SISG
2014

SISG Module 8:
Population Genetic Data Analysis

19th Summer Institute in Statistical Genetics

UNIVERSITY of WASHINGTON

(This page left intentionally blank.)

Population Genetic Data Analysis

Summer Institute in Statistical Genetics, July 9-11, 2014

Jérome Goudet: jerome.goudet@unil.ch

and

Bruce Weir: bsweir@uw.edu

Contents

Lectures on these topics by Bruce Weir will alternate with R exercises led by Jérome Goudet. Some of the software is at http://www2.unil.ch/popgen/teaching/SISG14/ .

GENETIC DATA

Genetic Data

"Genes are perpetuated as sequences of nucleic acid, but function by being expressed as proteins." (Lewin, *GENES*).

DNA
(transcription)
↓
RNA
(translation)
↓
Protein
↓
Phenotype

Sources of Data

Phenotype	Mendel's peas
	Blood groups
Protein	Allozymes
	Amino acid sequences
DNA	Restriction sites, RFLPs
	Length variants, VNTRs, STRs
	SNPs
	Nucleotide sequences

Mendel's Data

Dominant Form		Recessive Form	
	Seed characters		
5474	Round	1850	Wrinkled
6022	Yellow	2001	Green
	Plant characters		
705	Grey-brown	224	White
882	Simply inflated	299	Constricted
428	Green	152	Yellow
651	Axial	207	Terminal
787	Long	277	Short

Genetic Data

Human ABO blood groups discovered in 1900.

Elaborate mathematical theories constructed by Sewall Wright, R.A. Fisher, J.B.S. Haldane and others. This theory was challenged by data from new data from electrophoretic methods in the 1960's:

"For many years population genetics was an immensely rich and powerful theory with virtually no suitable facts on which to operate. ... Quite suddenly the situation has changed. The motherlode has been tapped and facts in profusion have been pored into the hoppers of this theory machine. ... The entire relationship between the theory and the facts needs to be reconsidered."

Lewontin (1974)

Electrophoretic Detection

Charge differences among allozymes lead to separation on electrophoretic gels.

Length differences among Restriction Fragment Length Polymorphisms, or Variable Number of Tandem Repeat polymorphisms also detected by electrophoresis.

AmpliType^{TM} Data

PCR based-DNA typing systems for forensic science. The loci involved are all sequence polymorphisms that are detected/delineated by hybridization to allele-specific oligonucleotide probes. Colometric detection used for the hybridized product.

AmpliType^{TM} Data - Caucasian

LDLR		GYPA		HBGG		D7S8		Gc	
a	a	B	B	A	B	A	B	C	C
A	a	b	b	A	B	A	A	A	B
A	a	B	B	A	A	A	A	A	A
a	a	B	B	A	A	A	B	B	B
A	a	B	b	A	B	B	B	A	C
A	A	B	b	A	A	A	B	B	C
a	a	B	b	A	B	A	B	A	C
A	a	b	b	A	A	A	B	A	C
a	a	b	b	A	A	A	A	A	B
A	a	B	B	A	B	A	A	C	C
A	a	B	B	A	B	A	B	C	C
a	a	B	B	A	A	A	B	A	A
A	A	B	b	A	A	A	B	A	B
A	a	B	B	A	A	B	B	C	C
A	A	B	B	A	A	A	A	B	B
A	a	b	b	A	A	A	A	C	C
A	a	B	b	A	B	A	A	B	B
a	a	B	b	A	B	A	A	A	C
A	a	B	B	A	B	A	B	A	C
A	a	b	b	B	B	A	B	B	B

AmpliType^{TM} Data - Afr. Amer.

LDLR		GYPA		HBGG		D7S8		Gc	
a	a	b	b	B	C	A	A	A	B
a	a	B	B	A	A	A	B	B	C
a	a	B	b	A	C	A	A	B	B
A	a	B	b	A	C	A	A	B	C
A	a	B	b	A	C	A	B	A	B
A	a	B	b	A	A	A	B	A	B
a	a	B	b	B	C	B	B	B	C
A	a	B	B	A	A	A	B	B	B
A	a	B	b	A	C	A	B	B	B
a	a	b	b	A	B	A	A	B	B
A	a	B	b	B	C	A	B	B	C
A	a	B	b	A	C	A	B	B	C
a	a	b	b	A	A	A	A	A	C
a	a	b	b	A	C	A	B	B	B
a	a	B	B	A	A	A	A	A	B
A	a	B	B	A	B	A	A	B	B
a	a	B	b	A	B	A	B	B	C
A	a	B	b	B	C	A	A	B	B
A	a	B	B	A	B	A	B	B	B
A	a	B	B	B	C	A	B	B	B

AmpliType^{TM} Data - Hispanic

LDLR		GYPA		HBGG		D7S8		Gc	
a	a	B	b	B	B	B	B	B	C
A	a	B	b	A	A	A	A	A	C
A	A	b	b	A	B	A	A	A	C
A	a	B	B	A	A	A	A	C	C
a	a	B	b	A	B	B	B	A	C
a	a	B	B	B	B	A	B	C	C
A	a	B	B	B	B	A	B	C	C
A	a	B	B	A	B	A	A	A	B
A	a	B	B	A	B	A	B	C	C
A	A	B	b	A	A	A	B	B	B
A	a	b	b	A	B	A	A	A	C
A	a	B	B	A	B	A	B	C	C
a	a	B	b	A	C	A	A	B	B
A	A	b	b	B	B	A	A	C	C
A	a	B	b	A	A	A	A	C	C
a	a	B	B	B	B	A	B	C	C
a	a	B	b	A	B	A	B	B	C
A	a	B	B	B	B	A	A	A	B
A	A	B	B	A	B	B	B	C	C
A	a	B	B	A	B	B	B	A	C

STR markers: CODIS set

(http://www.cstl.nist.gov/biotech/strbase/seq_info.htm)

Locus	Structure	Usual No. of repeats
CSF1PO	$[AGAT]_n$	6–16
TPOX	$[AATG]_n$	5–14
TH01*	$[AATG]_n$	3–14

* "9.3" is $[AATG]_6ATG[AATG]_3$

Length variants detected by capillary electrophoresis.

"CTT" Data - Caucasian

CSF1P0		TPOX		TH01	
11	12	8	11	7	8
11	13	8	8	6	7
11	12	8	11	6	7
10	12	8	8	6	9
11	12	8	12	9	9.3
10	12	9	11	6	7
10	13	8	11	6	6
11	12	8	8	6	9.3
9	10	8	9	7	9.3
11	12	8	8	6	8
11	13	8	11	7	9
11	12	8	11	6	9.3
10	11	8	8	7	9.3
10	10	8	11	7	9.3
9	10	8	8	6	9.3
11	12	9	11	9	9.3
9	11	9	11	9	9.3
11	12	8	8	6	7
10	10	9	11	6	9.3
10	13	8	8	8	9.3

"CTT" Data - Afr. Amer.

CSF1P0		TPOX		TH01	
10	11	8	8	7	8
10	10	8	11	6	7
8	10	8	10	7	8
12	12	7	8	6	8
11	12	9	11	7	9.3
10	10	11	11	7	9
11	12	8	9	7	9
10	12	9	9	7	7
11	12	8	12	7	9
10	11	8	9	7	9
7	11	8	8	6	7
10	11	9	10	8	9
12	12	11	11	9	9
10	12	9	11	6	6
10	10	8	9	7	9.3
11	13	9	12	8	8
10	12	7	10	7	8
11	11	8	9	9	9
8	11	10	11	8	9
10	12	9	9	6	9

"CTT" Data - Hispanic

CSF1P0		TPOX		TH01	
11	12	8	8	10	10
11	11	8	8	6	7
11	12	8	11	7	9.3
10	10	11	11	6	9.3
11	11	8	12	9.3	9.3
12	13	12	12	6	8
11	12	8	10	9.3	9.3
10	11	11	11	6	9
12	12	7	8	6	9.3
10	12	8	11	7	9.3
11	12	8	8	6	7
12	13	11	12	6	7
11	12	8	11	7	9.3
11	14	8	10	6	6
11	12	8	11	7	7
11	12	8	11	6	6
10	11	8	11	6	9
11	12	8	8	6	6
11	11	8	11	6	9
12	12	11	11	7	8

STR Typing Methodology

"Short tandem repeat (STR) typing methods are widely used today for human identity testing applications including forensic DNA analysis. Following multiplex PCR amplification, DNA samples containing the length-variant STR alleles are typically separated by capillary electrophoresis and genotyped by comparison to an allelic ladder supplied with a commercial kit. This article offers a brief perspective on the technologies and issues involved in STR typing."

Butler JM. Short tandem repeat typing technologies used in human identity testing. BioTechniques 43:Sii-Sv (October 2007) doi 10.2144/000112582

Single Nucleotide Polymorphisms (SNPs)

"Single nucleotide polymorphisms (SNPs) are the most frequently occurring genetic variation in the human genome, with the total number of SNPs reported in public SNP databases currently exceeding 9 million. SNPs are important markers in many studies that link sequence variations to phenotypic changes; such studies are expected to advance the understanding of human physiology and elucidate the molecular bases of diseases. For this reason, over the past several years a great deal of effort has been devoted to developing accurate, rapid, and cost-effective technologies for SNP analysis, yielding a large number of distinct approaches. "

Kim S. Misra A. 2007. SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng. 2007;9:289-320.

Current 1000Genomes Data

Tuesday June 24, 2014
The Initial Phase 3 variant list and phased genotypes.

The initial call set from the 1000 Genomes Project Phase 3 analysis is now available on our ftp site in the directory release/20130502/.

These release contains more than 79 million variant sites and includes not just biallelic snps but also indels, deletions, complex short substitutions and other structural variant classes. It is based on data from 2535 individuals from 26 different populations around the world.

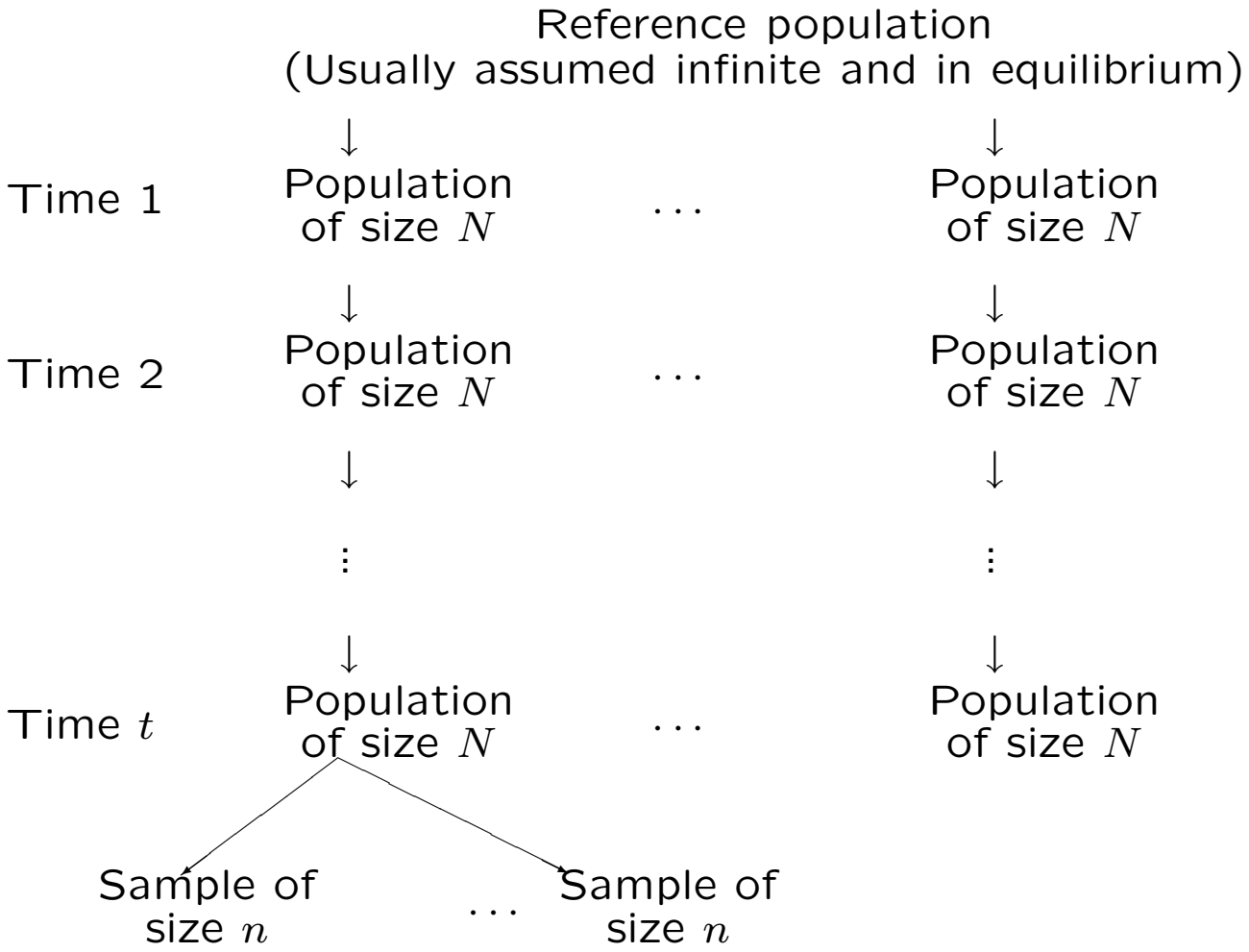
www.1000genomes.org

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.

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

$$\begin{aligned}\Pr(\text{Seed is round and is type AA}) \;=\; &\Pr(\text{Seed is round}|\text{Seed is type AA}) \\ &\times \Pr(\text{Seed is type AA})\end{aligned}$$

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

$$\begin{aligned}\Pr(\text{Seed is round}) = {} & \Pr(\text{Round}|\text{Type AA})\Pr(\text{Type AA}) \\ & + \Pr(\text{Round}|\text{Type Aa})\Pr(\text{Type Aa}) \\ & + \Pr(\text{Round}|\text{Type aa})\Pr(\text{Type aa})\end{aligned}$$

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

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

Bayes' Theorem Example

Suppose G is event that a man has genotype A_1A_2 and H is the event that he transmits allele A_1 to his child. Then $\Pr(H|G) = 0.5$.

Now what is the probability that a man has genotype A_1A_2 given that he transmits allele A_1 to his child?

$$\begin{aligned} \Pr(G|H) &= \frac{\Pr(H|G)\Pr(G)}{\Pr(H)} \\ &= \frac{0.5 \times 2p_1p_2}{p_1} = p_2 \end{aligned}$$

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:

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

Seed Characters

As an experimental check on this last result, and therefore on Mendel's theory, Mendel selfed a round-seeded F_2 plant and noted the F_3 seed shape (observed on the F_2 parental plant).

If all the F_3 seeds are round, the F_2 must have been AA. If some F_3 seed are round and some are wrinkled, the F_2 must have been Aa. Possible to observe many F_3 seeds for an F_2 parental plant, so no doubt that all seeds were round. Data supported theory: one-third of F_2 plants gave only round seeds and so must have had genotype AA.

Plant Characters

Model for stem length is

Genotype	Phenotype
GG	Long
Gg	Long
gg	Short

To check this model it is necessary to grow the F_3 seed to observe the F_3 stem length.

F_2 Plant Character

Mendel grew only 10 F_3 seeds per F_2 parent. If all 10 seeds gave long stems, he concluded they were all GG, and F_2 parent was GG. This could be wrong. The probability of a Gg F_2 plant giving 10 long-stemmed F_3 offspring (GG or Gg), and therefore wrongly declared to be homozygous GG is $(3/4)^{10} = 0.0563$.

Fisher's 1936 Criticism

The probability that a long-stemmed F_2 plant is declared to be homozygous (event V) is

$$\begin{aligned} \Pr(V) &= \Pr(V|U)\Pr(U) + \Pr(V|\bar{U})\Pr(\bar{U}) \\ &= 1 \times (1/3) + 0.0563 \times (2/3) \\ &= 0.3709 \neq 0.3333 \end{aligned}$$

where U is the event that a long-stemmed F_2 is actually homozygous and $\bar{U}$ is the event that it is actually heterozygous.

Fisher claimed Mendel's data closer to the 0.3333 probability appropriate for seed shape than to the correct 0.3709 value.

Weldon's 1902 Doubts

In Biometrika, Weldon said:

"Here are seven determinations of a frequency which is said to obey the law of Chance. Only one determination has a deviation from the hypothetical frequency greater than the probable error of the determination, and one has a deviation sensible equal to the probable error; so that a discrepancy between the hypothesis and the observations which is equal to or greater than the probable error occurs twice out of seven times, and deviations much greater than the probable error do not occur at all. These results then accord so remarkably with Mendel's summary of them that if they were repeated a second time, under similar conditions and on a similar scale, the chance that the agreement between observation and hypothesis would be worse than that actually obtained is about 16 to 1."

Edwards' 1986 Criticism

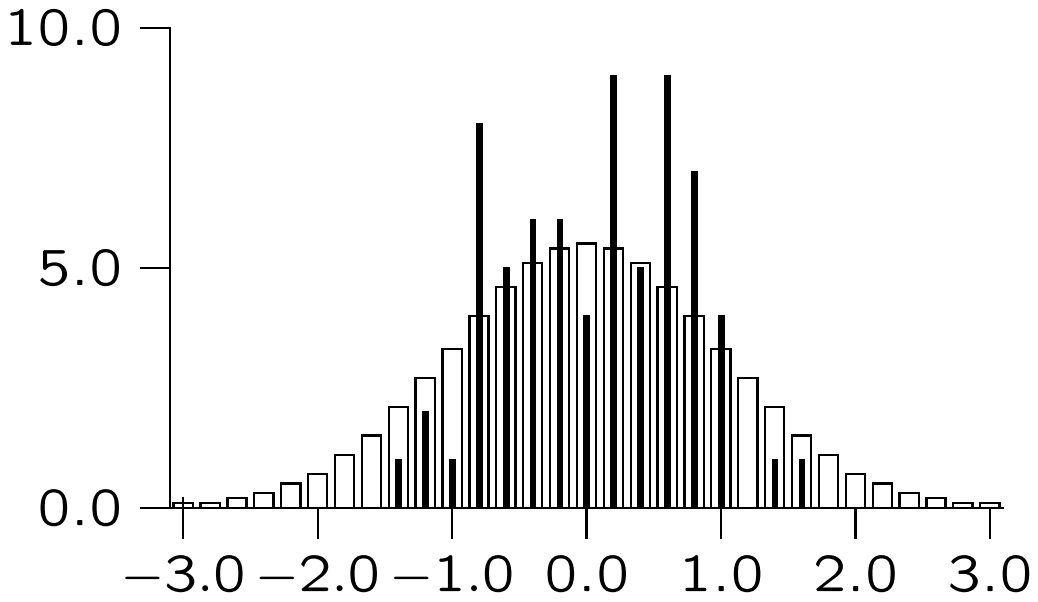
Mendel had 69 comparisons where the expected ratios were correct. Each set of data can be tested with a chi-square test:

		Category 1	Category 2	Total
Observed	(o)	a	n-a	n
Expected	(e)	b	n-b	n

$$
\begin{aligned}
X^2 &= \frac{(a-b)^2}{b} + \frac{[(n-a)-(n-b)]^2}{(n-b)} \\
&= \frac{n(a-b)^2}{b(n-b)}
\end{aligned}
$$

Edwards' Criticism

If the hypothesis giving the expected values is true, the X^2 values follow a chi-square distribution, and the X values follow a normal distribution. Edwards claimed Mendel's values were too small – not as many large values as would be expected by chance.



Novitski's 2004 Response

For six experiments on plant characters, Mendel planned to have 10 F_3 plants for each of 100 dominant phenotype F_2 plants.

With a 2% error rate, the probability that a family of 10 do not all survive, i.e. at least one error or one minus the chance of no error, is $1 - (0.98)^{10} = 0.18$. The expected number of failing families is $6 \times 100 \times 0.18 = 110$. If such a family still showed at least one recessive phenotype it would probably still have been included in the results. Otherwise the family would have been discarded, or augmented and this can result in an under-counting of dominant homozygotes.

This may account for the under-counting of homozygotes noted by Fisher.

Hartl and Fairbanks 2007 Rebuttals

In "Mud sticks: On the alleged falsification of Mendel's data" (Genetics 175:975-979) Hartl and Fairbanks criticized Fisher and Novitski. They did not comment on Edwards.

Against Novitski, they claimed that Mendel did indeed collect data from exactly 10 plants.

Against Fisher, they noted that one experiment was repeated and the combined data agreed with Fisher's expectations.

Pires and Branco, 2010

"A statistical model to explain the Mendel-Fisher controversy." Statistical Science 15:545–565, 2010.

This paper concentrates on the 84 p-values of Mendel's experiments and suggests an unconscious bias in Mendel's work.

Refers to the 2008 book:
Franklin, A., Edwards AWF, Fairbanks DJ, Hartl DL, Seidenfeld T. "Ending the Mendel-Fisher Controversy." University of Pittsburgh Press, Pittsburgh.

Continuing Debate

"This research was carried out in order to verify by simulation Mendel's laws and seek for the clarification, from the author"s point of view, the Mendel–Fisher controversy. ... It was concluded, that Mendel had no reason to manipulate his data in order to make them to coincide with his beliefs. Therefore, in experiment with a single trait, and in experiments with two traits assuming complete dominance, segregation ratios are 3:1; and 9:3:3:1, respectively. Consequently, Mendel's laws, under the conditions as were described are absolutely valid and universal."

REVISTA CIENTIFICA-FACULTAD DE CIENCIAS VETERINARIAS 24:38-46, 2014.

ALLELIC 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

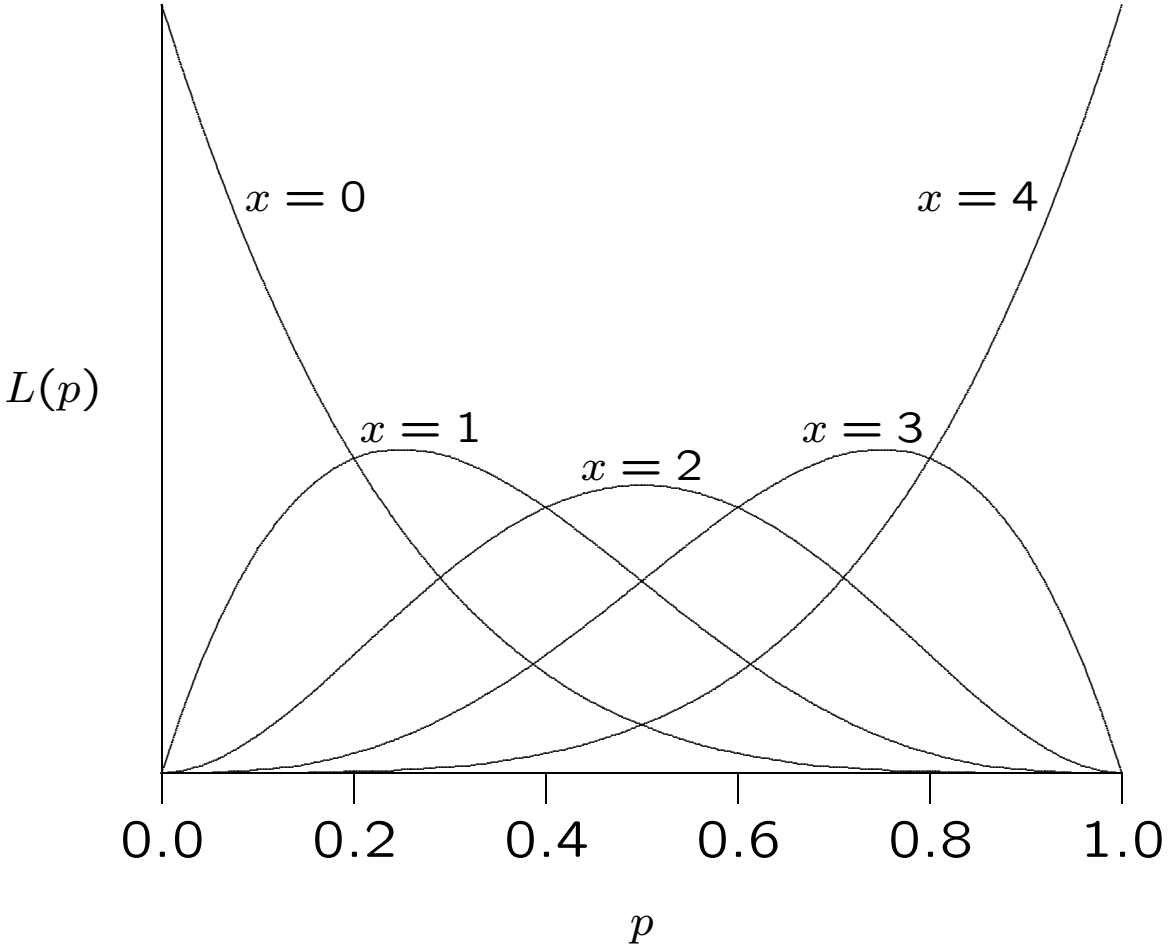
The quantity $\Pr(x|p)$ is the *probability of the data*, x successes in n trials, when each trial has probability p of success.

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

$$n \geq 10,000$$

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 at most as many as 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 at most as many as 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 as many as 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

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

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.

Within-population Variance

The variance of the sample allele frequency can be written as

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

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

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

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:

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

and

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

so

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

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

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

so that, solving 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:

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

Three-allele Case

This formulation preserves the relation between allele and genotype frequencies:

$$
\begin{aligned}
p_1 &= P_{11} + \frac{1}{2}P_{12} + \frac{1}{2}P_{13} \\
p_2 &= P_{22} + \frac{1}{2}P_{12} + \frac{1}{2}P_{23} \\
p_3 &= P_{33} + \frac{1}{2}P_{13} + \frac{1}{2}P_{23}
\end{aligned}
$$

Three-allele Case

The maximum likelihood estimates in the fully-parameterized case are

$$\begin{aligned}
\hat{p}_1 &= \tilde{P}_{11} + \frac{1}{2}\tilde{P}_{12} + \frac{1}{2}\tilde{P}_{13} \\
\hat{p}_2 &= \tilde{P}_{22} + \frac{1}{2}\tilde{P}_{12} + \frac{1}{2}\tilde{P}_{23} \\
\hat{p}_3 &= \tilde{P}_{33} + \frac{1}{2}\tilde{P}_{13} + \frac{1}{2}\tilde{P}_{23}
\end{aligned}$$

and

$$\begin{aligned}
\hat{f}_{12} &= 1 - \tilde{P}_{12}/\tilde{p}_1\tilde{p}_2 \\
\hat{f}_{13} &= 1 - \tilde{P}_{13}/\tilde{p}_1\tilde{p}_3 \\
\hat{f}_{23} &= 1 - \tilde{P}_{23}/\tilde{p}_2\tilde{p}_3
\end{aligned}$$

Three-allele Likelihood

For neutral genes, no reason to expect differences among f's, so an alternative parameterization would be

$$\begin{aligned}
P_{11} &= p_1^2 + f p_1(1-p_1) \\
P_{12} &= 2p_1p_2 - 2f p_1p_2 \\
P_{22} &= p_2^2 + f p_2(1-p_2) \\
P_{13} &= 2p_1p_3 - 2f p_1p_3 \\
P_{23} &= 2p_2p_3 - 2f p_2p_3 \\
P_{33} &= p_3^2 + f p_3(1-p_3)
\end{aligned}$$

Three-allele Likelihood

No explicit analytic expressions for the MLEs of p's and f. Numerical methods needed to maximize the log-likelihood

$$\begin{aligned}
\ln L \;=\; & n_{11}[\ln(p_1) + \ln(p_1 + f - fp_1)] \\
& + n_{12}[\ln(p_1) + \ln(p_2) + \ln(1 - f)] \\
& + n_{22}[\ln(p_2) + \ln(p_2 + f - fp_2)] \\
& + n_{13}[\ln(p_1) + \ln(p_3) + \ln(1 - f)] \\
& + n_{23}[\ln(p_2) + \ln(p_3) + \ln(1 - f)] \\
& + n_{33}[\ln(p_3) + \ln(p_3 + f - fp_3)]
\end{aligned}$$

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:

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

Arrange gamete frequencies as two-way table:

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

For loci LDLR and GYPA in the Hispanic Sample, the 9 observed genotype counts are:

	BB	Bb	bb	Total
AA	$n_{AABB} = 1$	$n_{AABb} = 1$	$n_{AAbb} = 2$	$n_{AA} = 4$
Aa	$n_{AaBB} = 7$	$n_{AaBb} = 2$	$n_{Aabb} = 1$	$n_{Aa} = 10$
aa	$n_{aaBB} = 2$	$n_{aaBb} = 4$	$n_{aabb} = 0$	$n_{aa} = 6$
Total	$n_{BB} = 10$	$n_{Bb} = 7$	$n_{bb} = 3$	$n = 20$

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

	B	b	Total
A	$n_{AB} = x$	$n_{Ab} = 18 - x$	$n_A = 18$
a	$n_{aB} = 27 - x$	$n_{ab} = x - 5$	$n_a = 22$
Total	$n_B = 27$	$n_b = 13$	$2n = 40$

$$18 \geq x \geq 5$$

Example

EM iterative equation:

$$
\begin{aligned}
x' &= 2n_{AABB} + n_{AABb} + n_{AaBB} + n_{AB/ab} \\
&= 2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{2p_{AB}p_{ab}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} n_{AaBb} \\
&= 2 + 1 + 7 + \frac{2x(x-5)}{2x(x-5) + 2(18-x)(27-x)} 2 \\
&= 10 + \frac{2x(x-5)}{x(x-5) + (18-x)(27-x)}
\end{aligned}
$$

Example

A good starting value would assume independence of A and B alleles: $x = 2n * p_A * p_B = (18 \times 27/40) = 12.15$.

Successive iterates are:

Iterate	x value
0	12.1500
1	12.0000
2	11.9310
3	11.8994
4	11.8849
5	11.8783
6	11.8753
7	11.8739
8	11.8733
9	11.8730
10	11.8728

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

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

and we can use "Delta method" to find

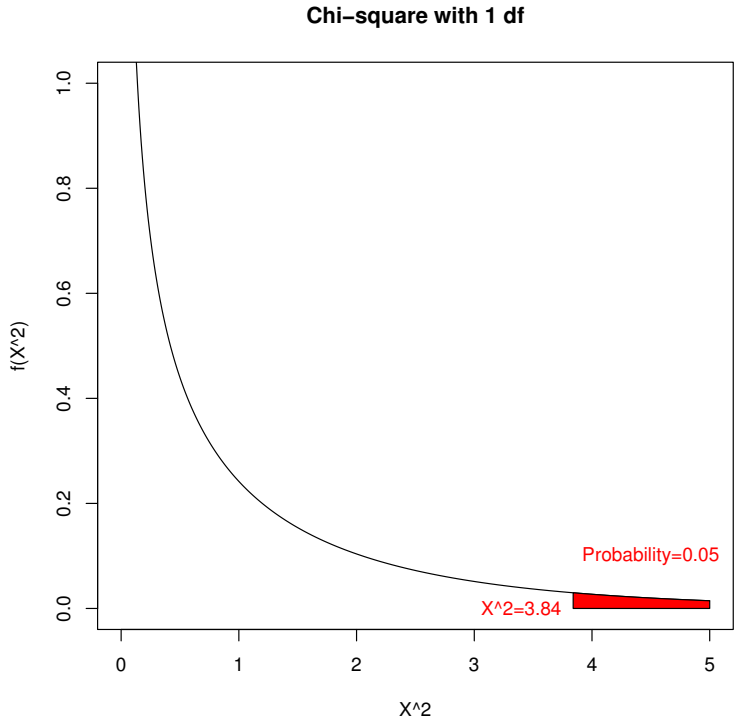
$$\begin{aligned} \mathcal{E}(\hat{f}) &= f \\ \text{Var}(\hat{f}) &\approx \frac{1}{2np_Ap_a}(1-f)[2p_Ap_a(1-f)(1-2f) + f(2-f)] \end{aligned}$$

If $\hat{f}$ is assumed to be normally distributed then, $(\hat{f}-f)/\sqrt{\text{Var}(\hat{f})} \sim N(0,1)$. Since $\text{Var}(\hat{f}) = 1/n$ when $f = 0$:

$$X^2 = \frac{\hat{f}^2}{\text{Var}(\hat{f})} = n\hat{f}^2$$

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:

$$\begin{aligned} n_{AA} &= n(\tilde{p}_A)^2 = 5.625 \\ n_{Aa} &= 2n\tilde{p}_A\tilde{p}_a = 2.750 \\ n_{aa} &= n(\tilde{p}_a)^2 = 0.625 \end{aligned}$$

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	2.750	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

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

(using the null-hypothesis value of the variance in the denominator). Suppose $\widehat{f} > 0$ 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.

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

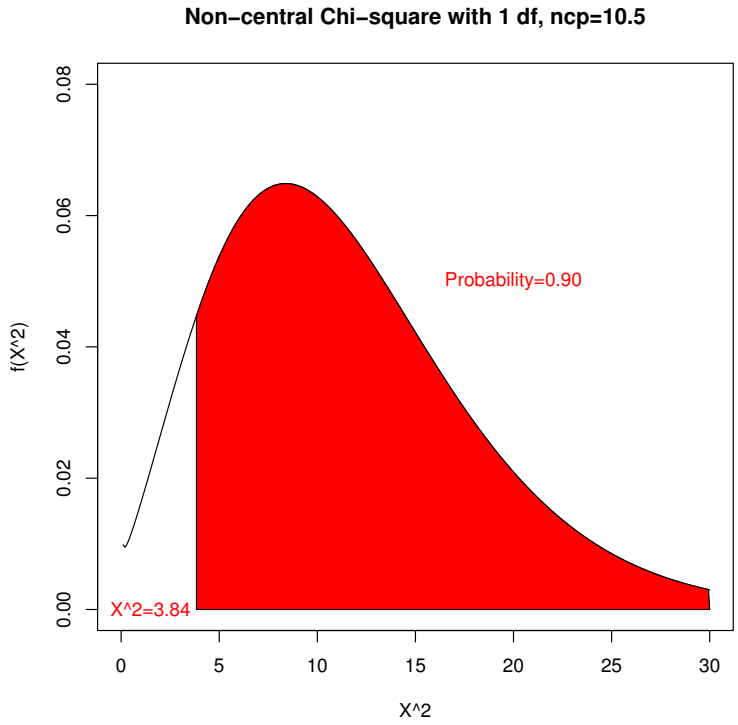
For 5% significance level $-z_{\alpha/2} = 1.96$, and for 90% power $z_\beta = -1.28$ so we need $nf^2 \geq (-1.96 - 1.28)^2 = 10.5$. i.e. n has to be over 100,000 when $f = 0.01$.

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

$$\Pr(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a) = \frac{\frac{n!}{n_{AA}! n_{Aa}! n_{aa}!}(p_A^2)^{n_{AA}}(2p_A p_a)^{n_{Aa}}(p_a^2)^{n_{aa}}}{\frac{(2n)!}{n_A! n_a!}(p_A)^{n_A}(p_a)^{n_a}}$$

$$= \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!}\frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

Reject the Hardy-Weinberg hypothesis if this quantity, the probability of the genotypic array conditional on the allelic array, 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:

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. (As $\tilde{P}_{Aa} = 0, \hat{f} = 1, X^2 = n$.)

In general, 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.

Example

For a sample of size $n = 100$ with minor allele frequency of 0.07, there are 8 sets of possible genotype counts:

			Exact		Chi-square	
n_{AA}	n_{Aa}	n_{aa}	Prob.	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	0.0002*	29.07	0.0000*
89	8	3	0.0051	0.0053*	14.87	0.0001*
88	10	2	0.0602	0.0654	5.38	0.0204*
87	12	1	0.3209	0.3863	0.61	0.4348
86	14	0	0.6136	1.0000	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 0.0053 for an exact test that rejects when $n_{Aa} \leq 8$.

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 rejection region for a nominal value of $\alpha = 0.05$? In other words, we decide to reject HWE for any sample with a p-value of 0.05 or less, and we find the total probability of all such datasets. We would hope that this 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 0.0131 for an exact test that rejects when $n_{Aa} \leq 10$.

n_{AA}	n_{Aa}	n_{aa}	$\Pr(n_{Aa}\|n_a)$	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	.0008
87	10	3	.0123	.0131
86	12	2	.0974	.1105
85	14	1	.3681	.4786
84	16	0	.5215	1.0000

$n_a = 15$ minor alleles

When the minor allele frequency is 0.075, for a nominal 5% significance level, the actual significance level is 0.0085 for an exact test that rejects when $n_{Aa} \leq 9$.

n_{AA}	n_{Aa}	n_{aa}	$\Pr(n_{Aa}\|n_a)$	p −value
92	1	7	.0000	.0000
91	3	6	.0000	.0000
90	5	5	.0000	.0000
89	7	4	.0004	.0004
88	9	3	.0081	.0085
87	11	2	.0776	.0862
86	13	1	.3464	.4326
85	15	0	.5675	1.0000

$n_a = 13$ 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 $n_{Aa} \leq 9$.

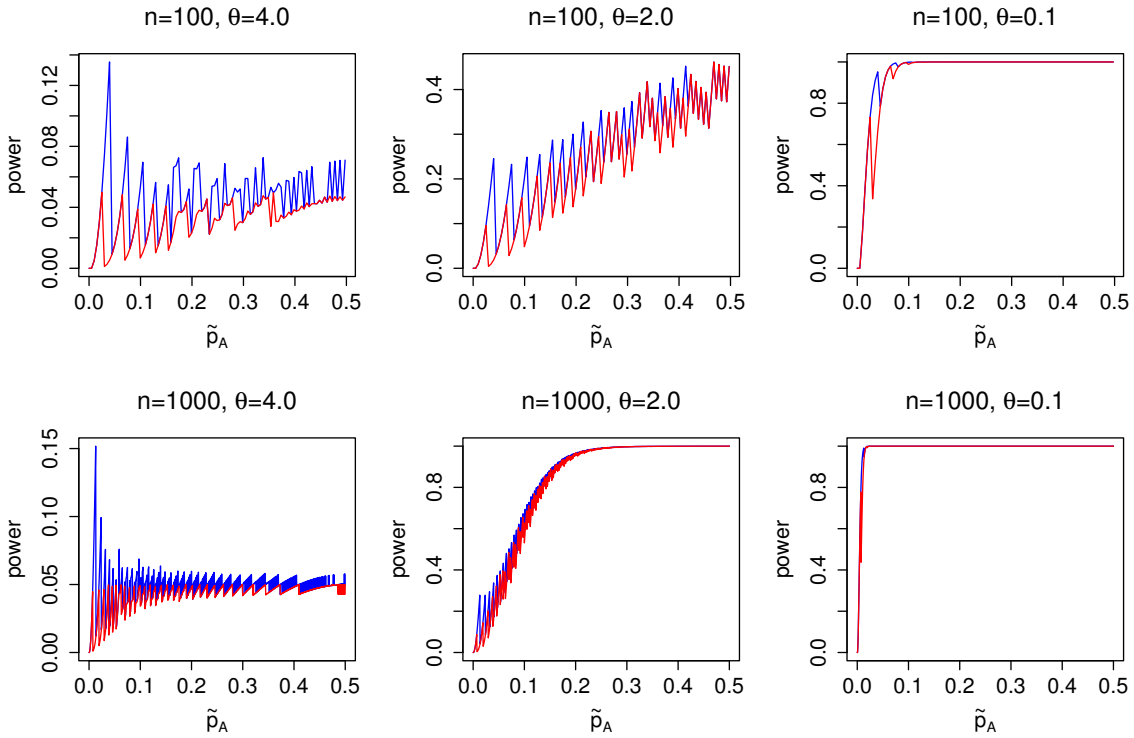
n_{AA}	n_{Aa}	n_{aa}	$\Pr(n_{Aa}\|n_a)$	p −value
93	1	6	.0000	.0000
92	3	5	.0000	.0000
91	5	4	.0001	.0001
90	7	3	.0030	.0031
89	9	2	.0452	.0483
88	11	1	.2923	.3405
87	13	0	.6595	1.0000

$n_a = 12$ 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}	$\Pr(n_{Aa}\|n_a)$	p −value
94	0	6	.0000	.0000
93	2	5	.0000	.0000
92	4	4	.0000	.0000
91	6	3	.0017	.0017
90	8	2	.0327	.0344
89	10	1	.2612	.2956
88	12	0	.7045	1.0000

Rohlfs and Weir, 2008



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 frequencies are $\tilde{p}_A = 0.75, \tilde{p}_a = 0.25$? Note that $1/C = 1/(5!5!0!) + 1/(6!3!1!) + 1/(7!1!2!)$.

			$\Pr(n_{Aa}\|n_A, n)$	
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 shows that the rejection region is $n_{Aa} = 1$. The $\psi = 1$ column shows that the power (the probability $n_{Aa} = 1$ when $\psi = 1$) is 37.4%.

Power Examples

For given values of n, n_a, the rejection region is determined from null hypothesis and the power is determined from the multinomial distribution.

		$\Pr(n_{Aa} \mid n_a = 16)$						
	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

$n_a = 15$ **Probabilities**

		$\Pr(n_{Aa} \mid n_a = 15)$						
	ψ	.250	.500	1.000	2.000	4.000	8.000	16.000
n_{Aa}	f	.622	.389	.150	.000	$-.058$	$-.075$	$-.080$
1		.0338	.0006	.0000	.0000	.0000	.0000	.0000
3		.2269	.0150	.0001	.0000	.0000	.0000	.0000
5		.3871	.1027	.0026	.0000	.0000	.0000	.0000
7		.2592	.2750	.0273	.0004	.0000	.0000	.0000
9		.0801	.3400	.1352	.0081	.0002	.0000	.0000
11		.0120	.2040	.3245	.0776	.0074	.0005	.0000
13		.0008	.0569	.3620	.3464	.1314	.0367	.0094
15		.0000	.0058	.1482	.5674	.8610	.9627	.9905
	Power	.9871	.7333	.1652	.0085	.0002	.0000	.0000

$n_a = 14$ **Probabilities**

		$\Pr(n_{Aa} \mid n_a = 14)$						
	ψ	.250	.500	1.000	2.000	4.000	8.000	16.000
n_{Aa}	f	.613	.378	.143	.000	$-.054$	$-.070$	$-.074$
0		.0062	.0001	.0000	.0000	.0000	.0000	.0000
2		.1256	.0051	.0000	.0000	.0000	.0000	.0000
4		.3610	.0582	.0010	.0000	.0000	.0000	.0000
6		.3422	.2207	.0156	.0002	.0000	.0000	.0000
8		.1375	.3547	.1002	.0051	.0001	.0000	.0000
10		.0255	.2631	.2973	.0602	.0054	.0004	.0000
12		.0021	.0877	.3964	.3209	.1150	.0316	.0081
14		.0001	.0105	.1895	.6136	.8795	.9680	.9919
	Power	.9723	.6387	.1168	.0053	.0001	.0000	.0000

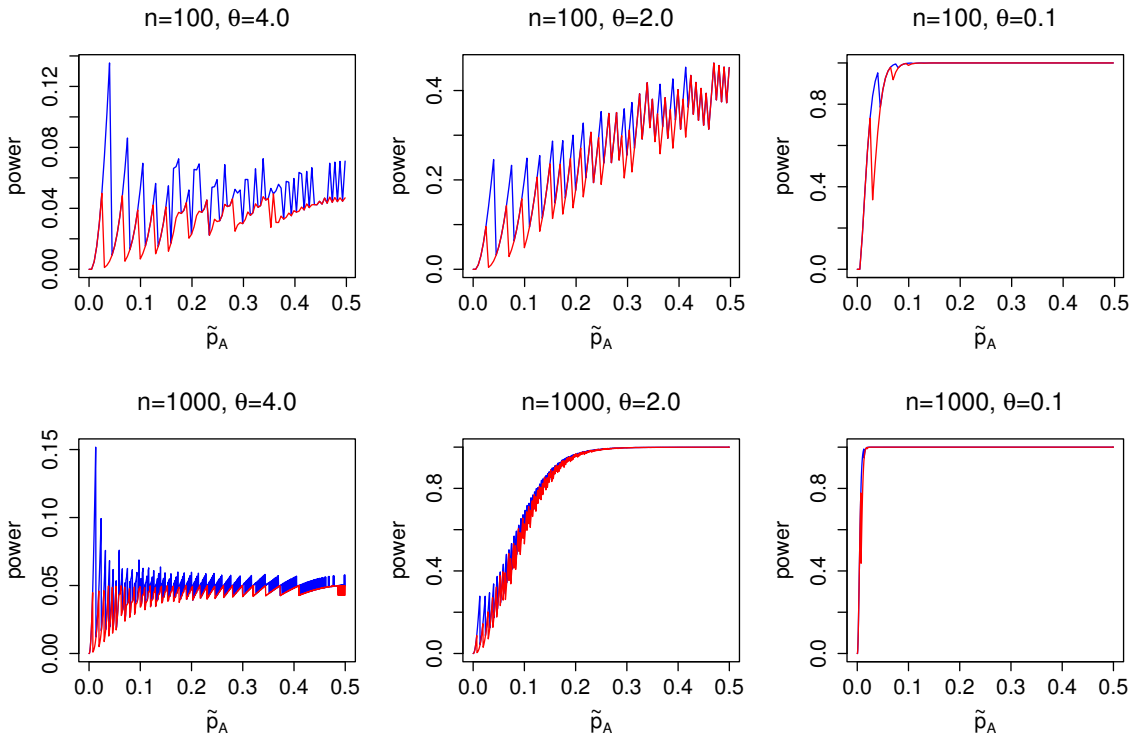
$n_a = 13$ **Probabilities**

		$\Pr(n_{Aa} \mid n_a = 13)$						
	ψ	.250	.500	1.000	2.000	4.000	8.000	16.000
n_{Aa}	f	.603	.366	.136	.000	$-.050$	$-.065$	$-.068$
1		.0479	.0012	.0000	.0000	.0000	.0000	.0000
3		.2786	.0275	.0003	.0000	.0000	.0000	.0000
5		.4004	.1583	.0080	.0001	.0000	.0000	.0000
7		.2169	.3430	.0696	.0030	.0001	.0000	.0000
9		.0508	.3216	.2611	.0452	.0038	.0003	.0000
11		.0051	.1301	.4225	.2923	.0994	.0269	.0069
13		.0002	.0183	.2383	.6595	.8967	.9728	.9931
	Power	.9947	.8516	.3391	.0483	.0039	.0003	.0000

$n_a = 12$ **Probabilities**

		$\Pr(n_{Aa} \mid n_a = 12)$						
	ψ	.250	.500	1.000	2.000	4.000	8.000	16.000
n_{Aa}	f	.592	.353	.128	.000	$-.046$	$-.059$	$-.063$
0		.0095	.0001	.0000	.0000	.0000	.0000	.0000
2		.1674	.0102	.0001	.0000	.0000	.0000	.0000
4		.4053	.0991	.0037	.0000	.0000	.0000	.0000
6		.3108	.3039	.0449	.0017	.0000	.0000	.0000
8		.0947	.3703	.2188	.0326	.0026	.0002	.0000
10		.0118	.1852	.4376	.2612	.0846	.0226	.0058
12		.0005	.0312	.2950	.7044	.9127	.9772	.9942
	Power	.9877	.7836	.2674	.0344	.0027	.0002	.0000

Rohlfs and Weir, 2008



Permutation Test

For large sample sizes and many alleles per locus, there are too many genotypic arrays for a complete enumeration and a determination of which are the least probable 5% arrays.

A large number of the possible arrays is generated by permuting the alleles among genotypes, and calculating the proportion of these permuted genotypic arrays that have a smaller conditional probability than the original data. If this proportion is small, the Hardy-Weinberg hypothesis is rejected.

Permutation Test

Mark a set of five index cards to represent five genotypes:

Card 1:	A	A
Card 2:	A	A
Card 3:	A	A
Card 4:	a	a
Card 5:	a	a

Tear the cards in half to give a deck of 10 cards, each with one allele. Shuffle the deck and deal into 5 pairs, to give five genotypes.

Permutation Test

The permuted set of genotypes fall into one of four types:

AA	Aa	aa	Number of times
3	0	2	
2	2	1	
1	4	0	

Permutation Test

Find the theoretical values for the proportions of each of the three types, from the expression:

$$\frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} \times \frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

AA	Aa	aa	Conditional Probability
3	0	2	
2	2	1	
1	4	0	

These should match the proportions found by repeating shufflings of the deck of 10 allele cards.

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.

Combining p-values

Methods for combining p-values rest on p having a uniform distribution when the hypothesis is true. Suppose the test statistic X has the continuous distribution $f(x)$ when the hypothesis is true. The p-value when $X = x$ is

$$p \;=\; \int_x^\infty f(y)dy$$

(A one-tailed test when large values of X cause rejection.) The distribution function $F(x)$ for X is

$$F(x) \;=\; \int_{-\infty}^x f(y)dy = 1 - p$$

The density function for $U = F(x)$ is

$$f(u) \;=\; f(x)\frac{dx}{du} = f(x)/\frac{du}{dx} = f(x)/f(x) = 1$$

Therefore U is uniform on [0,1], and so is p.

Fisher's Method

If U has a uniform distribution, transform to

$$\begin{aligned} V &= -2\ln(U) \\ U &= e^{-V/2} \\ f(v) &= f(u)\frac{du}{dv} \\ &= (1/2)e^{-v/2} \end{aligned}$$

and this is the density function of a chi-square distribution with 2 d.f.

Fisher's Method

If a hypothesis is true, and the significance level is p, then $X^2 = -2\ln(p)$ is distributed as $\chi^2_{(2)}$. (If $p = 0.05$ then $X^2 = 5.99$.) Therefore

$$t = -2\sum_{i=1}^{L} \ln p_i = -2\ln(\prod_{i=1}^{L} p_i)$$

has a chi-square distribution with $2L$ df when all L hypotheses are true.

Therefore, the p-value for the hypothesis H_T that all H_i are true is the probability of a $\chi^2_{(2L)}$ variable being greater or equal to the observed value t.

Example

HW-test p-values for previous data.

Locus	Cauc.	Afr.Am.	Hisp.	$-2\sum \ln p$
LDLR	1.00	0.14	0.69	4.67 ns
GYPA	0.22	1.00	0.42	4.76 ns
HBGG	1.00	0.44	0.21	4.76 ns
D7S8	1.00	0.64	0.69	1.63 ns
Gc	0.01	1.00	0.59	10.27 ns
$-2\sum \ln p$	12.24 ns	6.47 ns	7.40 ns	26.10 ns

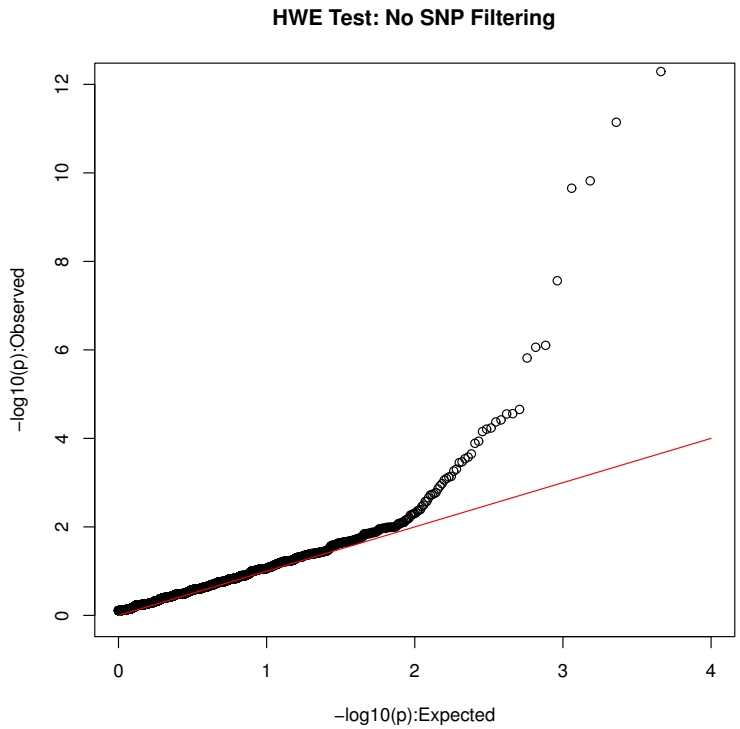
Do not reject overall hypothesis of no departures from HW over populations for each locus, or over loci for each population, or over all locus-population pairs.

QQ-Plots

An alternative 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.

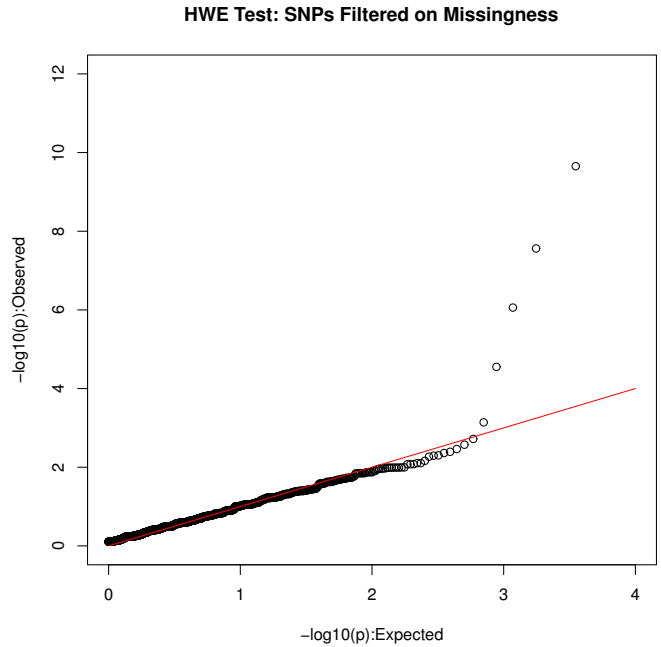
For a set of n tests, we would expect to see p values at $1/(n+1), 2/n, \ldots, n/(n+1), 1$. We plot the observed p-values against these expected values: the smallest against $1/(n+1)$ and the largest against 1. It is more convenient to transform to $-\log_{10}(p)$ to accentuate the extremely small p 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. 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.

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

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

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

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

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, $X_{AB}^2 \sim \chi_{(1)}^2$.

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 loci LDLR and GYPA in the Hispanic Sample:

	BB	Bb	bb	Total
AA	1	1	2	4
Aa	7	2	1	10
aa	2	4	0	6
Total	10	7	3	20

$$
\begin{aligned}
n_{AB} &= 2 \times 1 + 1 + 7 + \frac{1}{2}(2) = 11 \\
n_A = 18 &\quad , \quad \tilde{p}_A = 0.450 \\
n_B = 27 &\quad , \quad \tilde{p}_B = 0.675 \\
\widehat{f}_A &= \frac{4}{20} - \left(\frac{18}{40}\right)^2 = -\frac{1}{400} = -0.0025 \\
\widehat{f}_B &= \frac{10}{20} - \left(\frac{27}{40}\right)^2 = \frac{71}{1600} = 0.0444
\end{aligned}
$$

Example

$$\hat{\Delta}_{AB} = \frac{11}{20} - 2\frac{18}{4}\frac{27}{40} = -\frac{46}{800} = -0.0575$$

$$X^2_{AB} = \frac{20(-0.0575)^2}{(0.450)(0.550)(0.9975)(0.675)(0.325)(1.0444)} = 1.11$$

Previous work on EM algorithm estimated p_{AB} as 11.88/40=0.2970 so

$$\hat{D}_{AB} = 0.2970 - (0.450)(0.675) = -0.008$$

$$X^2_{AB} = \frac{40(-0.0068)^2}{(0.450)(0.550)(0.675)(0.325)} = 0.03$$

LD vs Composite LD Estimates

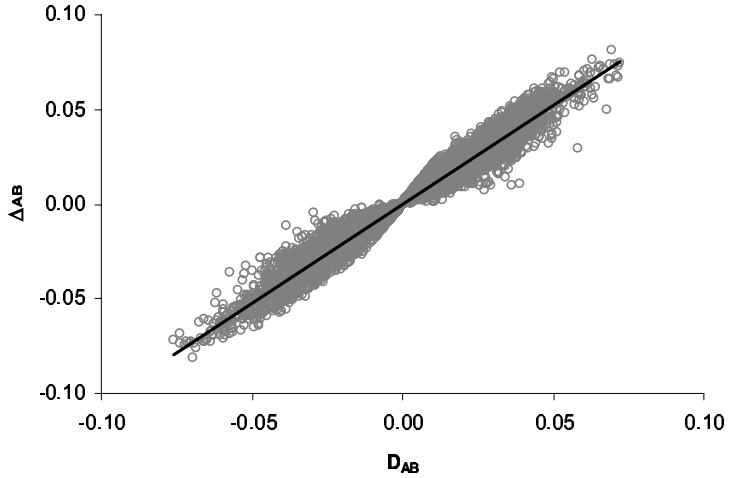


Figure 6: Linkage disequilibrium estimates

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

LD vs Composite LD Tests

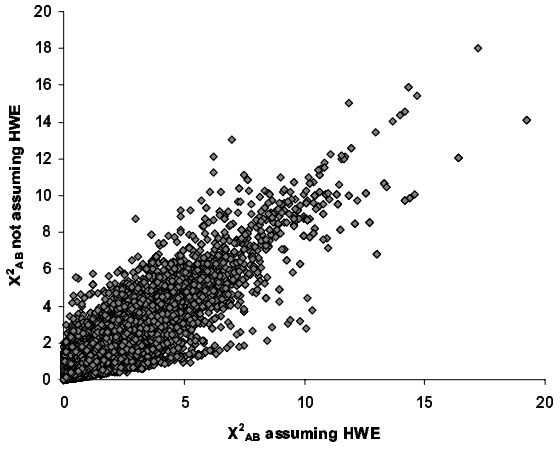


Figure 5: Linkage disequilibrium test statistics (Caucasian)

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

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

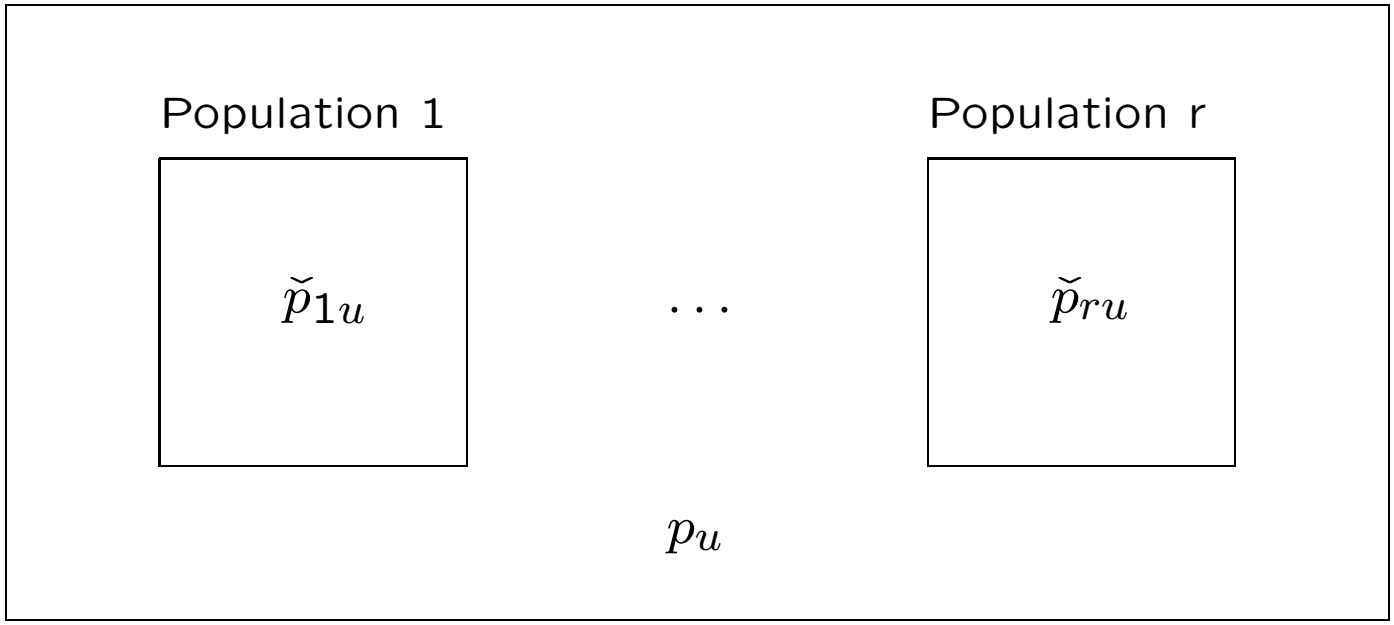
Possible to approach these data from purely statistical viewpoint.

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.

Frequencies of Allele A_u

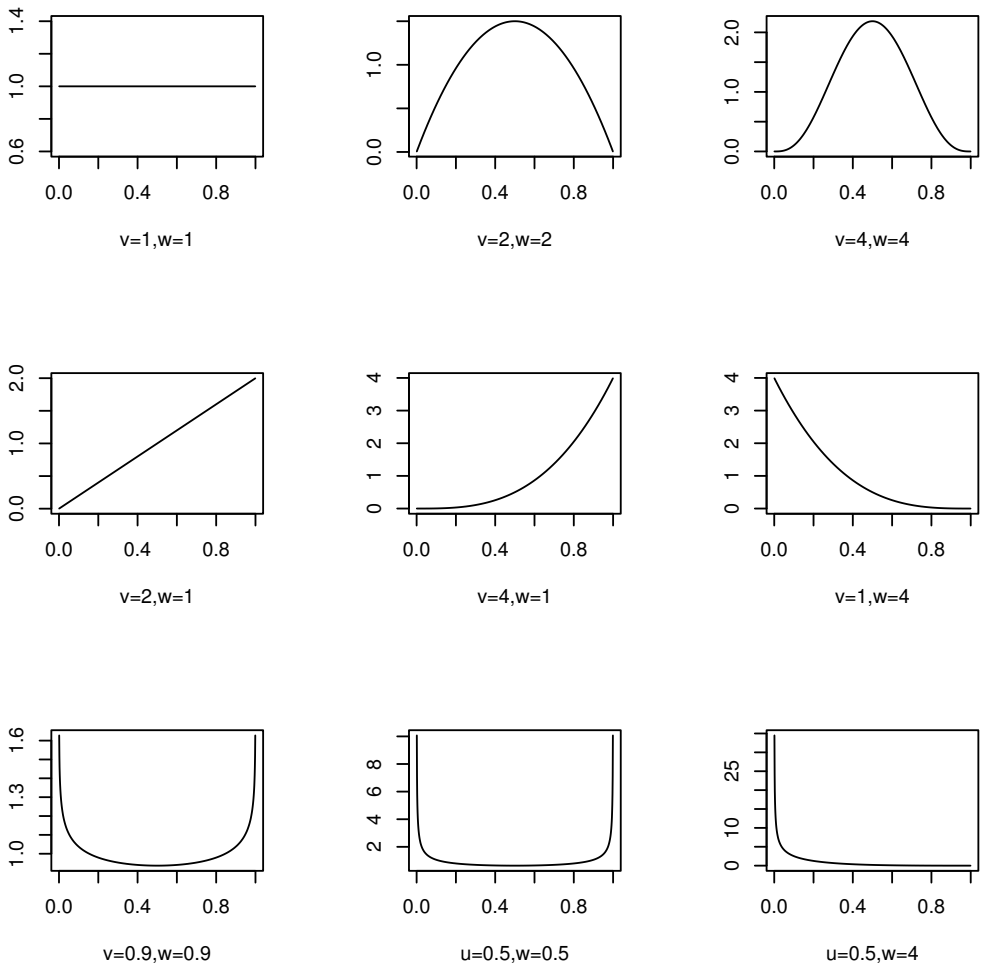


Among samples from a population: $\tilde{p}_{iu} \sim$ Binomial$(n, \breve{p}_{iu})$.

Among replicates of a population: $\breve{p}_{iu} \sim$ Beta $\left(\frac{(1-\theta_i)p_u}{\theta_i}, \frac{(1-\theta_i)(1-p_u)}{\theta_i}\right)$.

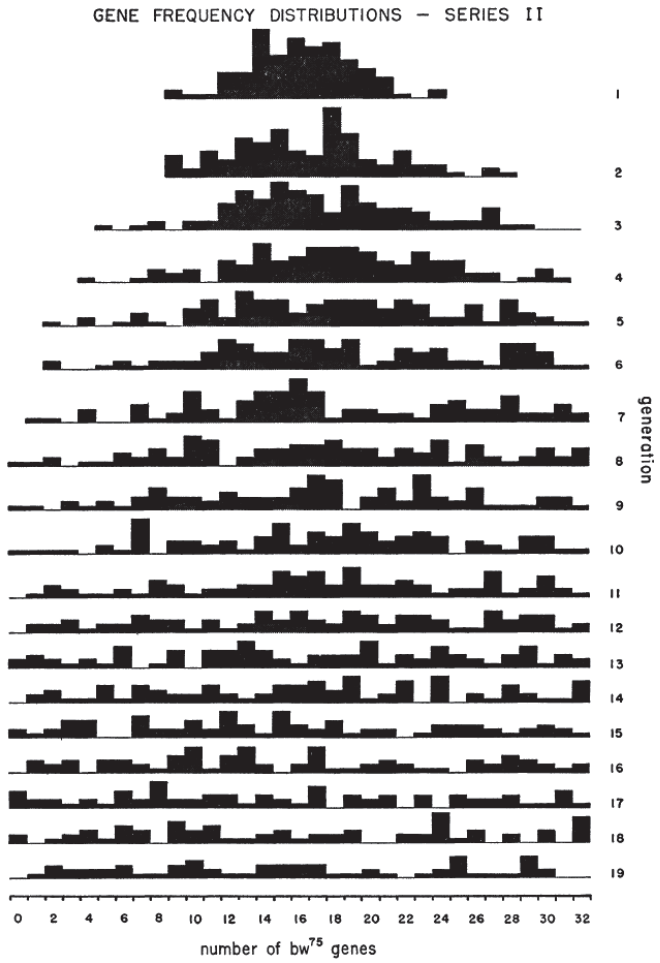
Beta distribution: Theoretical

The beta distribution can take a variety of shapes.



Beta distribution: Experimental

The beta distribution is suggested by a *Drosophila* cage experiment by P. Buri (Evolution 10:367, 1956).



Model Details

It is useful to attach indicator variables to each allele. For the jth allele sampled from the ith population

$$x_{ij} = \begin{cases} 1 & \text{allele is of type } A_u \\ 0 & \text{otherwise} \end{cases}$$

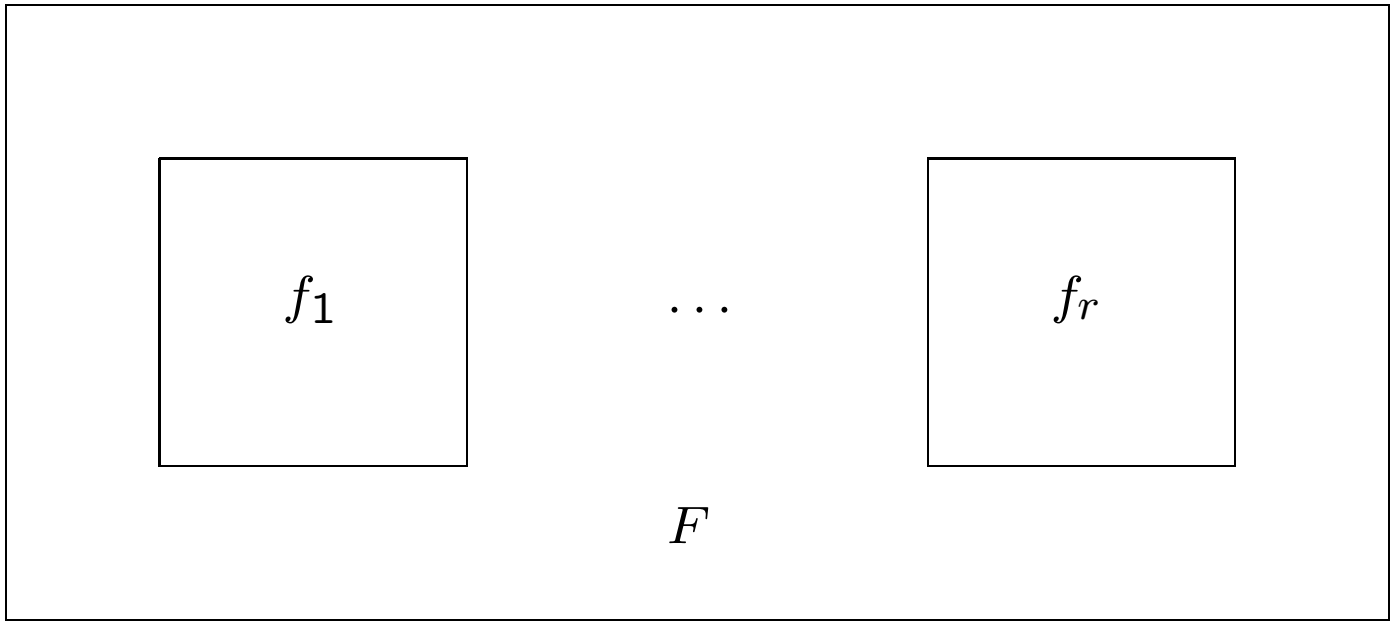
The model specifies expectations for these indicators. Over samples from one population:

$$\begin{aligned} \mathcal{E}(x_{ij}) &= \breve{p}_{iu} \\ \mathcal{E}(x_{ij}x_{ij'}) &= \breve{p}_{iu}^2 \;\; j \neq j', \text{ large } n_i \end{aligned}$$

Over samples from each population *and* over replicates of each population:

$$\begin{aligned} \mathcal{E}(x_{ij}) &= p_u \\ \mathcal{E}(x_{ij}x_{ij'}) &= \theta_i p_u + (1-\theta_i)p_u^2, \;\; j \neq j' \\ \mathcal{E}(x_{ij}x_{i'j'}) &= \theta_{ii'} p_u + (1-\theta_{ii'})p_u^2, \;\; i \neq i' \end{aligned}$$

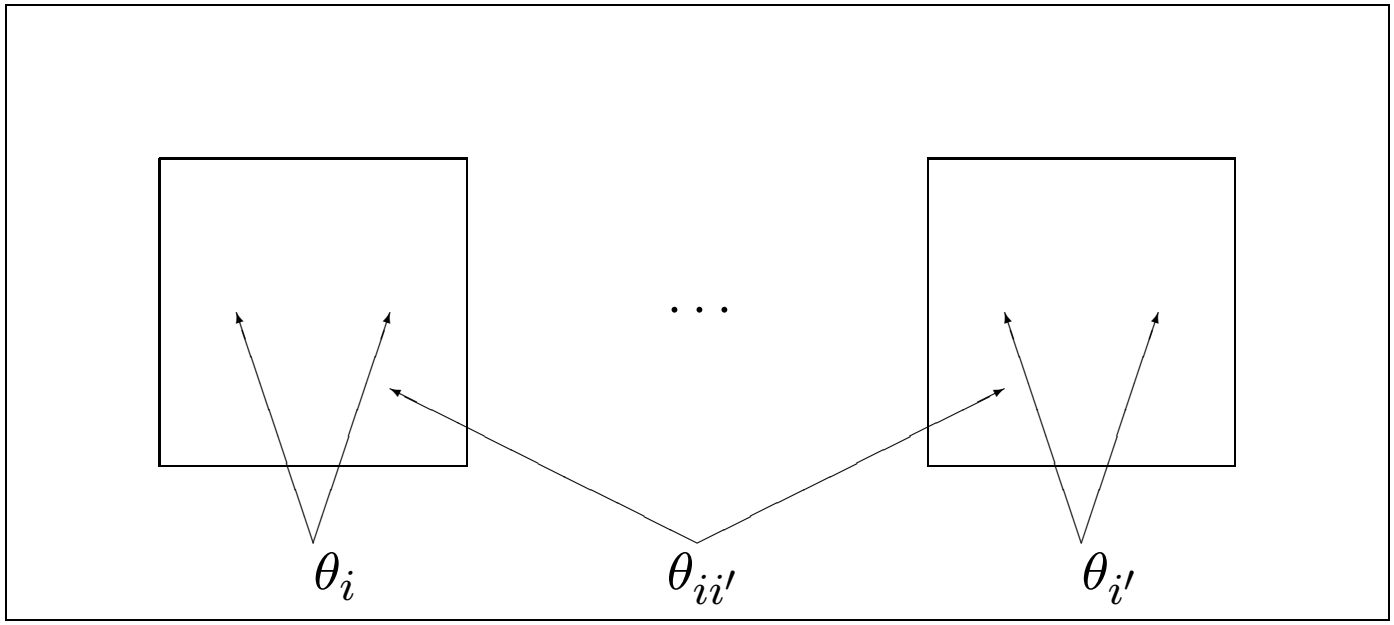
Inbreeding Coefficients



Within a population: $\check{P}_{uu_i} = \check{p}_{iu}^2 + f_i \check{p}_{iu}(1 - \check{p}_{iu})$.

Over replicates of a population: $\mathcal{E}(\check{P}_{uu_i}) = p_u^2 + F_i p_u(1 - p_u)$.

Coancestry Coefficients



For a population i, using average allele frequencies:
$\mathcal{E}(\check{P}_{uu_i}) = p_u^2 + \theta_i p_u(1-p_u)$, if $F_i = \theta_i$.

For two populations i, i': $\mathcal{E}(\check{P}_{u,u_{ii'}}) = p_u^2 + \theta_{ii'} p_u(1-p_u)$.

Genotypic Frequencies

Within a population:

$$\breve{P}_{uu_i} = \breve{p}_{iu}^2 + f_i \breve{p}_{iu}(1 - \breve{p}_{iu})$$

Take expected values over replicates of a population:

$$\begin{aligned}
\mathcal{E}(\breve{p}_{iu}) &= p_u \\
\mathcal{E}(\breve{p}_{iu}^2) &= p_u^2 + \theta_i p_u(1 - p_u) \\
\mathcal{E}(\breve{P}_{uu_i}) &= [p_u^2 + \theta_i p_u(1 - p_u)] + f_i[p_u - p_u^2 - \theta_i p_u(1 - p_u)] \\
&= p_u^2 + [\theta_i + f_i(1 - \theta_i)]p_u(1 - p_u) \\
&= p_u^2 + F_i p_u(1 - p_u)
\end{aligned}$$

since $f_i = (F_i - \theta_i)/(1 - \theta_i)$ or $F_{IS} = (F_{IT} - F_{ST})/(1 - F_{ST})$ and this is zero for HWE within a population.

Sample Allele Frequencies: Weir & Hill, 2002

The sample frequency for allele A_u for population i can be written as an average of indicator variables:

$$\tilde{p}_{iu} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

and this leads to variances and covariances of frequencies over samples from populations and over replicates of populations:

$$\begin{aligned}
\mathrm{Var}(\tilde{p}_{iu}) &= p_u(1-p_u)\left(\theta_i + \frac{1-\theta_i}{n_i}\right) \\
\mathrm{Cov}(\tilde{p}_{iu}, \tilde{p}_{iu'}) &= -p_u p_{u'}\left(\theta_i + \frac{1-\theta_i}{n_i}\right), \quad u \neq u' \\
\mathrm{Cov}(\tilde{p}_{iu}, \tilde{p}_{i'u}) &= p_u(1-p_u)\theta_{ii'}, \quad i \neq i' \\
\mathrm{Cov}(\tilde{p}_{iu}, \tilde{p}_{i'u'}) &= -p_u p_{u'}\theta_{ii'}, \quad u \neq u', i \neq i'
\end{aligned}$$

Weir & Cockerham 1984 Restricted Model

If populations have equal evolutionary histories ($\theta_i = \theta$, all i) and are independent ($\theta_{ii'} = 0$, all $i \neq i'$)

$$\begin{aligned}
\text{Var}(\tilde{p}_{iu}) &= p_u(1-p_u)\left(\theta + \frac{1-\theta}{n_i}\right) \\
\text{Cov}(\tilde{p}_{iu}, \tilde{p}_{iu'}) &= -p_u p_{u'}\left(\theta + \frac{1-\theta}{n_i}\right), \quad u \neq u' \\
\text{Cov}(\tilde{p}_{iu}, \tilde{p}_{i'u}) &= 0, \quad i \neq i' \\
\text{Cov}(\tilde{p}_{iu}, \tilde{p}_{i'u'}) &= 0, \quad u \neq u', i \neq i'
\end{aligned}$$

The overall allele frequencies were weighted by sample sizes

$$\bar{p}_A = \frac{1}{\sum_i n_i} \sum_i n_i \tilde{p}_{iu}$$

If $\theta = 0$, these weighted means have minimum variance.

Weir & Cockerham 1984 Restricted Model

Under this model, the sampled populations can be regarded as evolutionary realizations of the process that led to any one of them. The sampled populations are equivalent to the population replicates that have been discussed in this treatment.

Weir & Cockerham 1984 Restricted Model

Two mean squares were constructed for each allele:

$$\begin{aligned}
\mathrm{MSB}_u &= \frac{1}{r-1}\sum_{i=1}^{r} n_i(\tilde{p}_{iu}-\bar{p}_u)^2 \\
\mathrm{MSW}_u &= \frac{1}{\sum_i (n_i-1)}\sum_i n_i\tilde{p}_{iu}(1-\tilde{p}_{iu})
\end{aligned}$$

These have expected values

$$\begin{aligned}
\mathcal{E}(\mathrm{MSB}_u) &= p_u(1-p_u)[(1-\theta)+n_c\theta] \\
\mathcal{E}(\mathrm{MSW}_u) &= p_u(1-p_u)(1-\theta)
\end{aligned}$$

where $n_c = (\sum_i n_i - \sum_i n_i^2/\sum_i n_i)/(r-1)$. The Weir & Cockeram estimator of θ (or F_{ST}) is

$$\hat{\theta} = \frac{\sum_u(\mathrm{MSB}_u - \mathrm{MSW}_u)}{\mathrm{MSB}_u + (n_c-1)\mathrm{MSW}_u}$$

Unweighted Model

Bhatia et al., Genome Research 23:1514-1521, 2013, returned to early work by Nei. They used heterozygosities instead of mean squares:

$$\tilde{H}_i = \frac{n_i}{n_i - 1} \sum_u \tilde{p}_{iu}(1 - \tilde{p}_{iu}) = \frac{n_i}{n_i - 1}(1 - \sum_u \tilde{p}_{iu}^2)$$

$$\tilde{H}_{ii'} = 1 - \sum_u \tilde{p}_{iu}\tilde{p}_{i'u}$$

Unweighted averages over populations are

$$\tilde{H}_W = \frac{1}{r}\sum_i \tilde{H}_i \quad , \quad \theta_W = \frac{1}{r}\sum_i \theta_i$$

$$\tilde{H}_B = \frac{1}{r(r-1)} \sum_{i \neq i'} \tilde{H}_{ii'} \quad , \quad \theta_B = \frac{1}{r(r-1)} \sum_{i \neq i'} \theta_{ii'}$$

Unweighted Model Expectations

Under the general model of Weir & Hill, these within- and between-population heterozygosities have expectations

$$\begin{aligned} \mathcal{E}(\tilde{H}_i) &= H(1-\theta_i) \\ \mathcal{E}(\tilde{H}_{ii'}) &= H(1-\theta_{ii'}) \end{aligned}$$

$$\begin{aligned} \mathcal{E}(\tilde{H}_W) &= H(1-\theta_W) \\ \mathcal{E}(\tilde{H}_B) &= H(1-\theta_B) \end{aligned}$$

where $H = \sum_u p_u(1-p_u) = 1 - \sum_u p_u^2$.

These equations suggest estimation by the method of moments.

Unweighted Model Estimates

Can only estimate θ_i and $\theta_{ii'}$ "relative to" θ_B:

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

$$\hat{\beta}_W = 1 - \frac{\tilde{H}_W}{\tilde{H}_B} \quad , \quad \mathcal{E}(\hat{\beta}_W) = \frac{\theta_W - \theta_B}{1 - \theta_B}$$

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

Estimation of θ_B is not possible unless the whole set of populations is replicated. Different values of β_i or $\beta_{ii'}$ imply different values of θ_i or $\theta_{ii'}$.

Weir & Cockerham Estimator

Under the Weir and Hill model, the Weir and Cockerham estimator has expectation

$$\mathcal{E}(\hat{\theta}_{\mathrm{WC}}) = \frac{\theta_W^c - \theta_B^c + Q}{1 - \theta_B^c + Q} \quad \text{instead of} \quad \frac{\theta_W - \theta_B}{1 - \theta_B}$$

where

$$\theta_W^c = \frac{\sum_i n_i^c \theta_i}{\sum_i n_i^c} \quad , \quad \theta_B^c = \frac{\sum_{i \neq i'} n_i n_{i'} \theta_{ii'}}{\sum_{i \neq i'} n_i n_{i'}}$$

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

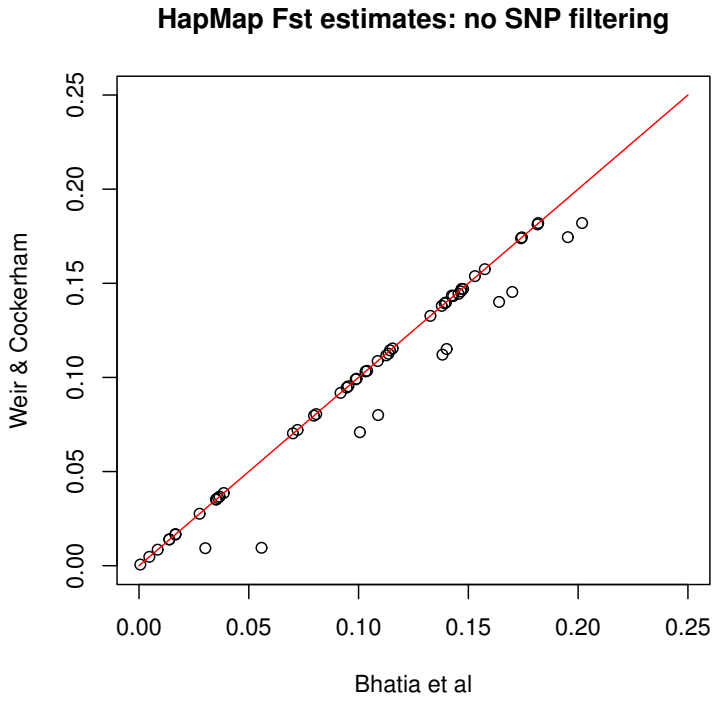
$$Q = \frac{1}{(r-1)n_c} \sum_i \left(\frac{n_i}{\bar{n}} - 1 \right) \theta_i$$

If the Weir and Cockerham model holds ($\theta_i = \theta$), or if $n_i = n$, or if n_c is large, then $Q = 0$.

HapMap III Data

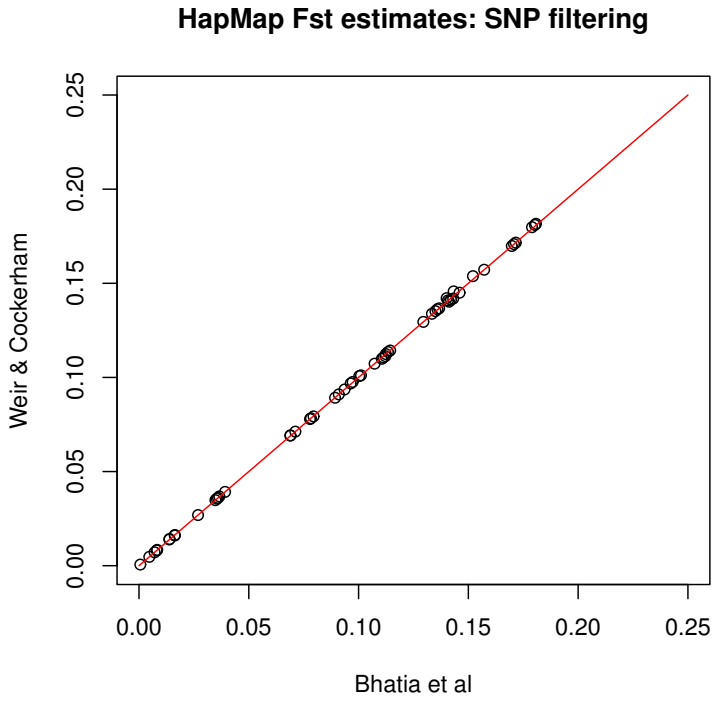
Code	Population Description	Sample size	Number of SNPs typed
ASW	African ancestry in Southwest USA	142	61566
CEU	Utah residents with Northern and Western European ancestry from CEPH collection	324	56218
CHB	Han Chinese in Beijing, China	160	51946
CHD	Chinese in Metropolitan Denver, Colorado	140	50517
GIH	Gujarati Indians in Houston, Texas	166	55710
JPT	Japanese in Tokyo, Japan	168	53020
LWK	Luhya in Webuye, Kenya	166	60026
MXL	Mexican ancestry in Los Angeles, California	142	57937
MKK	Maasai in Kinyawa, Kenya	342	61057
TSI	Toscani in Italia	154	55881
YRI	Yoruba in Ibadan, Nigeria	326	58559

Weir & Cockerham vs Unweighted Estimator



F_{ST} estimates for HapMap III, using all 87,592 SNPs on chromosome 1.

Weir & Cockerham vs Unweighted Estimator



F_{ST} estimates for HapMap III, using the 42,463 SNPs on chromosome 1 that have at least five copies of the minor allele in samples from all 11 populations.

Multiple Loci: Weir & Cockerham

The Weir & Cockerham estimators for locus l are of the form

$$\text{Estimator}_l = \frac{\sum_u \text{Numerator}_{lu}}{\sum_u \text{Denominator}_{lu}}$$

With several loci, these can be extended to

$$\text{Estimator} = \frac{\sum_{l,u} \text{Numerator}_{lu}}{\sum_{l,u} \text{Denominator}_{lu}}$$

If every locus has the same value of θ, this multi-locus estimator gives an estimate of the common value. Otherwise it estimates a weighted average of the different θ values, where the weights are functions of the allele frequencies at the loci in the sum.

Multiple Loci: Unweighted

The unweighted estimators for locus l are of the form

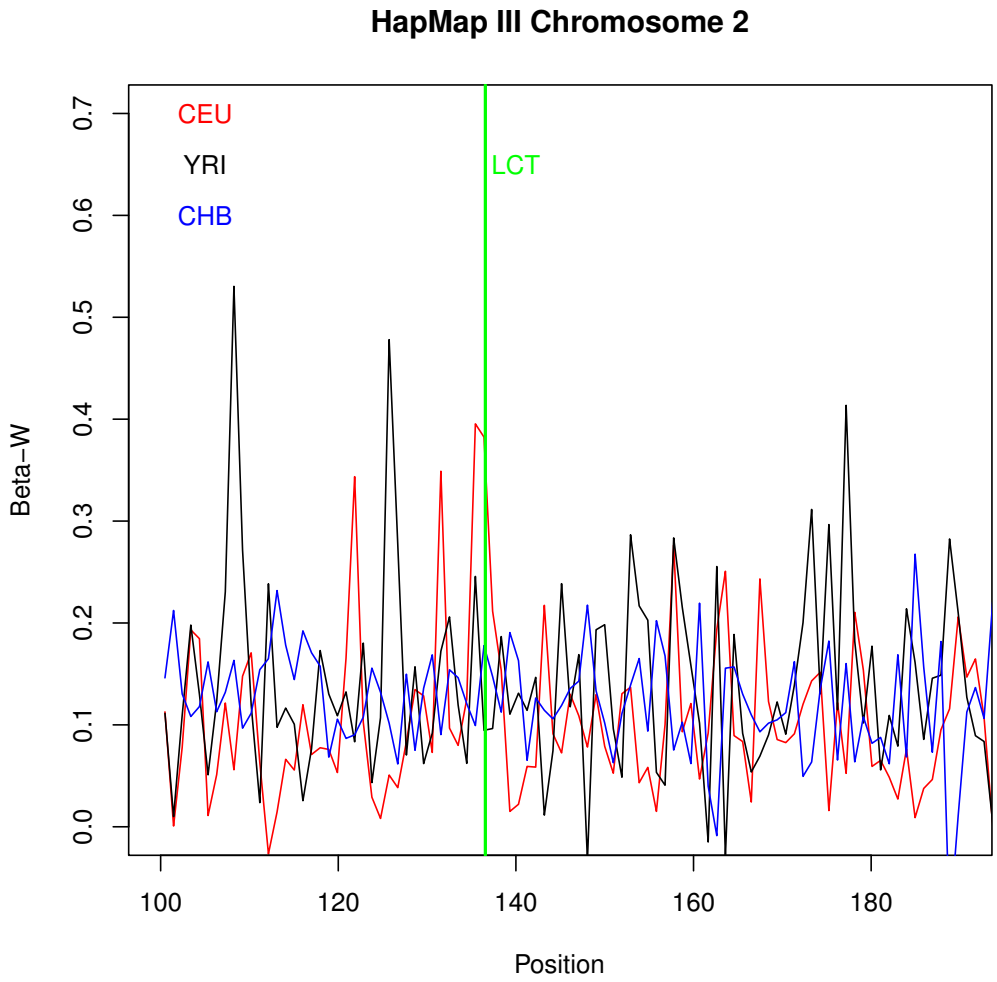
$$\text{Estimator}_l \;=\; \frac{\tilde{H}_{Bl} - \tilde{H}_{xl}}{\tilde{H}_{Bl}},\;\; x = i, W, ii'$$

With several loci, these can be extended to

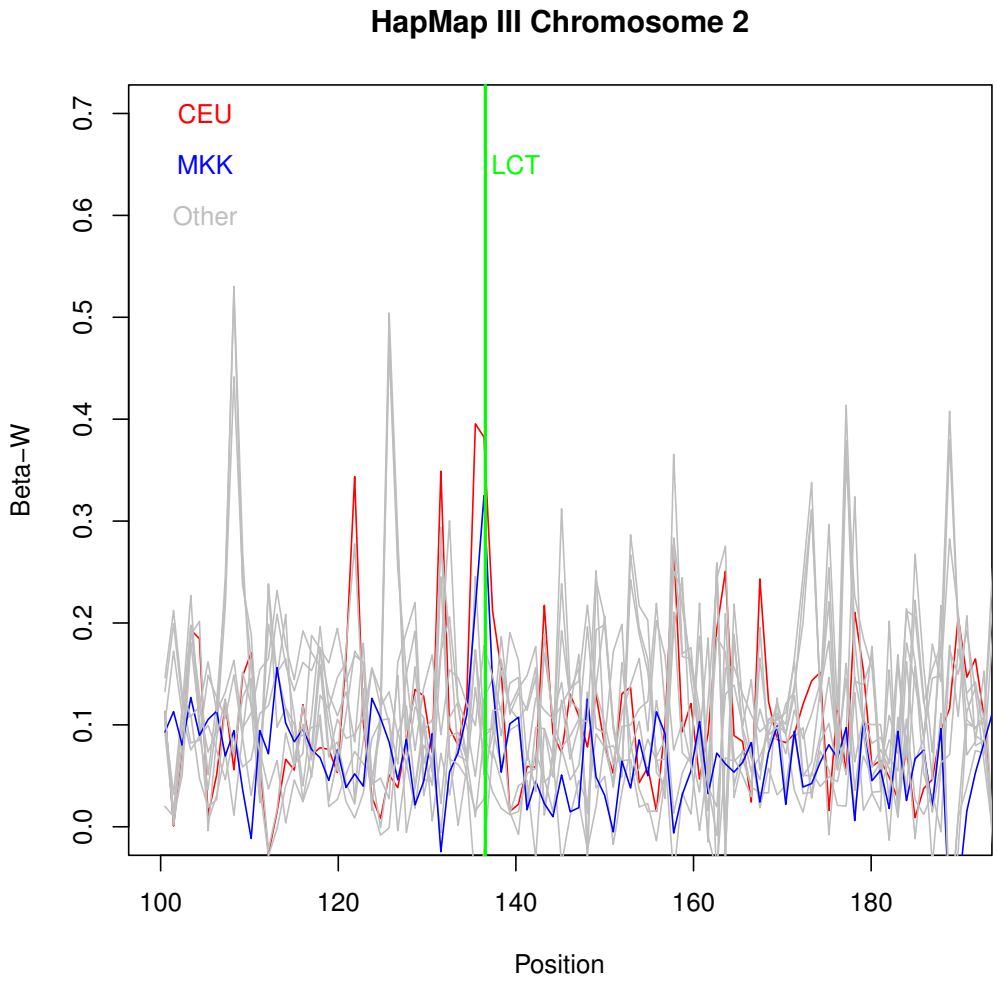
$$\text{Estimator} \;=\; \frac{\sum_l (\tilde{H}_{Bl} - \tilde{H}_{xl})}{\sum_l \tilde{H}_{Bl}} \;\; x = i, ii', W$$

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

$\widehat{\beta}_W$ in LCT Region: 3 Populations



$\widehat{\beta}_W$ in LCT Region: 11 Populations

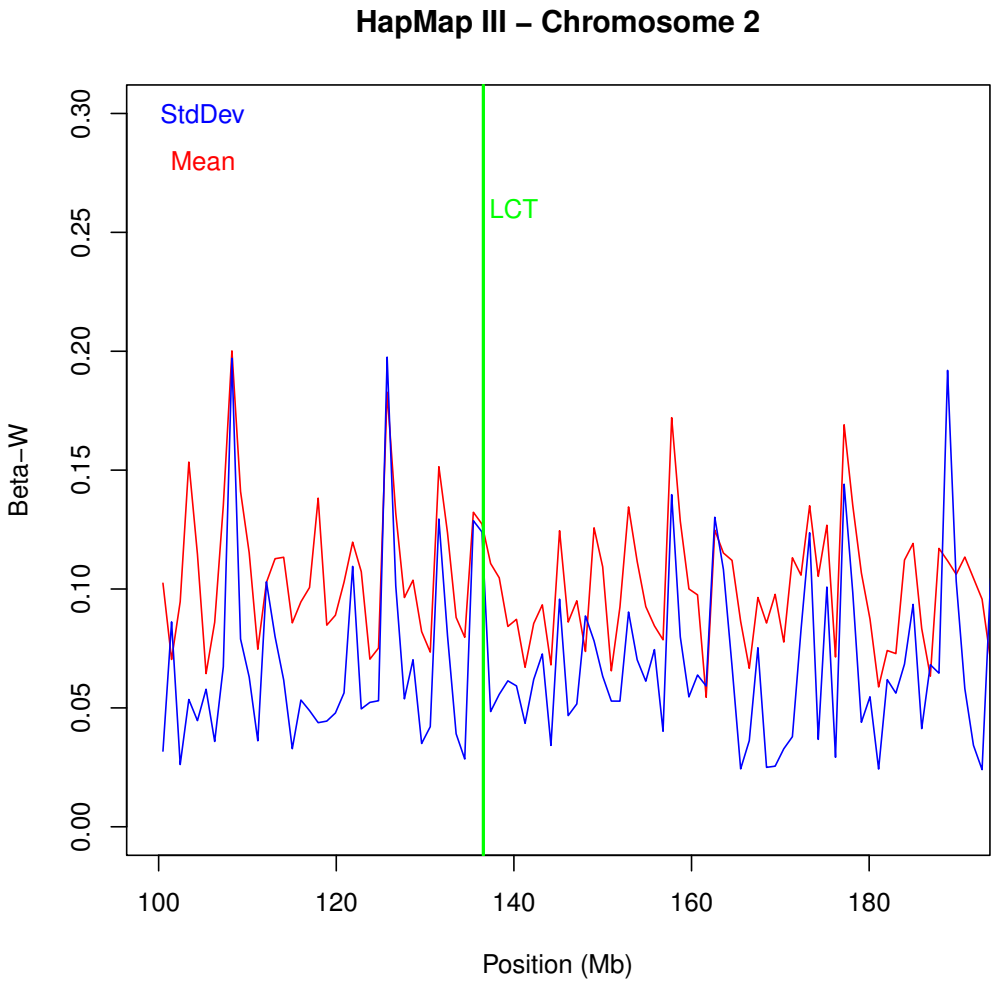


MKK Population

"The Maasai are a pastoral people in Kenya and Tanzania, whose traditional diet of milk, blood and meat is rich in lactose, fat and cholesterol. 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 and cholesterol regulation in 143156 founder individuals 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

Mean and Variance of $\widehat{\beta}_W$ in LCT Region



Pure Drift Model

Within a species, divergence among populations is driven primarily by genetic drift – the random choice of one of the two parental alleles for transmission from an individual to an offspring. Mutation is likely to be of less importance in the short term, and natural selection may not affect genetic markers.

Suppose a parental population has n_u of $2N$ alleles of type A_u. Conditional on $p_u = n_u/2N$, the number of alleles of this type in the offspring generation of N individuals is $B(2N, p_u)$.

Drift Model Variance

Write initial allele frequency as p_0.

Genera -tion	Allele freq.	Conditional on previous gen. Mean	Conditional on previous gen. Variance	Total over all generations Mean	Total over all generations Variance
0	p_0	p_0	0	p_0	0
1	p_1	p_0	$\frac{p_0(1-p_0)}{2N}$	p_0	$p_0(1-p_0)\left[1-\left(1-\frac{1}{2N}\right)\right]$
2	p_2	p_1	$\frac{p_1(1-p_1)}{2N}$	p_0	$p_0(1-p_0)\left[1-\left(1-\frac{1}{2N}\right)^2\right]$
3	p_3	p_2	$\frac{p_2(1-p_2)}{2N}$	p_0	$p_0(1-p_0)\left[1-\left(1-\frac{1}{2N}\right)^3\right]$
...	...	...	...	...	...
t	p_t	p_{t-1}	$\frac{p_{t-1}(1-p_{t-1})}{2N}$	p_0	$p_0(1-p_0)\left[1-\left(1-\frac{1}{2N}\right)^t\right]$

Drift Model IBD

Could also define θ as

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

so that

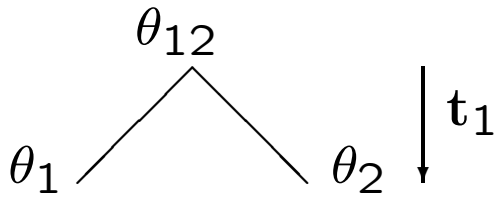
$$\text{Var}(p_t) \;=\; p_0(1-p_0)\theta_t$$

We can see that

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

and θ can be interpreted as the probability that two alleles are identical by descent. This probability interpretation is not necessary. With this drift model, θ serves as a measure of the number of generations since the founding population ($\theta = 0$).

Drift Model: Two Populations



$$\theta_1 = 1 - (1 - \theta_{12})X_1^{t_1} \quad , \quad \theta_2 = 1 - (1 - \theta_{12})X_2^{t_1}$$

where $X_i = (2N_i - 1)/2N_i$ and N_i is the population size for population i. Therefore

$$\beta_i = \frac{\theta_i - \theta_{12}}{1 - \theta_{12}} \;=\; 1 - X_i^{t_1} \approx \frac{t_1}{2N_i}$$

and this is the quantity estimated by the estimate $\widehat{\beta}_i$. Note that the times from each population to their most recent common ancestral population must be the same.

Drift Model: Two Populations

The unweighted within-population average estimator $\widehat{\beta}_W$ has expectation

$$\begin{aligned}\beta_W = \frac{\theta_W - \theta_B}{1-\theta_B} &= \frac{\frac{\theta_1+\theta_2}{2} - \theta_{12}}{1-\theta_{12}} \\ &= 1 - \frac{X_1^{t_1} + X_2^{t_1}}{2} \\ &\approx \frac{1}{2}\left(\frac{1}{2N_1} + \frac{1}{2N_2}\right) t_1 = \frac{t_1}{2N_h}\end{aligned}$$

so time is being estimated in terms of the harmonic mean of the two population sizes.

Drift Model: Two Populations

The Weir & Cockerham estimator has an expectation of

$$\mathcal{E}(\hat{\theta}_{WC}) = \frac{\frac{\theta_1+\theta_2}{2} - \theta_{12} + \frac{(n_1-n_2)(\theta_1-\theta_2)}{2n_1n_2}}{1 - \theta_{12} + \frac{(n_1-n_2)(\theta_1-\theta_2)}{2n_1n_2}}$$

which reduces to

$$\mathcal{E}(\hat{\theta}_{WC}) = \frac{\frac{\theta_1+\theta_2}{2} - \theta_{12}}{1 - \theta_{12}}$$

if $n_1 = n_2$ or $\theta_1 = \theta_2$.

Within-populations Example

FBI: African American (AA), Caucasian (CA) and Hispanic (HI) data.

	$\widehat{\beta}_i$				
Locus	AA	CA	HI	Average	$\widehat{\beta}_{WC}$
D3S135	.010	−.030	.069	.017	.019
vWA	.000	−.002	.050	.017	.019
FGA	.007	.012	−.008	.003	.006
D8S117	.026	.003	.009	.012	.015
D21S11	−.012	−.003	.047	.012	.014
D18S51	.011	.008	.010	.010	.012
D5S818	−.018	.061	.012	.019	.021
D13S31	.132	.028	−.042	.036	.040
D7S820	.014	−.016	.026	.008	.011
CSF1PO	−.048	.015	.051	.006	.008
TPOX	−.118	.090	.112	.027	.030
THO1	.078	.008	.041	.043	.045
D16S53	−.011	.028	.024	.014	.016
All loci	.010	.017	.032	.020	.020

Between-populations Example

FBI: African American (AA), Caucasian (CA) and Hispanic (HI) data.

	$\widehat{\beta}_{ii'}$		
Locus	AA&CA	AA&HI	CA&HI
D3S135	−.018	.026	−.009
vWA	−.018	.006	.010
FGA	.006	−.002	−.004
D8S117	−.012	.002	.009
D21S11	−.008	−.008	.015
D18S51	−.004	−.008	.011
D5S818	.003	−.039	.033
D13S31	.058	−.021	−.032
D7S820	.001	.004	−.006
CSF1PO	−.024	−.009	.034
TPOX	−.043	−.053	.097
THO1	−.034	.028	.006
D16S53	−.009	−.009	.018
Total	−.008	−.006	.014

Allele Frequency Distribution

The moment estimators of θ's and β's made use only of the mean and variance of the distribution of allele frequencies over populations. The advantages of maximum likelihood estimation are available if the whole distribution is known.

A convenient distribution is the normal, although it can be only an approximation. Under a class of evolutionary models (allelic exchangeability and stationarity) the Beta or Dirichlet distribution may be used.

Normal Distribution

If allele frequencies are not too extreme, it may be satisfactory to assume they have a normal distribution over populations.

In the case of equal values of θ_i and zero values of $\theta_{ii'}$, the sample frequencies $\tilde{p}_{iu}$ for alleles A_u in sample i have these properties:

$$\begin{aligned}
\mathcal{E}(\tilde{p}_{iu}) &= p_u \\
\text{Var}(\tilde{p}_{iu}) &= p_u(1-p_u)\left[\theta + \frac{1-\theta}{2n}\right] \\
\text{Cov}(\tilde{p}_{iu}, \tilde{p}_{iu'}) &= -p_u p_{u'}\left[\theta + \frac{1-\theta}{2n}\right], u \neq u' \\
\text{Cov}(\tilde{p}_{iu}, \tilde{p}_{i'u'}) &= 0, i \neq i'
\end{aligned}$$

Multiple alleles suggest a multivariate normal approach.

Multivariate Normality

$$Q = \sum_{i=1}^{r} (\tilde{\mathbf{p}}_i - \bar{\mathbf{p}})' \mathbf{V}^{-1} (\tilde{\mathbf{p}}_i - \bar{\mathbf{p}})$$

$$= \sum_{i=1}^{r} \sum_{u=1}^{m} \frac{(\tilde{p}_{iu} - \bar{p}_u)^2}{\bar{p}_u}$$

($\bar{p}_u = \sum_{i=1}^{r} \tilde{p}_{iu}/r$) has distribution

$$Q \sim \theta \chi^2_{(r-1)(m-1)}$$

suggesting that

$$(r-1)(m-1)\hat{\theta} = Q$$

or

$$\hat{\theta}_N \approx \frac{1}{(m-1)(r-1)} \sum_{i=1}^{r} \sum_{u=1}^{m} \frac{(\tilde{p}_{iu} - \bar{p}_u)^2}{\bar{p}_u}$$

Properties of Normal Estimate

From the chi-distribution of Q:

$$\begin{aligned} \mathcal{E}(Q) &= (r-1)(m-1)\theta \\ \mathrm{Var}(Q) &= 2(r-1)(m-1)\theta^2 \end{aligned}$$

so that

$$\mathcal{E}(\hat{\theta}_N) = \theta$$

$$\begin{aligned} \mathrm{Var}(\hat{\theta}_N) &= \frac{2[1+(2n-1)\theta)]^2}{(r-1)(m-1)(2n-1)^2} \\ &\approx \frac{2\theta^2}{(r-1)(m-1)}, \text{ (large } n) \end{aligned}$$

Confidence Intervals

For large sample sizes,

$$\frac{(r-1)(m-1)\widehat{\theta}_N}{\theta} \sim \chi^2_{(r-1)(m-1)}$$

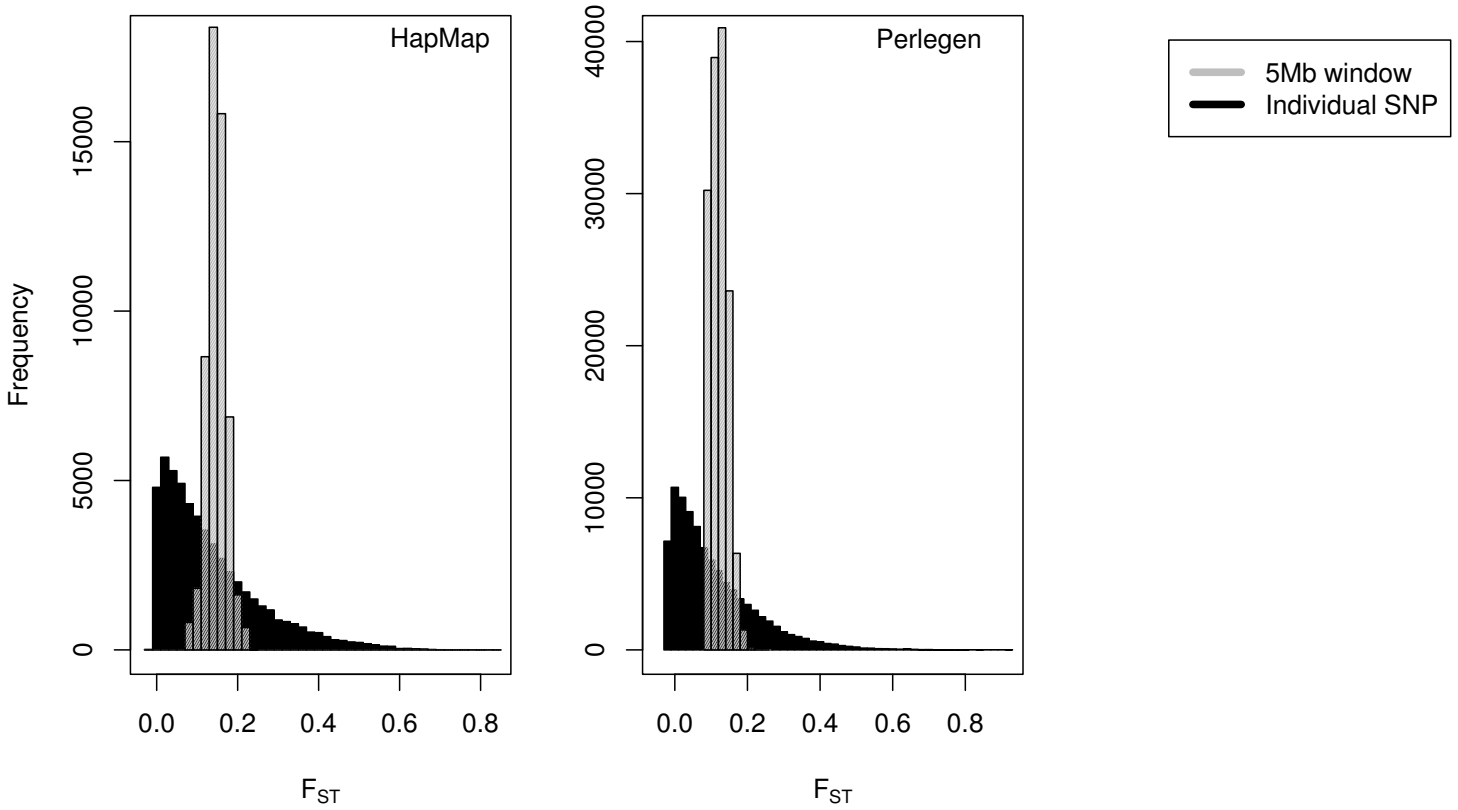
so a 95% confidence interval for θ is

$$\left(\frac{(r-1)(m-1)\widehat{\theta}_N}{\chi^2_{0.975}} , \frac{(r-1)(m-1)\widehat{\theta}_N}{\chi^2_{0.025}} \right)$$

An estimated θ of 0.01 from data from five populations for a locus with five alleles, for example, leads to a 95% confidence interval of (0.001, 0.036).

Analytical approach more convenient than bootstrapping.

Effect of Number of Loci



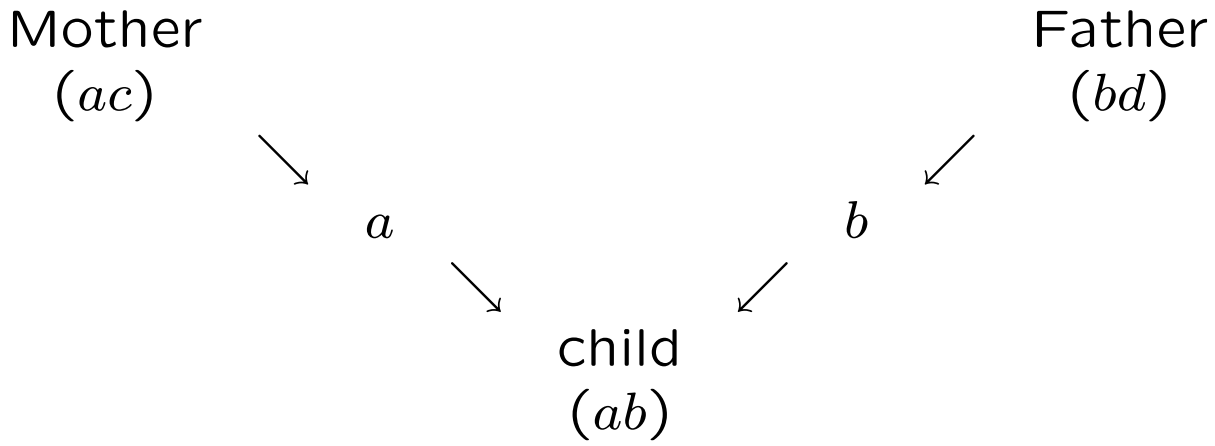
INBREEDING AND RELATEDNESS

Inbreeding

Two alleles that descend from the same allele are said to be **identical by descent (ibd)**.

The probability that two parents transmit ibd alleles to a child is the **inbreeding coefficient F** of the child.

Use small letters for alleles, and capital letters for their particular type: i.e. parents might transmit alleles a and b to their child, and these alleles might both be type A.



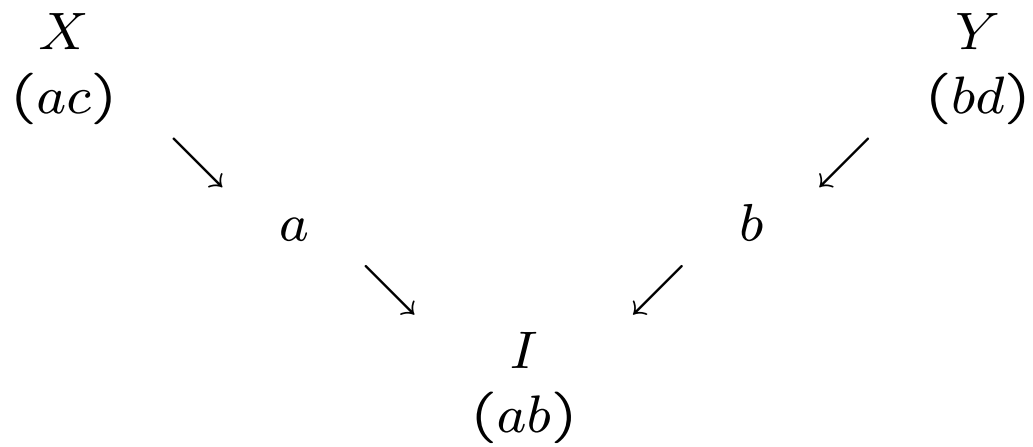
Relatedness

Two individuals that have ibd alleles are said to be **related**.

The probability that an allele taken at random from one individual is ibd to an allele taken at random from another individual is the **coancestry coefficient** θ of those two individuals.

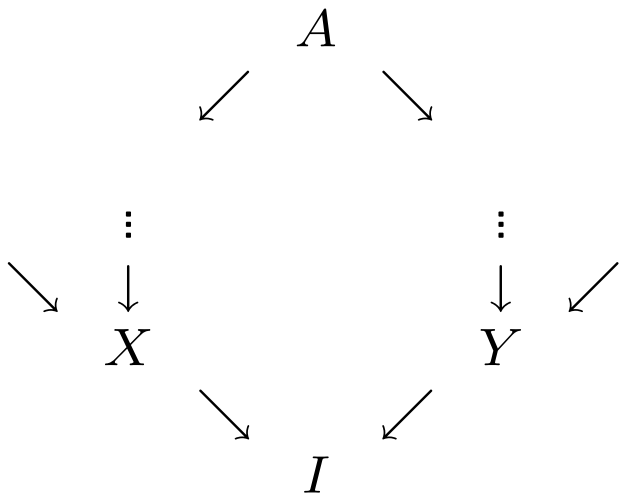
The inbreeding coefficient of an individual is the coancestry of its parents.

Relatedness



$$F_I = \theta_{XY}$$

Path Counting



Identify the path linking the parents of I to their common ancestor(s).

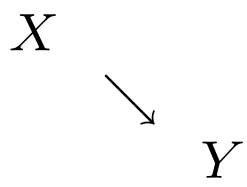
Path Counting

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.

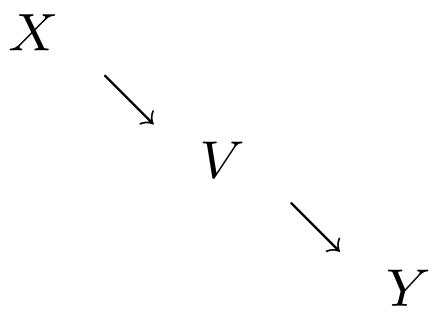
Parent-Child



The common ancestor of parent X and child Y is X. The path linking X, Y to their common ancestor is YX and this has $n = 2$ individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

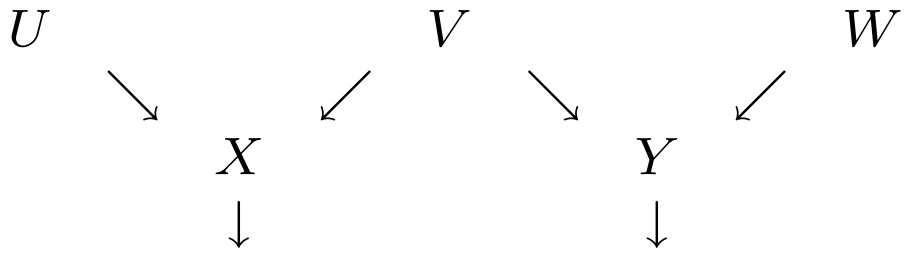
Grandparent-grandchild



The common ancestor of grandparent X and grandchild Y is X. The path linking X, Y to their common ancestor is YVX and this has $n = 3$ individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

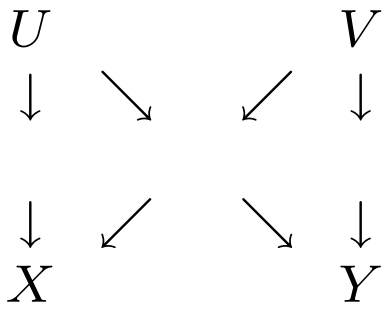
Half sibs



The common ancestor of half sibs X and Y is V. The path linking X, Y to their common ancestor is XVY and this has $n = 3$ individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

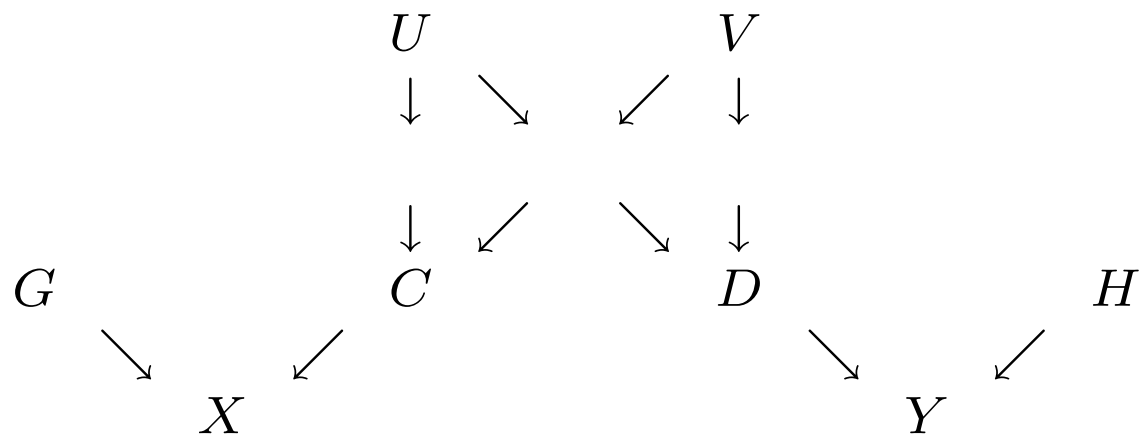
Full sibs



The common ancestors of full sibs X and Y are U and V. The paths linking X, Y to their common ancestors are XUY and XVY and these each have $n = 3$ individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \frac{1}{4}$$

First cousins



The common ancestors of cousins X and Y are U and V. The paths linking X, Y to their common ancestors are $XCUDY$ and $XCVDY$ and these each have $n = 5$ individuals. Therefore

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

Genotype frequencies

Consider an individual with inbreeding coefficient F. The probability it has two ibd alleles is F, and the probability of two non-ibd alleles is $1 - F$. The probability that any allele is type A is p_A, the population allele frequency. So, the probability the individual is homozygous is

$$\begin{aligned} P_{AA} &= \Pr(AA|\text{ibd})\Pr(\text{ibd}) + \Pr(AA|\text{not ibd})\Pr(\text{not ibd}) \\ &= p_A \times F + p_A^2 \times (1 - F) \end{aligned}$$

There is an evolutionary perspective here: F reflects the history that leads to alleles being ibd. Identity by descent has meaning only with reference to previous generations. The allele frequency p_A is the expected value over evolutionary histories.

Genotype frequencies

To emphasize the difference from Hardy-Weinberg:

$$P_{AA} = p_A^2 + Fp_A(1 - p_A)$$

Because heterozygous individuals must have non-ibd alleles:

$$\begin{aligned} P_{Aa} &= 2(1 - F)p_A p_a \\ &= 2p_A p_a - 2Fp_A p_a \end{aligned}$$

First Cousin Example

For individuals with parents who were first cousins, $F = 1/16 = 0.0625$. For a locus with allele frequencies $\{p_i\}$ that were all 0.10:

$$\begin{aligned} P_{ii} &= (0.1)^2 + 0.0625(0.1)(0.10) = 0.015625 \\ P_{ij} &= 2(0.1)(0.1) - 2(0.0625)(0.1)(0.1) = 0.018750 \end{aligned}$$

Individual Inbreeding Coefficients

Earlier slides considered the estimation of the within-population inbreeding coefficient f that made use of sample allele frequencies from the particular population under consideration. This f is a description of the population.

Individual-specific (total) inbreeding coefficients F can be estimated under the assumption that all loci have the same coefficient, now interpreted as the probability of identity by descent (ibd). Many loci are needed.

At locus l, write p_l for the frequency of A_l and code the genotypes A_lA_l, A_la_l, a_la_l as $X_l = 2, 1, 0$. These new variables have the properties $\mathcal{E}(X_l) = 2p_l, \mathrm{Var}(X_l) = 2p_l(1-p_l)(1+F)$.

Individual Inbreeding Coefficients

Then one moment estimator is formed by summing over loci $l, l = 1, 2, \ldots L$:

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

but an estimate with smaller variance is

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

In practice, we need to use sample allele frequencies.

MLE for Individual Inbreeding Coefficients

To avoid having to choose among MoM estimates 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,

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

From Bayes' theorem then

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

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.

Descent measures for four alleles

Much of population and quantitative genetics theory rests on the comparison of pairs of individuals - either on their genotypes or on their trait values, or both. For single-locus analyses this requires a discussion of four-allele descent measures.

One-locus descent measures δ for any four alleles a, b, c, d identify which subsets of the four are ibd. The measures give the probabilities of the specified alleles being ibd, and other alleles of the four being not-ibd.

Complete Set of IBD Measures

A complete description of the ibd status among four alleles a, b, c, d carried by two individuals requires 15 measures (as opposed to the two, F, $1 - F$, for one individual):

Alleles ibd*	Probability	Alleles ibd*	Probability
a, b, c, d	δ_{abcd}	a, b	δ_{ab}
a, b, c	δ_{abc}	a, c	δ_{ac}
a, b, d	δ_{abd}	a, d	δ_{ad}
a, c, d	δ_{acd}	b, c	δ_{bc}
b, c, d	δ_{bcd}	b, d	δ_{bd}
a, b and c, d	$\delta_{ab.cd}$	c, d	δ_{cd}
a, c and b, d	$\delta_{ac.bd}$	none	δ_0
a, d and b, c	$\delta_{ad.bc}$		

*Alleles not listed are not ibd to those listed

Nine-parameter IBD Set

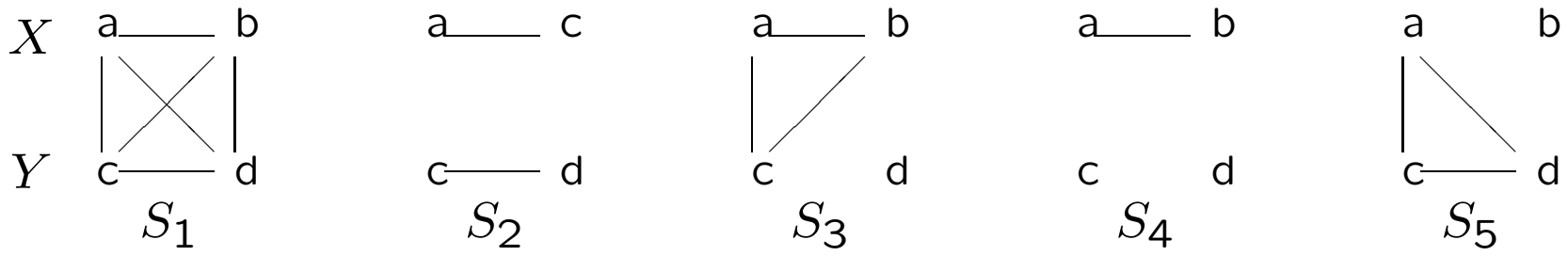
In most applications there is no need to distinguish between maternal and paternal alleles and the 15 ibd states can be collapsed into nine:

	Notation	
Alleles ibd*	Cockerham, 1971	Jacquard, 1970
a, b, c, d	δ_{abcd}	Δ_1
a, b, c or a, b, d	$\delta_{abc} + \delta_{abd}$	Δ_3
a, c, d or b, c, d	$\delta_{acd} + \delta_{bcd}$	Δ_5
a, b and c, d	$\delta_{ab.cd}$	Δ_2
(a, c and b, d) or (a, d and b, c)	$\delta_{ac.bd} + \delta_{ad.bc}$	Δ_7
a, b	δ_{ab}	Δ_4
c, d	δ_{cd}	Δ_6
a, c or a, c or b, c or b, d	$\delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd}$	Δ_8
none	δ_0	Δ_9

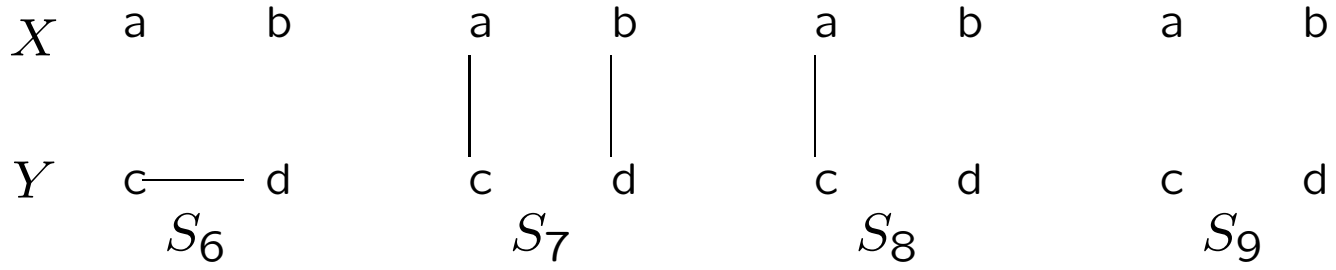
*Alleles not listed are not ibd to those listed

Nine-parameter IBD Set

Solid lines join pairs of ibd alleles: top row is the pair of alleles for X, bottom row the pair of alleles for Y.



S_1 S_2 S_3 S_4 S_5



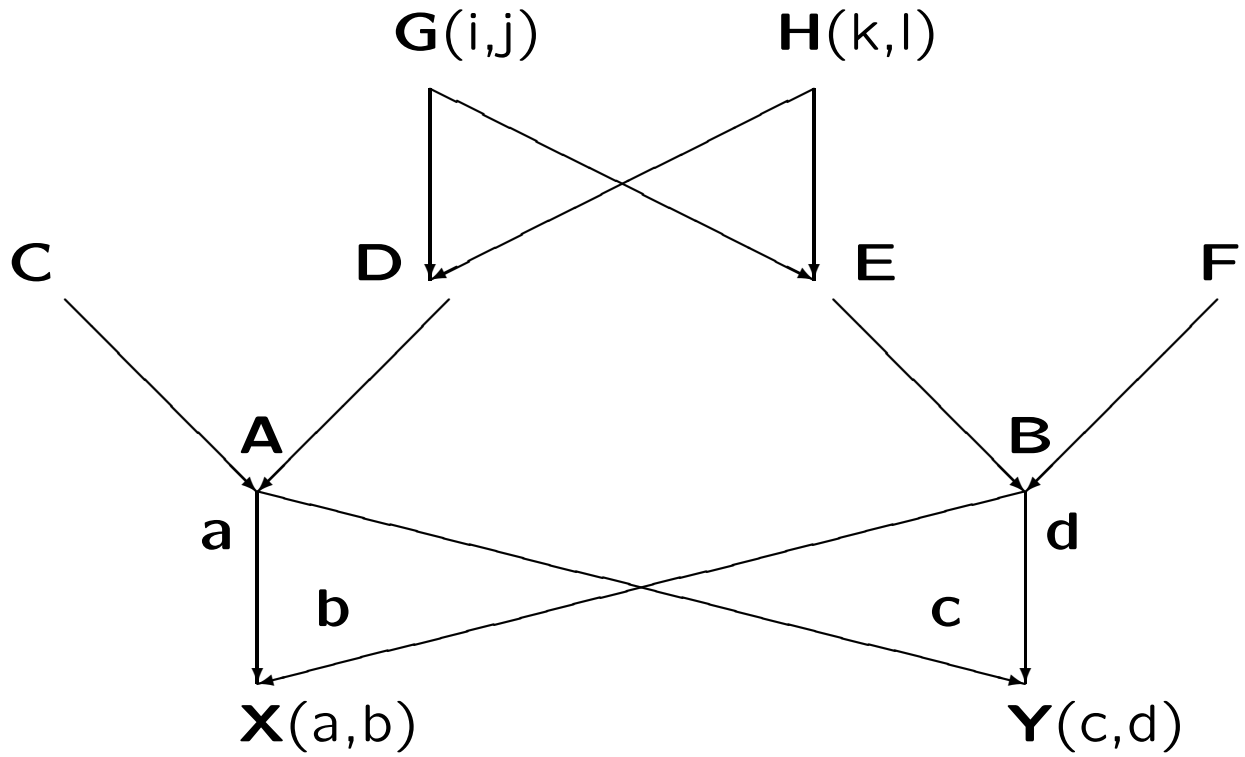
S_6 S_7 S_8 S_9

Coancestry Coefficient

The coancestry coefficient θ_{XY} is the probability that a random allele from $X(ab)$ is ibd to a random allele from $Y(cd)$:

$$\begin{aligned}
\theta_{XY} &= \frac{1}{4}[\Pr(a \equiv c) + \Pr(a \equiv d) + \Pr(b \equiv c) + \Pr(b \equiv d)] \\
&= \delta_{abcd} + \frac{1}{2}(\delta_{abc} + \delta_{abd} + \delta_{acd} + \delta_{bcd}) + \frac{1}{2}(\delta_{ac.bd} + \delta_{ad.bc}) \\
&\quad + \frac{1}{4}(\delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd}) \\
&= \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8
\end{aligned}$$

Siblings whose parents are first cousins



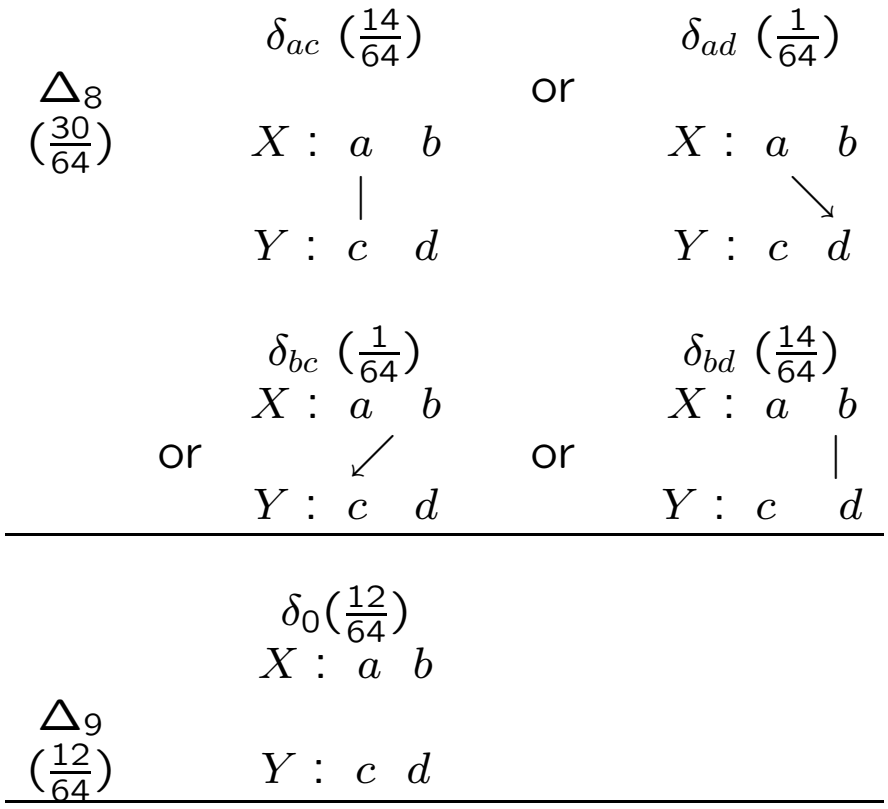
Siblings whose parents are first cousins

	$\delta_{abcd}\ (\frac{1}{64})$ $X:\ a$—b $\quad\quad\ \vert\quad\ \vert$		$\delta_{ab.cd}\ (0)$ $X:\ a$—b
Δ_1 $(\frac{1}{64})$	$Y:\ c$—d	Δ_2 (0)	$Y:\ c$—d
	$\delta_{abc}\ (\frac{1}{64})$ $X:\ a$—b $\quad\quad\ \vert$		$\delta_{abd}\ (\frac{1}{64})$ $X:\ a$—b $\quad\quad\quad\ \vert$
Δ_3 $(\frac{2}{64})$	$Y:\ c\quad d$	or	$Y:\ c\quad d$
	$\delta_{ab}\ (\frac{1}{64})$ $X:\ a$—b		$\delta_{cd}\ (\frac{1}{64})$ $X:\ a\quad b$
Δ_4 $(\frac{1}{64})$	$Y:\ c\quad d$	Δ_6 $(\frac{1}{64})$	$Y:\ c$—d

Siblings whose parents are first cousins

Δ_5 $(\frac{2}{64})$	δ_{acd} $(\frac{1}{64})$ $X:\ a \quad b$ $\quad\ \vert$ $Y:\ c\text{—}d$	or	δ_{bcd} $(\frac{1}{64})$ $X:\ a \quad b$ $\quad\ \ \vert$ $Y:\ c\text{—}\ d$
Δ_7 $(\frac{15}{64})$	$\delta_{ac.bd}$ $(\frac{15}{64})$ $X:\ a \quad b$ $\quad\ \vert \quad \vert$ $Y:\ c \quad d$	or	$\delta_{ad.bc}$ (0) $X:\ a \quad b$ $\quad\ X$ $Y:\ c \quad d$

Siblings whose parents are first cousins



Siblings whose parents are first cousins

$$
\begin{aligned}
\theta &= \frac{1}{4}(\delta_{abcd} + \delta_{abc} + \delta_{acd} + \delta_{ac.bd} + \delta_{ac}) \\
&\quad + \frac{1}{4}(\delta_{abcd} + \delta_{abd} + \delta_{acd} + \delta_{ad.bc} + \delta_{ad}) \\
&\quad + \frac{1}{4}(\delta_{abcd} + \delta_{abc} + \delta_{bcd} + \delta_{ad.bc} + \delta_{bc}) \\
&\quad + \frac{1}{4}(\delta_{abcd} + \delta_{abd} + \delta_{bcd} + \delta_{ac.bd} + \delta_{bd}) \\
&= \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8 \\
&= \frac{9}{32}
\end{aligned}
$$

Non-inbred Relatives

There is a final reduction when neither individual is inbred, as then neither a, b nor c, d are ibd. There are only three states and the three probabilities are often written as $k_2 = \Delta_7, k_1 = \Delta_8$ or $k_0 = \Delta_9$ to indicate the number of pairs of ibd alleles carried by the two individuals. Values of these three probabilities for some common relationships are:

Relationship	k_2	k_1	k_0	$\theta = \frac{1}{2}k_2 + \frac{1}{4}k_1$
Identical twins	1	0	0	1/2
Full sibs	1/4	1/2	1/4	1/4
Parent-child	0	1	0	1/4
Double first cousins	1/16	3/8	9/16	1/8
Half sibs*	0	1/2	1/2	1/8
First cousins	0	1/4	34	1/16
Unrelated	0	0	1	0

* Also grandparent-grandchild and avuncular (e.g. uncle-niece).

Joint genotypic probabilities

Genotypes	General	Non – inbred
ii, ii	$\Delta_1 p_i + (\Delta_2 + \Delta_3 + \Delta_5 + \Delta_7)p_i^2$ $+ (\Delta_4 + \Delta_6 + \Delta_8)p_i^3 + \Delta_9 p_i^4$	$k_2 p_i^2 + k_1 p_i^3 + k_0 p_i^4$
ii, jj	$\Delta_2 p_i p_j + \Delta_4 p_i p_j^2$ $+ \Delta_6 p_i^2 p_j + \Delta_9 p_i^2 p_j^2$	$k_0 p_i^2 p_j^2$
ii, ij	$\Delta_3 p_i p_j + (2\Delta_4 + \Delta_8)p_i^2 p_j$ $+ 2\Delta_9 p_i^3 p_j$	$k_1 p_i^2 p_j + 2k_0 p_i^3 p_j$
ii, jk	$2\Delta_4 p_i p_j p_k + 2\Delta_9 p_i^2 p_j p_k$	$2k_0 p_i^2 p_j p_k$
ij, ij	$2\Delta_7 p_i p_j + \Delta_8 p_i p_j (p_i + p_j)$ $+ 4\Delta_9 p_i^2 p_j^2$	$2k_2 p_i p_j + k_1 p_i p_j (p_i + p_j)$ $+ 4k_0 p_i^2 p_j^2$
ij, ik	$\Delta_8 p_i p_j p_k + 4\Delta_9 p_i^2 p_j p_k$	$k_1 p_i p_j p_k + 4k_0 p_i^2 p_j p_k$
ij, kl	$4\Delta_9 p_i p_j p_k p_l$	$4k_0 p_i p_j p_k p_l$

Example: Non-inbred full sibs

Genotypes	Probability
ii, ii	$p_i^2(1+p_i)^2/4$
ii, jj	$p_i^2 p_j^2/4$
ii, ij	$p_i p_j(p_i+p_j)/2$
ii, jk	$p_i^2 p_j p_k/2$
ij, ij	$p_i p_j(1+p_i+p_j+2p_i p_j)/2$
ij, ik	$p_i p_j p_k(1+2p_i)/2$
ij, kl	$p_i p_j p_k p_l$

Method of Moments for Relatedness Coefficients

PLINK 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

ibs state	Genotypes	Probability
2	$(AA, AA), (aa, aa), (Aa, Aa)$	$(p^2+q^2)^2 k_0 + k_1(p^3+pq+q^3) + k_2$
1	$(AA, Aa), (Aa, AA), (aa, Aa), (Aa, aa)$	$4pq(p^2+q^2)k_0 + 2pqk_1$
0	$(AA, aa), (aa, AA)$	$2p^2q^2k_0$

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 to get

$$N_0 \;=\; \Pr(\text{ibd} = 0)\sum_{\text{loci}} 2p^2q^2$$

This gives a moment estimate

$$\Pr(\text{ibd} = 0) \;=\; \frac{N_0}{\sum_{\text{loci}} 2p^2q^2}$$

MoM Approach: k_1

From

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

sum over loci to get

$$N_1 = \Pr(\text{ibd}=0)\sum_{\text{loci}} 4pq(p^2+q^2) + \Pr(\text{ibd}=1)\sum_{\text{loci}} 2pq$$

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

$$\Pr(\text{ibd}=1) = \frac{N_1 - \sum_{\text{loci}} 4pq(p^2+q^2)\Pr(\text{ibd}=0)}{\sum_{\text{loci}} 2pq}$$

MoM Approach: k_2

From

$$\begin{aligned}\Pr(\text{ibd}=2) &= \Pr(\text{ibs}=2|\text{ibd}=0)\Pr(\text{ibd}=0)\\ &\quad + \Pr(\text{ibs}=2|\text{ibd}=1)\Pr(\text{ibd}=1)\\ &\quad + \Pr(\text{ibs}=2|\text{ibd}=2)\Pr(\text{ibd}=2)\end{aligned}$$

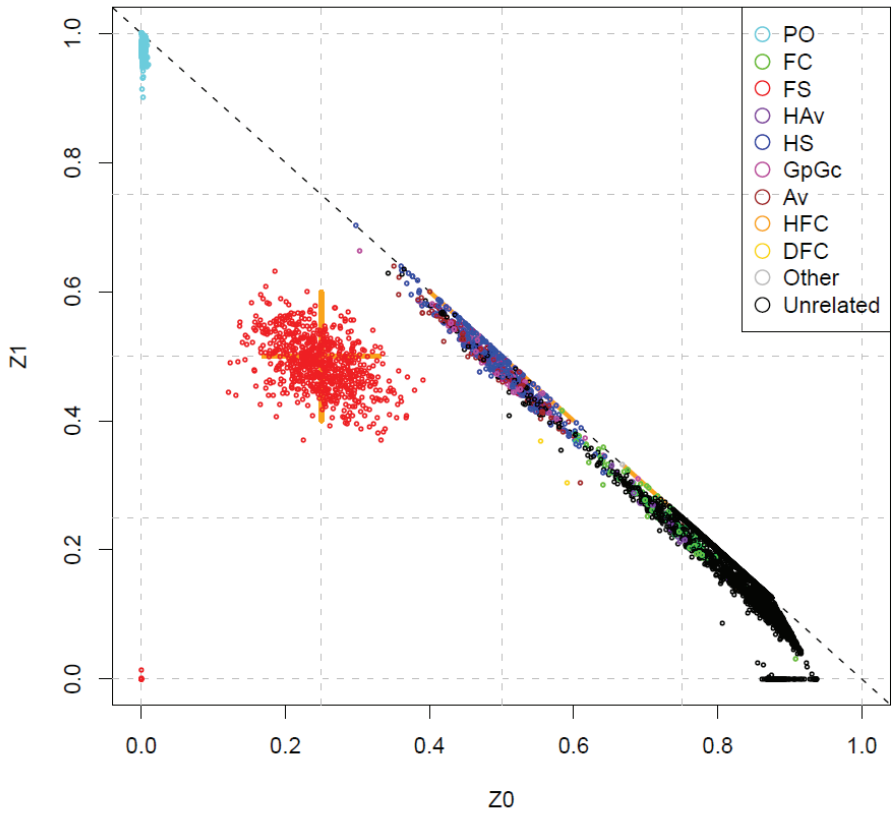
sum over loci to get

$$\begin{aligned}N_2 &= \Pr(\text{ibd}=0)\sum_{\text{loci}}(p^2+q^2)^2 + \Pr(\text{ibd}=1)\sum_{\text{loci}}(p^3+pq+q^3)\\ &\quad + \Pr(\text{ibd}=2)\end{aligned}$$

but we already have estimates of $\Pr(\text{ibd}=0)$ and $\Pr(\text{ibd}=1$. Therefore

$$\Pr(\text{ibd}=1) = \frac{N_2 - \sum_{\text{loci}}(p^2+q^2)^2\Pr(\text{ibd}=0) - \sum_{\text{loci}}(p^3+pq+q^3)\Pr(\text{ibd}=1)}{\sum_{\text{loci}}1}$$

Example


PO
FC
FS
HAv
HS
GpGc
Av
HFC
DFC
Other
Unrelated
Z1
Z0
0.0
0.2
0.4
0.6
0.8
1.0

MoM for Coancestry Coefficient

One moment estimator for the coancestry between individuals G and H is formed by summing over loci $l, l = 1, 2, \ldots L$:

$$\widehat{\theta}_{GH} = \frac{1}{L} \sum_{l=1}^{L} \frac{(X_l - 2p_l)(Y_l - 2p)}{2p_l(1 - p_l)}$$

where X_l, Y_l are $2, 1, 0$ if G, H are (AA, Aa, aa), respectively, at locus l. An estimate with smaller variance is

$$\widehat{\theta}_{GH} = \frac{\sum_{l=1}^{L}(x_l - 2p_l)(y_l - 2p)}{\sum_{l=1}^{L} 2p_l(1 - p_l)}$$

In practice, sample allele frequencies will need to be used.

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 S is the observed pair of genotypes, we know the conditional probabilities $\Pr(S|D_i)$ where the D_i represent the identity states (the relationship). The probability of ibd state D_i is k_i.

For example, if S is the state (AA, AA); $\Pr(S|D_0) = p_A^4$, $\Pr(S|D_1) = p_A^3$, and $\Pr(S|D_2) = p_A^2$ and

$$\Pr(S) = \sum_i \Pr(S|D_i)\Pr(D_i) = p_A^4 k_0 + p_A^3 k_1 + p_A^2 k_2$$

MLE for Relatedness Coefficients

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

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

The procedure begins with initial estimates of the $k_i = \Pr(D_i)$'s. The denominator is calculated from the law of total probability by adding over the three descent states:

$$\Pr(S_l) \;=\; \sum_i \Pr(S_l|D_i)\Pr(D_i) = \sum_i \Pr(S_l|D_i)k_i$$

MLE for Relatedness Coefficients

The updated estimates are obtained by averaging over m loci:

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

These updated values are then substituted into the right hand side and the process continued until the likelihood L no longer changes (or changes by less than some specified small amount) where

$$L = \prod_{l=1}^{m}\left[\sum_i \Pr(S_l|D_i)k_i\right]$$

It will be better to monitor changes in the log-likelihood.

"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:

$$
\begin{aligned}
L(AA, AA) &= k_0 + k_1/p_A + k_2/p_A^2 \\
L(BB, BB) &= k_0 + k_1/p_B + k_2/p_B^2 \\
L(AB, AB) &= k_0 + k_1/(4p_Ap_B) + k_2/(2p_Ap_B)
\end{aligned}
$$

$$
\begin{aligned}
L(AA, AB) &= k_0 + k_1/(2p_A) \\
L(BB, AB) &= k_0 + k_1/(2p_B)
\end{aligned}
$$

$$
L(AA, BB) = k_0
$$

Testing relationship

Hypotheses about alternative pairs of relationships may be tested with likelihood ratio test statistics. These ratios are the probability of the observed pair of genotypes under one hypothesis divided by the probability under the alternative hypothesis. These ratios are multiplied over (independent) loci.

Each hypothesis is described by a set of k's and there are three hypotheses likely to be of interest:
1. individuals are unrelated ($k_0 = 1$)
2. individuals have a specified (or annotated) relationship (k's specified)
or 3. individuals are related to an extent measured by the estimated k's.

Testing relationship

The three likelihoods may be written as L_0, L_S, L_E, respectively and each of the three ratios of these may be of interest. It may be satisfactory to regard the ratios that involve estimated k's as having chi-square distributions, but this is not justified for the unrelated versus specified situation:

H_0	H_1	LR	Null distribution
Unrelated	Specified	$-2\ln(L_0/L_S)$	$--$
Unrelated	Estimated	$-2\ln(L_0/L_E)$	$\chi^2_{(2)}$
Specified	Estimated	$-2\ln(L_S/L_E)$	$\chi^2_{(2)}$

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

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

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

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

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

$$\rho_{G,X} = \rho^2_{MT}$$

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

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

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.

It is common for investigators to assume a multiplicative disease model (i.e. additive on a log scale):

$$G_{TT} = \alpha\beta^2, G_{Tt} = \alpha\beta, G_{tt} = \alpha$$

so that

$$\begin{aligned} \mu_G &= \alpha(p_T\beta + p_t)^2 \\ \sigma^2_{A_T} &= 2p_T p_t \alpha^2(1-\beta)^2(p_T\beta + p_t)^2 \\ \sigma^2_{D_T} &= p^2_T p^2_t \alpha^2(1-\beta)^4 \end{aligned}$$

This leads to Hardy-Weinberg equilibrium at marker loci among cases since

$$2\mu_G\sigma_{D_T} = \sigma^2_{A_T}$$

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

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

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

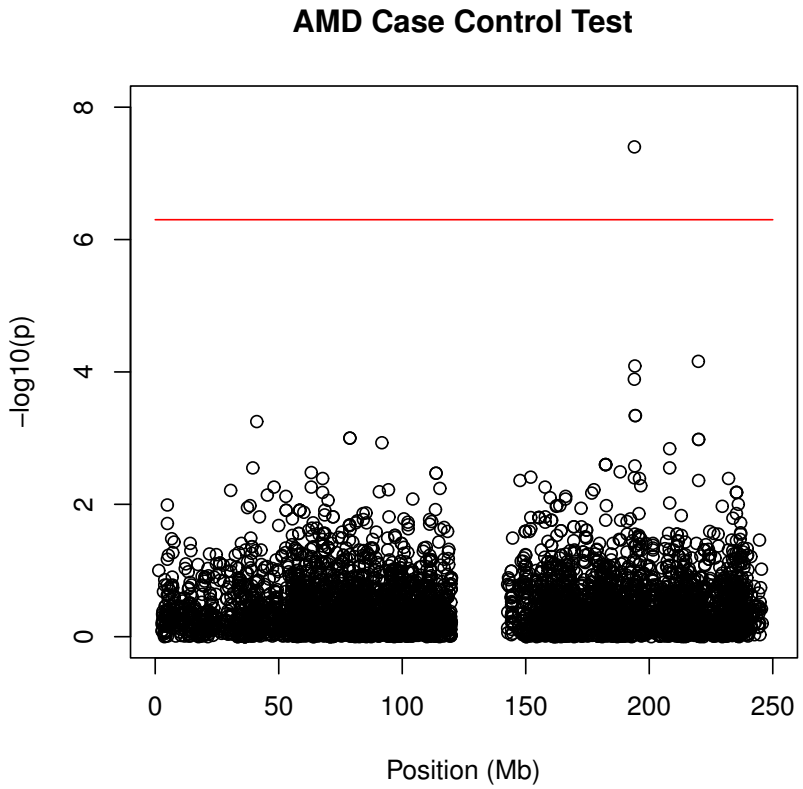
Allelic Case Control Test

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

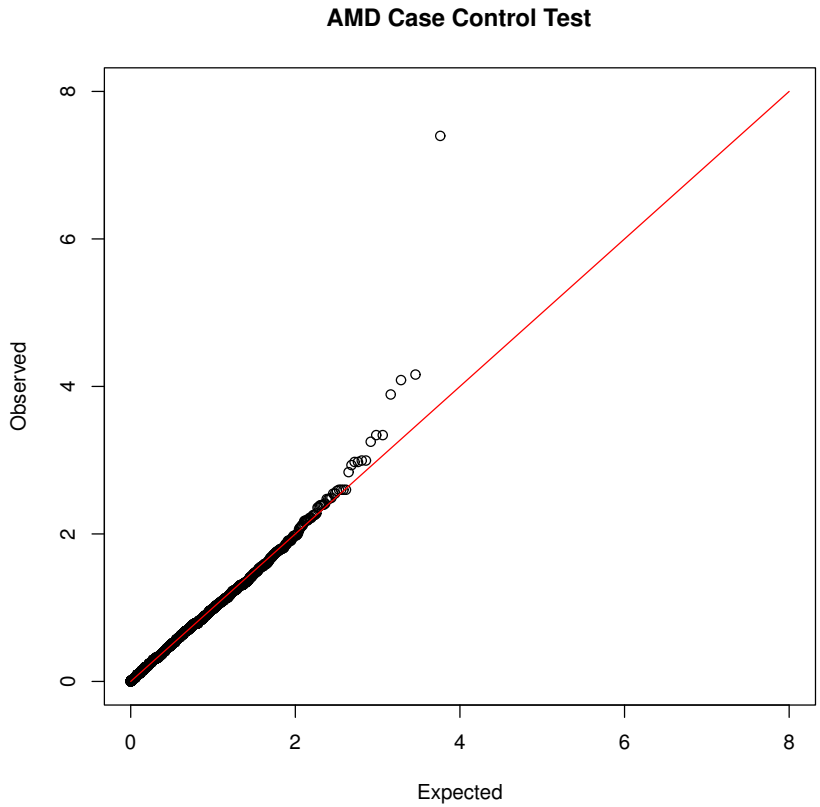
$$\mathcal{E}(X_A^2) \approx (1+f)$$

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.

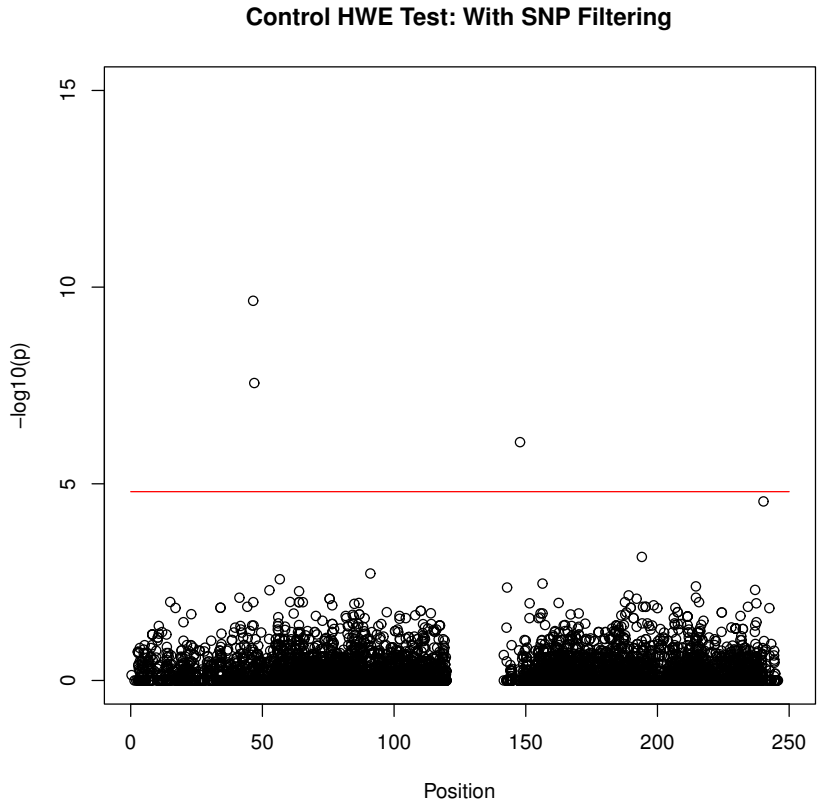
AMD Example: Case-control test statistics on chromosome 1



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.