0}{2p_2(1-p_2)^2 k_1 + 4p_2(1-p_2)^3 k_0} \\
& + \frac{4p_3(1-p_3)^3 k_0}{2p_3(1-p_3)^2 k_1 + 4p_3(1-p_3)^3 k_0} \\
& + \frac{4p_4^2(1-p_4^2) k_0}{2p_4(1-p_4) k_2 + p_4(1-p_4) k_1 + 4p_4^2(1-p_4)^2 k_0} \\
& + \frac{4p_5^3(1-p_5) k_0}{2p_5^2(1-p_5) k_1 + 4p_5^3(1-p_5) k_0} \Bigg)
\end{aligned}
$$

"RELPAIR" calculations

This approach compares the probabilities of two genotypes under alternative hypotheses; H_0: the individuals have a specified relationship, versus H_1: the individuals are unrelated. The alternative is that $k_0 = 1, k_1 = k_2 = 0$ so the likelihood ratios for the two hypotheses are:

$$\begin{aligned}
\mathrm{LR}(MM, MM) &= k_0 + k_1/p_M + k_2/p_M^2 \\
\mathrm{LR}(mm, mm) &= k_0 + k_1/p_m + k_2/p_m^2 \\
\mathrm{LR}(Mm, Mm) &= k_0 + k_1/(4p_M p_m) + k_2/(2p_M p_m)
\end{aligned}$$

$$\begin{aligned}
\mathrm{LR}(MM, Mm) &= k_0 + k_1/(2p_M) \\
\mathrm{LR}(mm, Mm) &= k_0 + k_1/(2p_m)
\end{aligned}$$

$$\mathrm{LR}(MM, mm) = k_0$$

Reality Check: Inbreeding and Relatedness

Inbreeding and relatedness estimates most easily expressed as allele dosages. For individual i the number of reference alleles A is X_i: these are 2,1,0 for AA, AB, BB.

The allelic matching proportions are:

$$\begin{aligned}
\tilde{M}_i &= (X_i - 1)^2 \\
\tilde{M}_{ii'} &= \frac{1}{2}[1 + (X_i - 1)(X_{i'} - 1)] \\
\tilde{M}_B &= \frac{1}{r(r-1)} \sum_{\substack{i=1 \\ i \neq i'}}^{r} \sum_{i'=1}^{r} \tilde{M}_B
\end{aligned}$$

The estimates and their expectations are

$$\hat{\beta}_i = \frac{\tilde{M}_i - \tilde{M}_B}{1 - \tilde{M}_B} \quad , \quad \mathcal{E}(\hat{\beta}_i) = \frac{F_i - \theta_B}{1 - \theta_B}$$

$$\hat{\beta}_{ii'} = \frac{\tilde{M}_{ii'} - \tilde{M}_B}{1 - \tilde{M}_B} \quad , \quad \mathcal{E}(\hat{\beta}_{ii'}) = \frac{\theta_{ii'} - \theta_B}{1 - \theta_B}$$

Example

What are the inbreeding and kinship estimates for individuals P, Q:

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
P	AA	AA	AA	AB	AB	AB	BB	BB	BB	BB
Q	AB	AB	AA	AB	AB	AB	AB	BB	BB	BB

$$\begin{aligned} \tilde{M}_P &= \\ \tilde{M}_Q &= \\ \tilde{M}_{PQ} &= \\ \hat{\beta}_P &= \\ \hat{\beta}_Q &= \\ \hat{\beta}_{PQ} &= \end{aligned}$$

Reality Check: Population Structure

First need to calculate allelic matching proportions within populations i. We start here from published reference allele frequencies $\tilde{p}_i$:

$$\tilde{M}_i = \tilde{p}_i^2 + (1-\tilde{p}_i)^2 \quad , \quad \tilde{M}_W = \frac{1}{r}\sum_{i=1}^{r} \tilde{M}_i$$

Then we need the allelic matching proportions between pairs of populations i and i':

$$\tilde{M}_{ii'} = \tilde{p}_i\tilde{p}_{i'} + (1-\tilde{p}_i)(1-\tilde{p}_{i'}) \quad , \quad \tilde{M}_B = \frac{1}{r(r-1)} \sum_{\substack{i=1 \\ i\neq i'}}^{r} \sum_{i'=1}^{r} \tilde{M}_B$$

The population-specific F_{ST} estimates are

$$\hat{\beta}_i = \frac{\tilde{M}_i - \tilde{M}_B}{1-\tilde{M}_B} \quad , \quad \hat{\beta}_W = F_{ST} = \frac{\tilde{M}_W - \tilde{M}_B}{1-\tilde{M}_B}$$

HapMap Data

	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MXL	MKK	TSI	YRI
SNP1	0.48	0.84	0.90	0.91	0.81	0.86	0.51	0.82	0.53	0.85	0.40
SNP2	0.07	0.09	0.47	0.44	0.28	0.41	0.19	0.04	0.14	0.11	0.05

Calculate population-specific F_{ST} for CEU, CHB and YRI, using each of SNP1 and SNP2.

	SNP1	SNP2	Both
$\tilde{M}_{\text{CEU}}$			
$\tilde{M}_{\text{CHB}}$			
$\tilde{M}_{\text{YRI}}$			
$\tilde{M}_{\text{CEU,CHB}}$			
$\tilde{M}_{\text{CEU,YRI}}$			
$\tilde{M}_{\text{CHB,YRI}}$			
$\tilde{M}_{\text{B}}$			
$\hat{\beta}_{\text{CEU}}$			
$\hat{\beta}_{\text{CHB}}$			
$\hat{\beta}_{\text{YRI}}$			
F_{ST}			

ASSOCIATION MAPPING

Association Mapping

Association methods use random samples from a population and are alternatives to methods based on pedigrees or crosses between inbred lines. The associations depend on linkage disequilibrium between marker and trait loci instead of depending on linkage between those loci as in pedigree or line cross methods.

A quantitative trait locus **T** contributes to a trait of interest. The QTL genotype cannot be observed but maybe it can be inferred, and the location of the QTL be estimated, from observations on the trait and the genotype at a genetic marker **M**.

Marker-Trait Genotype Frequencies

Each marker genotypic class M_iM_j is composed of a mixture of elements from each of the QTL classes, T_rT_s, where the proportion of QTL class T_rT_s contained within marker class M_iM_j is $Pr(T_rT_s|M_iM_j)$. With random mating, genotype frequencies are products of gamete frequencies. For example

$$\begin{aligned} \Pr(T_rT_r, M_iM_i) &= \Pr(T_rM_i)^2 \\ \Pr(T_rT_r|M_iM_i) &= \Pr(T_rM_i)^2/\Pr(M_i)^2 \end{aligned}$$

and gamete frequencies involve allele frequencies and linkage disequilibria:

$$\Pr(T_rM_i) = p_rp_i + D_{ri}$$

Two-allele Genotypes

	TT	Tt	tt
MM	P_{MT}^2	$2P_{MT}P_{Mt}$	P_{Mt}^2
Mm	$2P_{MT}P_{mT}$	$2P_{MT}P_{mt} + 2P_{Mt}P_{mT}$	$2P_{Mt}P_{mt}$
mm	P_{mT}^2	$2P_{mT}P_{mt}$	P_{mt}^2

Two-allele Gametes

	T	t
M	$P_{MT} = p_M p_T + D_{MT}$	$P_{Mt} = p_M p_t - D_{MT}$
m	$P_{mT} = p_m p_T - D_{MT}$	$P_{mt} = p_m p_t + D_{MT}$

$$\rho_{MT} = \frac{D_{MT}}{\sqrt{p_M p_m p_T p_t}}$$

$$\rho^2_{MT} = \frac{D^2_{MT}}{p_M p_m p_T p_t}$$

Marker and Trait Variables

Introduce variables X and G for loci **M** and **T**. The values of X will be assigned for the marker whereas the values G represent the genetic contributions to measured trait variables or to disease status. In either case, the Hardy-Weinberg assumption provides the following expressions for the means and variances:

$$\begin{aligned} \mathcal{E}(X) = \mu_X &= p_M^2 X_{MM} + 2p_M p_m X_{Mm} + p_m^2 X_{mm} \\ \mathcal{E}(G) = \mu_G &= p_T^2 G_{TT} + 2p_T p_t G_{Tt} + p_t^2 G_{tt} \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(X) &= \sigma^2_{A_M} + \sigma^2_{D_M} \\ \mathrm{Var}(G) &= \sigma^2_{A_T} + \sigma^2_{D_T} \end{aligned}$$

Components of Variance

The "additive" and "dominance" components of variance are

$$\begin{aligned}
\sigma^2_{A_M} &= 2p_M p_m[p_M(X_{MM} - X_{Mm}) + p_m(X_{Mm} - X_{mm})]^2 \\
\sigma^2_{A_T} &= 2p_T p_t[p_T(G_{TT} - G_{Tt}) + p_t(G_{Tt} - G_{tt})]^2
\end{aligned}$$

$$\begin{aligned}
\sigma^2_{D_M} &= p^2_M p^2_m(X_{MM} - 2X_{Mm} + X_{mm})^2 \\
\sigma^2_{D_T} &= p^2_T p^2_t(G_{TT} - 2G_{Tt} + G_{tt})^2
\end{aligned}$$

and these lead to the following expression for the covariance of X and G:

$$\mathrm{Cov}(G, X) = \rho_{MT}\sigma_{A_T}\sigma_{A_M} + \rho^2_{MT}\sigma_{D_T}\sigma_{D_M}$$

Correlation of Trait and Marker Variables

$$\text{Cov}(G, X) \;=\; \rho_{MT}\sigma_{A_T}\sigma_{A_M} + \rho^2_{MT}\sigma_{D_T}\sigma_{D_M}$$

If either X or G are purely additive, then their covariance is

$$\text{Cov}(G, X) \;=\; \rho_{MT}\sigma_{A_T}\sigma_{A_M}$$

If both X and G are purely additive, then their correlation is

$$\rho_{GX} \;=\; \rho_{MT}$$

If either X or G are purely non-additive, then their covariance is

$$\text{Cov}(G, X) \;=\; \rho^2_{MT}\sigma_{D_T}\sigma_{D_M}$$

If both X and G are purely non-additive, then their correlation is

$$\rho_{GX} \;=\; \rho^2_{MT}$$

Measured Traits

Suppose $Y = G + E$ where G is the genetic effect of locus **T** and E are all other effects. These other effects are supposed to have mean zero and to be independent of both G and the marker variable X. Then

$$\begin{aligned} \mathcal{E}(Y) &= \mathcal{E}(G) \\ \mathrm{Cov}(X,Y) &= \mathrm{Cov}(X,G) \\ \mathrm{Var}(Y) &= \sigma^2_{A_T} + \sigma^2_{D_T} + V_E \end{aligned}$$

Trait values Y may be regressed on marker variables X. The regression coefficient is

$$\beta_{YX} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \frac{\rho_{MT}\sigma_{A_T}\sigma_{A_M} + \rho^2_{MT}\sigma_{D_T}\sigma_{D_M}}{\sigma^2_{A_M} + \sigma^2_{D_M}}$$

Marker Variable

Variable X may be chosen to be additive, e.g $X_{MM} = 2, X_{Mm} = 1, X_{mm} = 0$ so that $\sigma^2_{A_M} = 2p_M p_m$, $\sigma^2_{D_M} = 0$, and then

$$\beta_{YX} = \rho_{MT} \frac{\sigma_{A_T}}{\sigma_{A_M}}$$

The marker variable can also be made to have a zero additive variance, e.g. $X_{MM} = p_m, X_{Mm} = 0, X_{mm} = p_M$ so that $\sigma^2_{A_M} = 0$, $\sigma^2_{D_M} = p^2_M p^2_m$, and

$$\beta_{YX} = \rho^2_{MT} \frac{\sigma_{D_T}}{\sigma_{D_M}}$$

A significant regression coefficient implies a significant linkage disequilibrium measure ρ_{MT} between marker and disease loci. The signal is expected to be stronger with an additive marker as $\rho_{MT} \geq \rho^2_{MT}$ and it is usual that $\sigma^2_{A_T} \geq \sigma^2_{D_T}$.

Analysis of Variance

Instead of regressing trait on marker, the trait means could be compared among marker classes. The expected trait means follow as

$$\begin{aligned}\mathcal{E}(Y|M_iM_j) &= \sum_{r,s} G_{rs} \Pr(T_rT_s|M_iM_j) \\ &= \sum_{r,s} Grs \Pr(T_rM_i, T_sM_j)/\Pr(M_iM_j)\end{aligned}$$

in general.

For a trait locus with only two alleles, T, t, for marker homozygote MM:

$$\mathcal{E}(Y|MM) = (G_{TT}P_{MT}^2 + 2G_{Tt}P_{MT}P_{Mt} + G_{tt}P_{Mt}^2)/p_M^2$$

Trait Means in Marker Classes

This last expression can be manipulated to show the effects of linkage disequilibrium.

Trait means among the three marker genotype classes:

$$\begin{aligned}
\mathcal{E}(Y|MM) &= \mu_G + 2\rho_{MT}\mathcal{A}/p_M + \rho_{MT}^2\mathcal{D}/p_M^2 \\
\mathcal{E}(Y|Mm) &= \mu_G + \rho_{MT}\mathcal{A}(1/p_M - 1/p_m) - \rho_{MT}^2\mathcal{D}/(p_M p_m) \\
\mathcal{E}(Y|mm) &= \mu_G - 2\rho_{MT}\mathcal{A}/p_m + \rho_{MT}^2\mathcal{D}/p_m^2
\end{aligned}$$

where $\mathcal{A} = \sigma_{A_T}\sqrt{(p_M p_m)}, \mathcal{D} = \sigma_{D_T}(p_M p_m)$, so that an analysis of variance will also test that $\rho_{MT} = 0$ and the test will be affected by both additive and dominance effects at the trait locus.

Dichotomous Traits: Case Only

The case-control approach starts with independent samples of people who are either affected or not affected with a disease and compares marker frequencies between the two groups. The MM marker frequency among cases is

$$\begin{aligned}
\Pr(MM|\text{Case}) &= p_M^2 + \frac{1}{\mu_G}\left[p_M\rho_{MT}\mathcal{A} + \rho_{MT}^2\sigma_{D_T}\mathcal{D}\right] \\
\Pr(Mm|\text{Case}) &= 2p_Mp_m + \frac{1}{\mu_G}\left[(p_m - p_M)\rho_{MT}\mathcal{A} - 2\rho_{MT}^2\mathcal{D}\right] \\
\Pr(mm|\text{Case}) &= p_m^2 + \frac{1}{\mu_G}\left[-p_m\rho_{MT}\mathcal{A} + \rho_{MT}^2\mathcal{D}\right]
\end{aligned}$$

Note that these three probabilities sum to one.

Case Allele Frequencies

Combining the genotypic frequencies to give allele frequencies:

$$\begin{aligned}\Pr(M|\text{Case}) &= p_M + \frac{\rho_{MT}\sigma_{A_T}}{2\mu_G}\sqrt{2p_Mp_m} \\ \Pr(m|\text{Case}) &= p_m - \frac{\rho_{MT}\sigma_{A_T}}{2\mu_G}\sqrt{2p_Mp_m}\end{aligned}$$

and these two sum to one.

The inbreeding coefficient at the marker locus in the case population follows from

$$\Pr(MM|\text{Case}) = \Pr(M|\text{Case})^2 + f_{\text{Case}}\Pr(M|\text{Case})[1-\Pr(M|\text{Case})]$$

or

$$f_{\text{Case}} = \frac{\rho_{MT}^2(2\mu_G\sigma_{D_T} - \sigma_{A_T}^2)}{(\mu_G\sqrt{2p_M/p_m} + \rho_{MT}\sigma_{A_T})(\mu_G\sqrt{2p_m/p_M} - \rho_{MT}\sigma_{A_T})}$$

Case-only HWE Testing

The power of this test depends on nf^2_{Case} which is proportional to ρ^4_{MT} so the power will decrease quickly as ρ_{MT} deceases.

Dichotomous Traits: Case-Control

An argument similar to that above provides the marker genotype frequencies among controls:

$$\begin{aligned}
\Pr(MM|\text{Control}) &= p_M^2 - \frac{1}{1-\mu_G}\left[p_M\rho_{MT}\mathcal{A} + \rho_{MT}^2\mathcal{D}\right] \\
\Pr(Mm|\text{Control}) &= 2p_Mp_m - \frac{1}{1-\mu_G}\left[(p_m - p_M)\rho_{MT}\mathcal{A} - 2\rho_{MT}^2\mathcal{D}\right] \\
\Pr(mm|\text{Control}) &= p_m^2 - \frac{1}{1-\mu_G}\left[-p_m\rho_{MT}\mathcal{A} + \rho_{MT}^2\mathcal{D}\right]
\end{aligned}$$

$$\begin{aligned}
\Pr(M|\text{Control}) &= p_M - \frac{\rho_{MT}\mathcal{A}}{2(1-\mu_G)} \\
\Pr(m|\text{Control}) &= p_m + \frac{\rho_{MT}\mathcal{A}}{2(1-\mu_G)}
\end{aligned}$$

Case-control Test

The simplest case-control test compares marker allele frequencies between the two samples and it is clearly equivalent to testing that $\rho_{MT} = 0$ since

$$\Pr(M|\text{Case}) - \Pr(M|\text{Control}) \propto \rho_{MT}\sigma_{A_T}\sqrt{2p_M p_m}$$

The test is not affected by non-additivity at the disease locus. If the allelic counts for M, m in cases and controls are laid out in a 2×2 table, the contingency-table chi-square test statistic has 1 df. An alternative is to work with the 3×2 table of marker genotype counts in cases and controls and calculate a 2 df chi-square test statistic. This test is affected by both additivity and non-additivity at the disease locus but it is sensitive to errors in genotype calls for rare alleles.

Allelic Case Control Test

Write the marker genotype counts in random samples of cases and controls as

Genotype	MM	Mm	mm	Total
Case counts	r_0	r_1	r_2	R
Control counts	s_0	s_1	s_2	S
Total counts	n_0	n_1	n_2	N

The allelic test statistic uses the allele counts

Observed	M	m	Total
Case counts	$2r_0 + r_1$	$2r_2 + r_1$	$2R$
Control counts	$2s_0 + s_1$	$2s_2 + s_1$	$2S$
Total counts	$2n_0 + n_1$	$2n_2 + n_1$	$2N$

Allelic Case Control Test

A contingency table test for independence of marker allele and disease status compares the observed allelic counts with the products of the marginal totals divided by the overall total:

Expected	M	m	Total
Case counts	$2R(2n_0+n_1)/2N$	$2R(2n_2+n_1)/2N$	$2R$
Control counts	$2S(2n_0+n_1)/2N$	$2S(2n_2+n_1)/2N$	$2S$
Total counts	$2n_0+n_1$	$2n_2+n_1$	$2N$

The test statistic is

$$X_A^2 = \sum \frac{(\text{Obs.} - \text{Exp.})^2}{\text{Exp.}}$$

$$= \frac{2N[N(r_1+2r_2) - R(n_1+2n_2)]^2}{SR[2N(n_1+2n_2) - (n_1+2n_2)^2]}$$

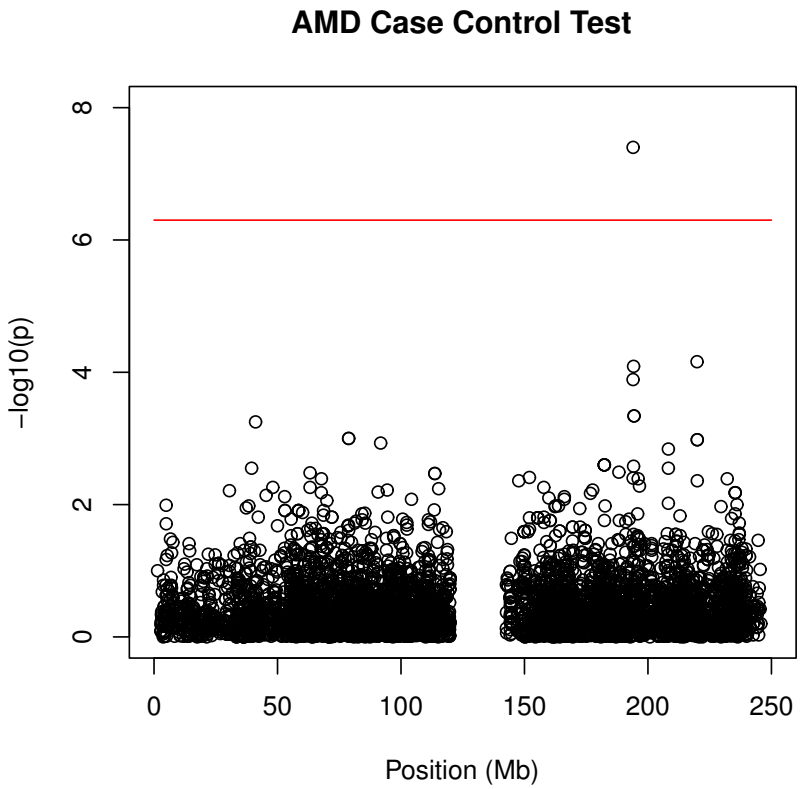
Allelic Case Control Test

Approximating the expectation of X_A^2 by the ratio of the expectations of the numerator and denominator:

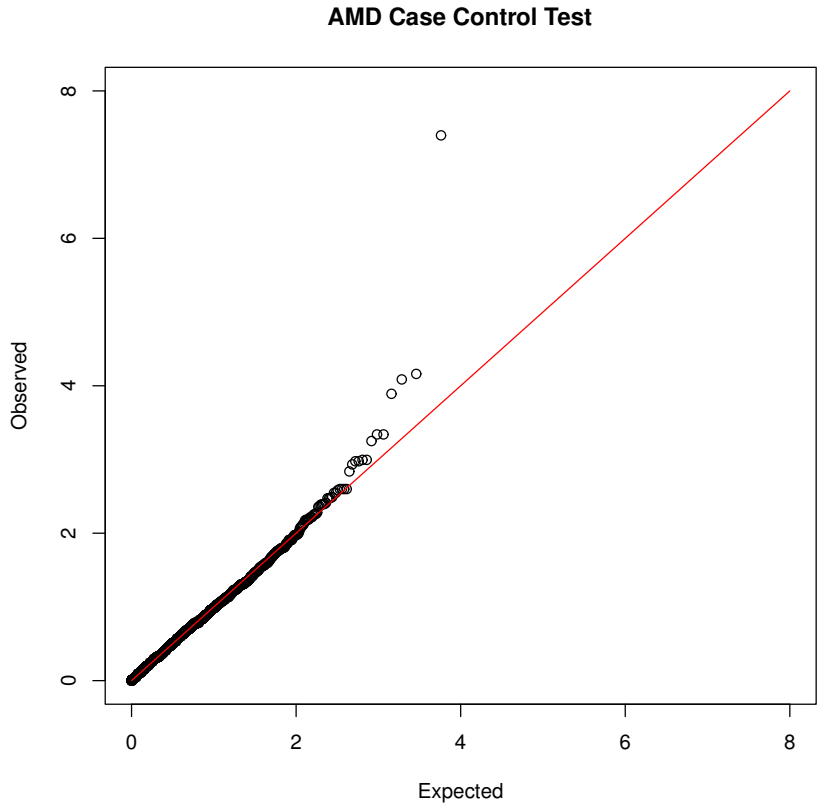
$$\mathcal{E}(X_A^2) \approx (1+f)$$

showing an inflation factor of $(1+f)$ when there is inbreeding. The expected value is 1 when $f = 0$ and the test statistic has a chi-square distribution with 1 df.

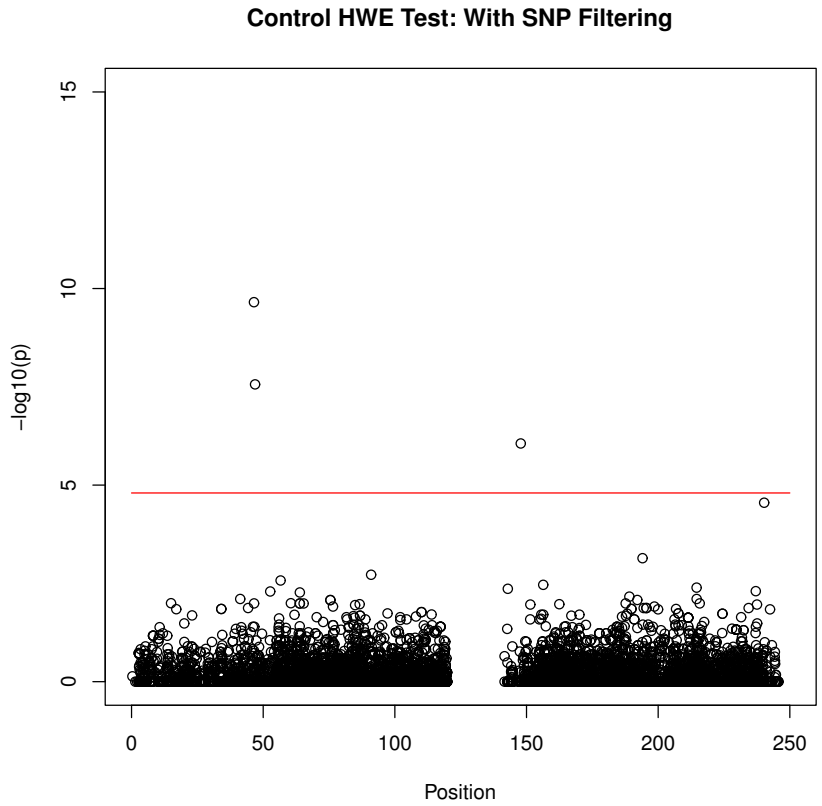
AMD Example: Case-control test statistics on chromosome 1



AMD Example: Case-control test statistics QQ plot



AMD Example: HWE test statistics on chromosome 1



Trend Test

		$i=0$	$i=1$	$i=2$	
Marker Genotype		MM	Mm	mm	Total
Marker Variable		X_0	X_1	X_2	
Case counts	$Y=1$	r_0	r_1	r_2	R
Control counts	$Y=0$	s_0	s_1	s_2	S
Total counts		n_0	n_1	n_2	N

The Armitage trend test is based on a score statistic U:

$$U = \sum_{i=0}^{2} X_i \left(\frac{S}{N} r_i - \frac{R}{N} s_i \right)$$

Trend Test for Additivity

Assuming normality for U, the test statistic

$$X_T^2 = \frac{U^2}{\widehat{\text{Var}}(U)} = \frac{N(N\sum_i r_i X_i - R\sum_i n_i X_i)^2}{SR[N\sum_i n_i X_i^2 - (\sum_i n_i X_i)^2]}$$

is distributed as $\chi^2_{(1)}$ under the hypothesis $H_0 : \rho_{MT} = 0$.

Usual to consider a linear trend test, with $X_0 = 0, X_1 = 1, X_2 = 2$, so that $\sigma^2_{D_M} = 0$ and

$$X_T^2 = \frac{N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{SR[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]}$$

This will provide a test for additive effects at the disease locus.

Case-control vs Trend Tests

The allelic case-control test statistic is

$$X_A^2 = \frac{2N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{SR[N(2n_1 + 4n_2) - (n_1 + 2n_2)^2]}$$

and the linear trend test statistic is

$$X_T^2 = \frac{N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{SR[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]}$$

In both cases, $\sigma^2_{D_M} = 0$ and the test is for linkage disequilibrium ρ_{MT} between trait and marker alleles, and is affected only by additive trait effects.

Unlike the allelic case-control test, the trend statistic has an expected value of 1 even when there are departures from Hardy-Weinberg equilibrium.

Trend Test for Non-Additivity

Setting $X_0 = p_m, X_1 = 0, X_2 = p_M$ gives $\sigma^2_{A_M} = 0$ and a test for non-additive effects. There is not an obvious simplification of the equation for the test statistic.