Population Genetic Data Analysis

Summer Institute in Statistical Genetics
University of Washington
July 12-14, 2017

Jérôme Goudet: jerome.goudet@unil.ch

Bruce Weir: bsweir@uw.edu

Contents

Lectures on these topics by Bruce Weir will alternate with R exercises led by Jérôme Goudet.
The R material is at http://www2.unil.ch/popgen/teaching/SISG17/

GENETIC DATA

Sources of Population Genetic Data

Phenotype	Mendel’s peas
	Blood groups
Protein	Allozymes
	Amino acid sequences
DNA	Restriction sites, RFLPs
	Length variants: VNTRs, STRs
	Single nucleotide polymorphisms
	Single nucleotide variants

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 RC. 1974. The Genetic Basis of Evolutionary Change. Columbia University Press.

STR markers: CTT set

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

			Usual No.
Locus	Structure	Chromosome	of repeats
CSF1PO	$[\mathrm{AGAT}]_n$	5q	6–16
TPOX	$[\mathrm{AATG}]_n$	2p	5–14
TH01*	$[\mathrm{AATG}]_n$	11p	3–14

* "9.3" is $[\mathrm{AATG}]_6\mathrm{ATG}[\mathrm{AATG}]_3$

Length variants detected by capillary electrophoresis.

“CTT” Data - Forensic Frequency Database

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

Sequencing of STR Alleles

“STR typing in forensic genetics has been performed traditionally using capillary electrophoresis (CE). Massively parallel sequencing (MPS) has been considered a viable technology in recent years allowing high-throughput coverage at a relatively affordable price. Some of the CE-based limitations may be overcome with the application of MPS ... generate reliable STR profiles at a sensitivity level that competes with current widely used CE-based method.”

Zeng XP, King JL, Stoljarova M, Warshauer DH, LaRue BL, Sajantila A, Patel J, Storts DR, Budowle B. 2015. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Science International: Genetics 16:38-47.

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.

AMD SNP Data

SNP	Individual														
rs6424140	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
rs1496555	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2
rs1338382	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
rs10492936	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
rs10489589	3	1	1	1	2	2	1	2	1	1	1	3	1	1	1
rs10489588	3	1	1	1	2	2	1	2	1	1	1	3	1	1	1
rs4472706	1	3	3	3	2	2	3	2	3	3	3	1	3	3	3
rs4587514	3	3	3	3	3	2	2	3	2	2	2	3	3	1	3
rs10492941	3	3	3	3	3	3	3	3	2	3	3	2	3	3	1
rs1112213	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
rs4648462	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
rs2455122	2	1	1	0	1	2	1	1	1	1	1	1	1	1	2
rs2455124	2	1	1	2	1	2	1	1	1	1	1	1	1	1	2
rs10492940	2	1	1	1	1	2	1	2	1	1	1	2	1	1	2
rs10492939	1	2	1	1	1	1	3	2	1	2	3	2	2	1	1
rs10492938	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
rs10492937	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3
rs7546189	1	2	3	3	1	3	2	2	3	3	2	2	2	2	2
rs1128474	3	2	3	2	3	3	2	3	3	3	3	3	2	1	3

Genotype key: 0 –; 1 AA; 2 AB; 3 BB.

Phase 3 1000Genomes Data

- 84.4 million variants

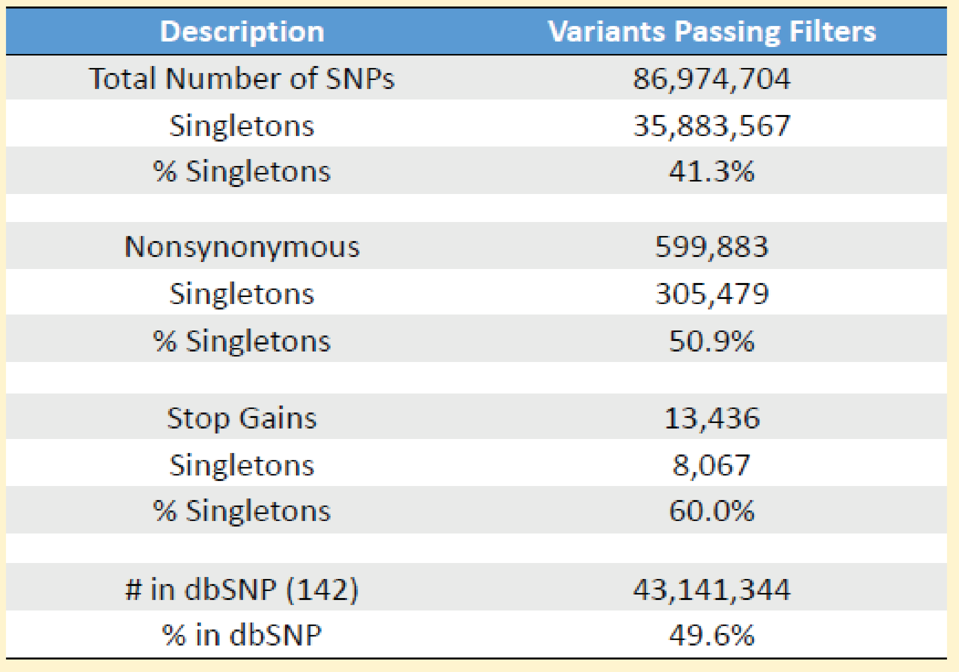
- 2504 individuals

- 26 populations

www.1000Genomes.org

Whole-genome Sequence Studies

One current study is the NHLBI Trans-Omics for Precision Medicine (TOPMed) project. www.nhlbiwgs.org
In the first data freeze of Phase 1 of this study:



Description	Variants Passing Filters
Total Number of SNPs	86,974,704
Singletons	35,883,567
% Singletons	41.3%
Nonsynonymous	599,883
Singletons	305,479
% Singletons	50.9%
Stop Gains	13,436
Singletons	8,067
% Singletons	60.0%
# in dbSNP (142)	43,141,344
% in dbSNP	49.6%

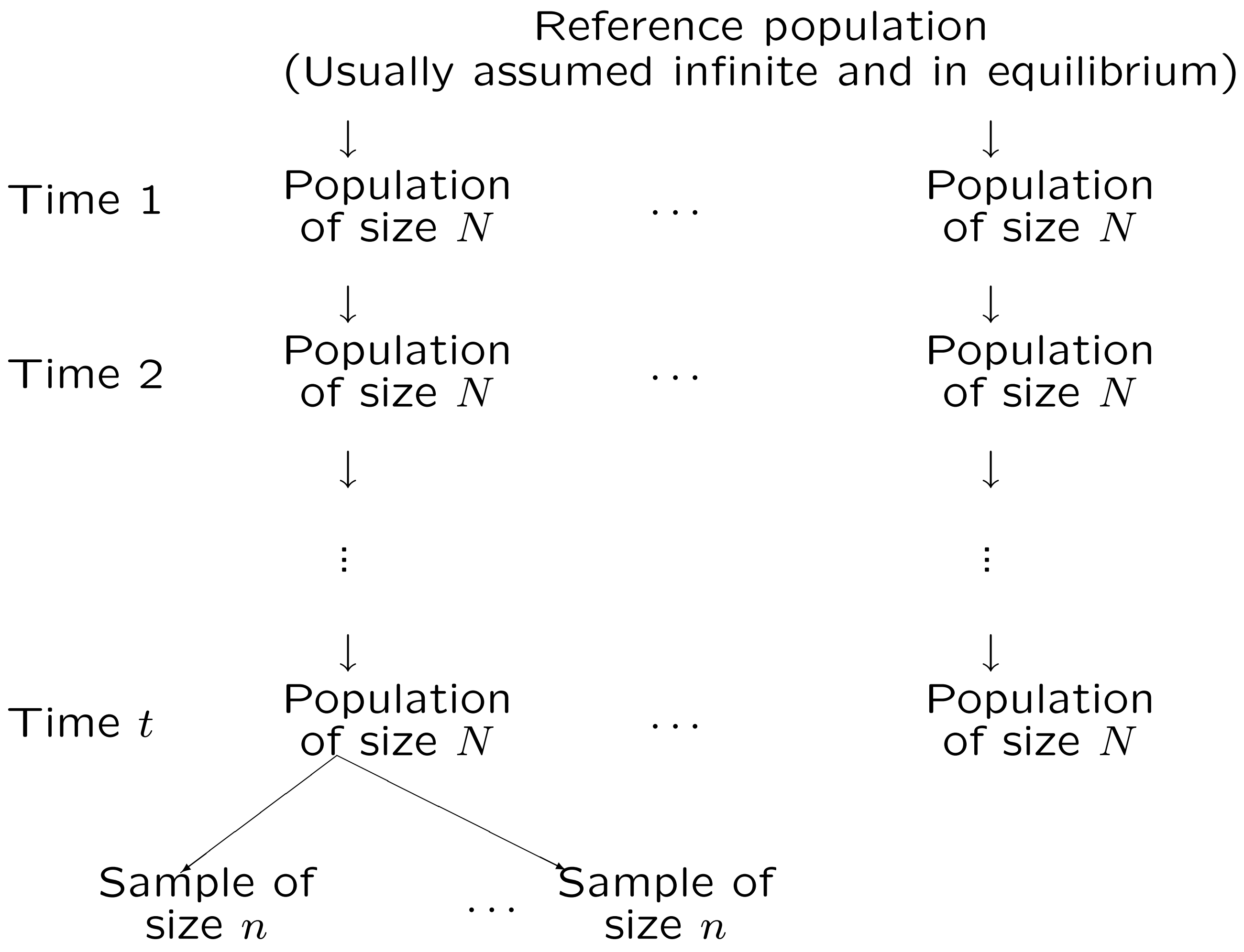
Abecasis et al. 2016. ASHG Poster. Currently 400 million SNVs found from 73,000 whole-genome sequences.

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



Reference population
(Usually assumed infinite and in equilibrium)

Time 1 — Population of size N … Population of size N

Time 2 — Population of size N … Population of size N

$\vdots$

Time t — Population of size N … Population of size N

Sample of size n … Sample of size n

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.

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

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

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

Second Law of Probability

If G and H are mutually exclusive events, then:

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

For example,

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

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

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

Complementary Probability

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

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

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

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

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:

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

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

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

For example

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

Independent Events

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

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

and

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

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

$$\text{Pr}(E) \;=\; \text{Pr}(E|G)\,\text{Pr}(G) + \text{Pr}(E|H)\,\text{Pr}(H)$$

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

$$\text{Pr}(E) \;=\; \sum_i \text{Pr}(E|S_i)\,\text{Pr}(S_i)$$

For example

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

Bayes' Theorem

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

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

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

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

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:

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

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

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

F_2 generation

What are the proportions of AA and Aa among F_2 plants with round seeds? From Bayes' Theorem the predicted probability of AA genotype, if the seed is round, is

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

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

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

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. Mendel's experiments were "a carefully planned demonstration of his conclusions."

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."

"Run Mendel's experiments again at the same scale, Weldon reckoned, and the chance of getting worse results is 16 to 1." Radick, Science 350:159-160, 2015.

Edwards' 1986 Criticism

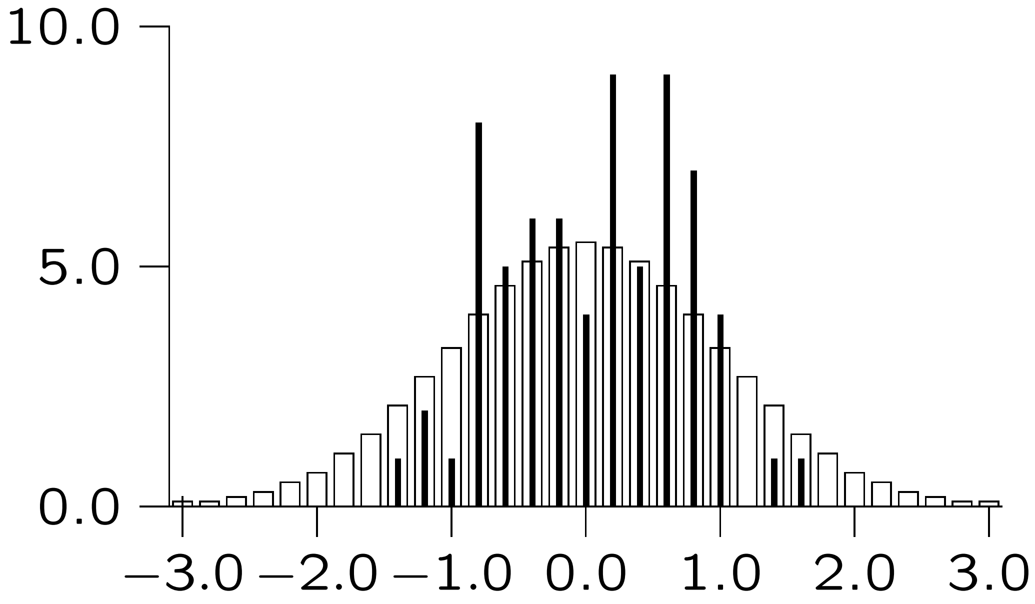
Mendel had 69 comparisons where the expected ratios were correct. Each set of data can be tested with a chi-square test:

		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.



Recent Discussions

Franklin A, Edwards AWF, Fairbanks DJ, Hartl DL, Seidenfeld T. “Ending the Mendel-Fisher Controversy.” University of Pittsburgh Press, Pittsburgh.

Smith MU, Gericke NM. 2015. Mendel in the modern classroom. Science and Education 24:151-172.

Radick G. 2015. Beyond the “Mendel-Fisher controversy.” Science 350:159-160.

Weeden NF. 2016. Are Mendel’s Data Reliable? The Perspective of a Pea Geneticist. Journal of Heredity 107:635-646. “Mendel’s article is probably best regarded as his attempt to present his model in a simple and convincing format with a minimum of additional details that might obscure his message.”

ALLELE FREQUENCIES

Properties of Estimators

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

Binomial Distribution

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

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

Binomial distribution

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

Probability of x heads in a row is

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

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

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

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

The binomial probability of x successes in n trials is

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

Binomial Likelihood

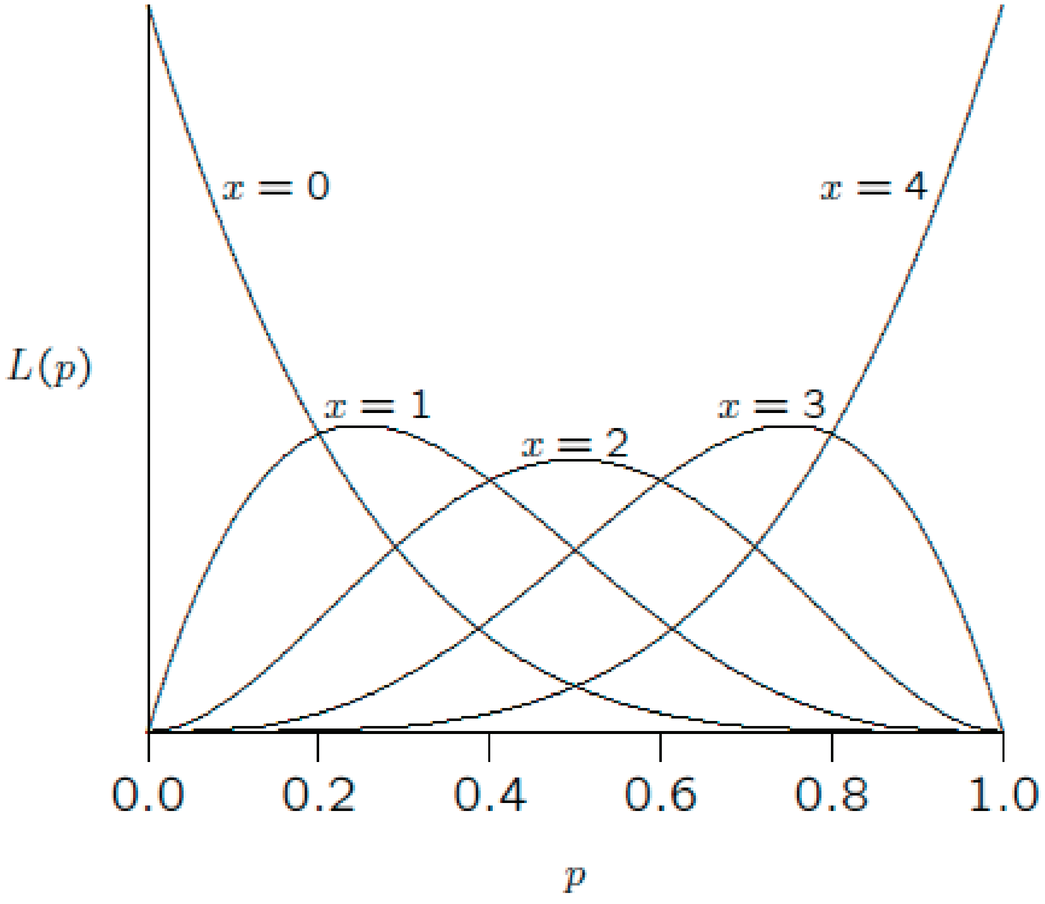
The quantity Pr($x|p$) is the *probability of the data*, x successes in n trials, when each trial has probability p of success.

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

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

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 endpoints which vary with the sample. Expect that 95% of samples will lead to an interval that includes the unknown true value p.

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:

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

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

Confidence Intervals

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

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

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

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

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

Exact Confidence Intervals: One-sided

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

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

$$\begin{aligned} \sum_{k=0}^{x} \mathrm{Pr}(k) &\geq 0.05 \\ \sum_{k=0}^{x} \binom{n}{k} p_U^k (1-p_U)^{n-k} &\geq 0.05 \end{aligned}$$

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

Exact Confidence Intervals: Two-sided

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

$$\begin{aligned} \sum_{k=0}^{x} \binom{n}{k} p_U^k (1-p_U)^{n-k} &\geq 0.025 \\ \sum_{k=x}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} &\geq 0.025 \end{aligned}$$

If $x = 0$, then $(1-p_U) \geq 0.025^{1/n}$ and this gives $p_U \leq 0.036$ when $n = 100$.
If $x = n$, then $p_L \geq 0.975^{1/n}$ and this gives $p_L \geq 0.964$ when $n = 100$.

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:

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

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

For the sample counts:

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

Allele Frequency Sampling Distribution

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

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

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

Within-population Variance

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

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

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

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

Within-population Variance

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

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

Then the variance can be written as

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

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

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

These expectations are the averages of values from many samples from the same population.

Intraclass Correlation

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

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

and

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

so

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

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?

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

Maximum Likelihood Estimation

The likelihood function for p_A, f is

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

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

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

Bailey's Method

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

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

Solving these equations (e.g. by adding the first equation to half the second equation to give solution for $\hat{p}_A$ and then substituting that into one equation):

$$\begin{aligned} \hat{p}_A &= \frac{2n_{AA}+n_{Aa}}{2n} = \tilde{p}_A \\ \hat{f} &= 1 - \frac{n_{Aa}}{2n\tilde{p}_A(1-\tilde{p}_A)} = 1 - \frac{\tilde{P}_{Aa}}{2\tilde{p}_A\tilde{p}_a} \end{aligned}$$

Three-allele Case

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

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

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

Method of Moments

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

Method of Moments

For the inbreeding coefficient at loci with m alleles, two different MoM estimates are (for large sample sizes)

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

These both have low bias. Their variances depend on the value of f.

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

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

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) as proportions of the total count of dominant phenotypes:

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

EM Algorithm for Recessive Alleles

Maximize the likelihood (using Bailey's method):

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

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

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

EM Algorithm for Two Loci

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

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

EM Algorithm for Two Loci

Gamete frequencies are marginal sums:

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

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

$$\begin{array}{cc|c}
p_{AB} & p_{Ab} & p_A \\
p_{aB} & p_{ab} & p_a \\
\hline
p_B & p_b & 1
\end{array}$$

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 as proportions of the total count of double heterozygotes:

$$\begin{aligned} n_{ab}^{AB} &= \frac{2p_{AB}p_{ab}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} n_{AaBb} \\ n_{aB}^{Ab} &= \frac{2p_{Ab}p_{aB}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} n_{AaBb} \end{aligned}$$

and then *Maximize* the likelihood by setting

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

Example

As an example, consider the data

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

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

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

$$21 \geq x \geq 16$$

Example

EM iterative equation:

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

In this case note that if $x = 16$ then $x' = 16$ so this is the MLE.

Example

If we did not recognize the solution, a good starting value would assume independence of A and B alleles: $x = 2n * p_A * p_B = (25 \times 21/30) = 17.5$.

Successive iterates are:

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

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

ALLELIC ASSOCIATION

Hardy-Weinberg Law

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

For a locus with two alleles, A, a:

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

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

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

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, $(\hat{f}-f)/\sqrt{\text{Var}(\hat{f})} \sim N(0,1)$. When H_0 is true, the square of this quantity has a chi-square distribution.

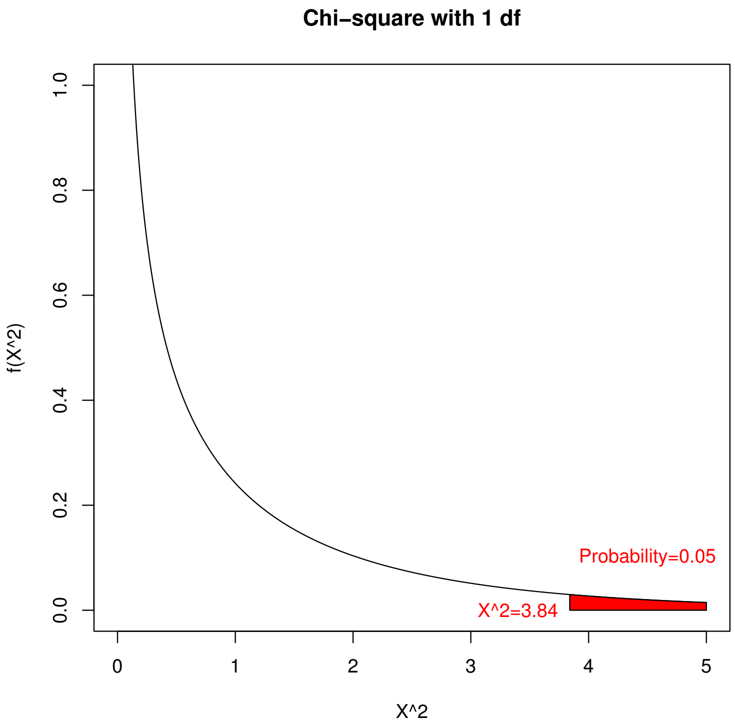
Inference about HWE

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

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

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

Significance level of HWE test



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

Goodness-of-fit Test

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

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

The test statistic is

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

Goodness-of-fit Test

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

First need to estimate allele frequencies:

$$\begin{aligned}\tilde{p}_A &= \tilde{P}_{AA} + \frac{1}{2}\tilde{P}_{Aa} = 0.75 \\ \tilde{p}_a &= \tilde{P}_{aa} + \frac{1}{2}\tilde{P}_{Aa} = 0.25\end{aligned}$$

Then form "expected" counts:

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

Goodness-of-fit Test

Perform the chi-square test:

Genotype	Observed	Expected	(Obs. – Exp.)^2/Exp.
AA	6	5.625	0.025
Aa	3	3.750	0.150
aa	1	0.625	0.225
Total	10	10	0.400

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

Sample size determination

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

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

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

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

Sample size determination

If the hypothesis is false, the normal test statistic is

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

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

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

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

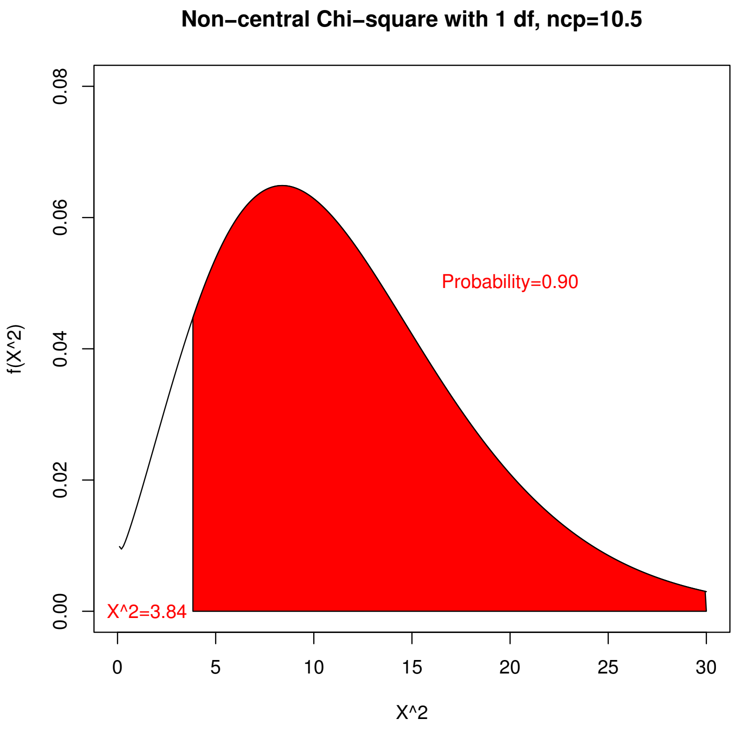
Sample size determination

More directly, when the Hardy-Weinberg hypothesis is not true, the test statistic $n\widehat{f}^2$ 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 $1 - \beta$, $\lambda = (z_{\alpha/2} + z_\beta)^2$.

```
> pchisq(3.84,1,0)
[1] 0.9499565
> pchisq(3.84,1,10.5)
[1] 0.1001356
> qchisq(0.95,1,0)
[1] 3.841459
```

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 $\lambda = 10.5$.

Population Structure: Departures from HWE

If a population consists of a number of subpopulations, each in HWE but with different allele frequencies, there will be a departure from HWE at the population level. This is the Wahlund effect.

Suppose there are two equal-sized subpopulations, each in HWE but with different allele frequencies, then

	Subpopn 1	Subpopn 2	Total Popn
p_A	0.6	0.4	0.5
p_a	0.4	0.6	0.5
P_{AA}	0.36	0.16	$0.26 > (0.5)^2$
P_{Aa}	0.48	0.48	$0.48 < 2(0.5)(0.5)$
P_{aa}	0.16	0.36	$0.26 > (0.5)^2$

Population Admixture: Departures from HWE

A population might represent the recent admixture of two parental populations. With the same two populations as before but now with 1/4 of marriages within population 1, 1/2 of marriages between populations 1 and 2, and 1/4 of marriages within population 2. If children with one or two parents in population 1 are considered as belonging to population 1, there is an excess of heterozygosity in the offspring population.

If the proportions of marriages within populations 1 and 2 are both 25% and the proportion between populations 1 and 2 is 50%, the next generation has

	Population 1	Population 2
P_{AA}	$0.09 + 0.12 = 0.21$	0.04
P_{Aa}	$0.12 + 0.26 = 0.38$	0.12
P_{aa}	$0.04 + 0.12 = 0.16$	0.09
	0.75	0.25

Population 2 is in HWE, but Population 2 has 51% heterozygotes instead of the expected 49.8%.

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:

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

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:

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

Exact HWE Test

Putting these together gives the conditional probability

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

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

Exact HWE Test Example

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 Example

As another example, the sample with $n_{AA} = 6, n_{Aa} = 3, n_{aa} = 1$ has allele counts $n_a = 15, n_a = 5$. There are two other sets of genotype counts possible and the probabilities of each set for a HWE population are:

n_{AA}	n_{Aa}	n_{aa}	n_A	n_a	$\Pr(n_{AA}, n_{Aa}, n_{aa} \mid n_A, n_a)$
5	5	0	15	5	$\frac{10!}{5!5!0!}\frac{2^5 15!5!}{20!} = \frac{168}{323} = 0.520$
6	3	1	15	5	$\frac{10!}{6!3!1!}\frac{2^3 15!5!}{20!} = \frac{140}{323} = 0.433$
7	1	2	15	5	$\frac{10!}{7!1!2!}\frac{2^1 15!5!}{20!} = \frac{15}{323} = 0.047$

Compare with chi-square p-value for $X^2 = 0.40$:

```
> pchisq(0.4,1)
[1] 0.4729107
>
```

Exact HWE Test 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	0.0655	5.38	0.0204*
87	12	1	0.3209	0.3864	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 0.0053 for an exact test that rejects when $n_{Aa} \leq 8$ and is 0.0204 for an exact test that rejects when $n_{Aa} \leq 10$.

Modified Exact HWE Test

Traditionally, the p-value is the (conditional) probability of the data plus the probabilities of all the less-probable datasets. The probabilities are all calculated assuming HWE is true. More recently (Graffelman and Moreno, Statistical Applications in Genetics and Molecular Biology 12:433-448, 2013) it has been shown that the test has a significance value closer to the nominal value if the p-value is half the probability of the data plus the probabilities of all datasets that are less probably under the null hypothesis. For the $(n_{AA} = 1, n_{Aa} = 0, n_{aa} = 49)$ example then, the p-value is 1/198.

Graffelman and Moreno, 2013

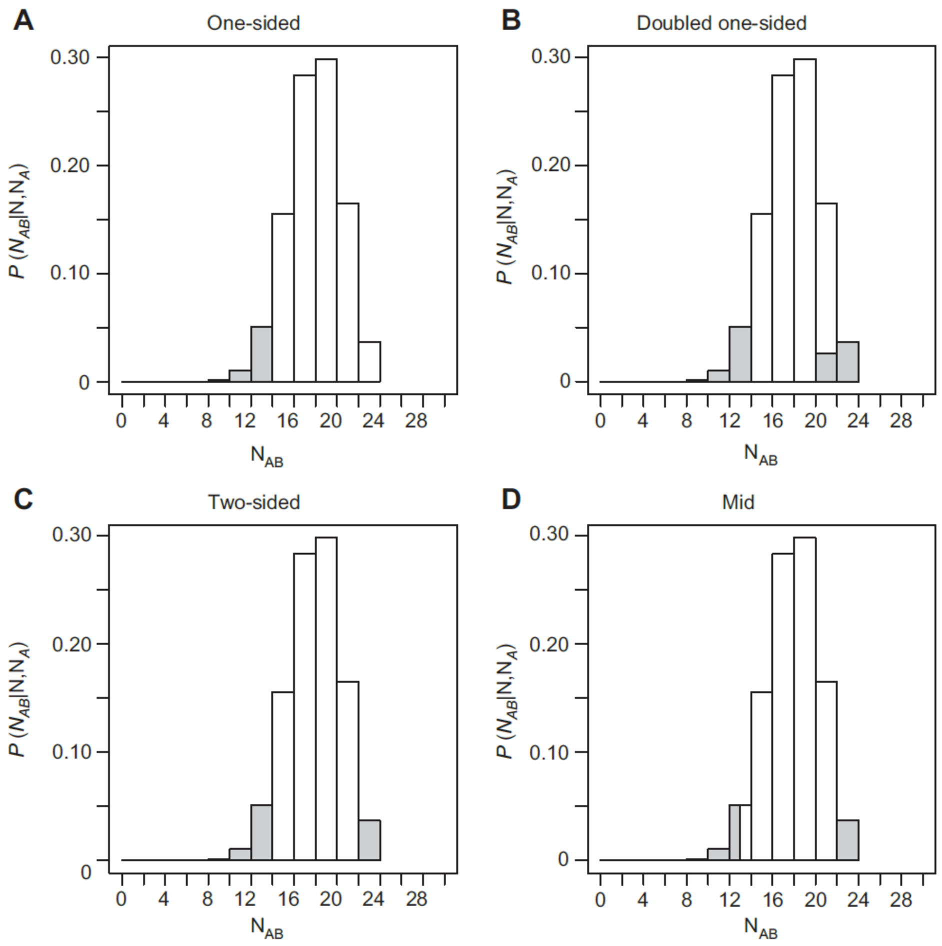


Figure 1 Computation of the *p*-value in an exact test for HWP, for a sample of 50 individuals with a minor allele count of 23, for which 13 heterozygotes were observed. (A) One-sided *p*-value in a test for heterozygote dearth. (B) *p*-value obtained by doubling the one-sided tail. (C) Standard two-sided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

Usual vs Mid p values

				p−value	
AA	Aa	aa	$\Pr(n_{Aa} \mid n, n_A)$	Usual	Mid
1	0	49	$\frac{50!}{1!0!49!}\frac{2^0 2!98!}{100!} = \frac{1}{99}$	$\frac{1}{99}$	$\frac{1}{198}$
0	2	48	$\frac{50!}{0!2!48!}\frac{2^2 2!98!}{100!} = \frac{98}{99}$	1	$\frac{50}{99}$
Average				0.99	0.50

Usual vs Mid p values

				p−value	
AA	Aa	aa	Pr$(n_{Aa} \mid n, n_A)$	Usual	Mid
5	5	0	0.520	1.000	0.740
6	3	1	0.433	0.480	0.287
7	1	2	0.047	0.047	0.023
Average				0.730	0.510

Modified Exact HWE Test 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.	Mid p-value	X^2	p-value
93	0	7	0.0000	0.0000*	100.00	0.0000*
92	2	6	0.0000	0.0000*	71.64	0.0000*
91	4	5	0.0000	0.0000*	47.99	0.0000*
90	6	4	0.0002	0.0002*	29.07	0.0000*
89	8	3	0.0051	0.0028*	14.87	0.0001*
88	10	2	0.0602	0.0353*	5.38	0.0204*
87	12	1	0.3209	0.2262	0.61	0.4348
86	14	0	0.6136	0.6832	0.57	0.4503

So, for a nominal 5% significance level, the actual significance level is 0.0353 for an exact test that rejects when $n_{Aa} \leq 10$ and is 0.0204 for a chi-square test that also rejects when $n_{Aa} \leq 10$.

Effect of Minor Allele Frequency

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

In particular, what is the size of the 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 0.0070 for an exact test that rejects when $n_{Aa} \leq 10$.

n_{AA}	n_{Aa}	n_{aa}	$\Pr(n_{Aa} \mid n_a)$	mid p −value
92	0	8	.0000	.0000
91	2	7	.0000	.0000
90	4	6	.0000	.0000
89	6	5	.0000	.0000
88	8	4	.0008	.0004
87	10	3	.0123	.0070
86	12	2	.0974	.0618
85	14	1	.3681	.2946
84	16	0	.5215	.7382

$n_a = 15$ minor alleles

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

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

$n_a = 13$ minor alleles

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

n_{AA}	n_{Aa}	n_{aa}	Pr($n_{Aa}|n_a$)	mid p −value
93	1	6	.0000	.0000
92	3	5	.0000	.0000
91	5	4	.0001	.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$(n_{Aa}\|n_a)$	mid 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	.0181
89	10	1	.2612	.1650
88	12	0	.7045	.2955

Graffelman and Moreno, 2013

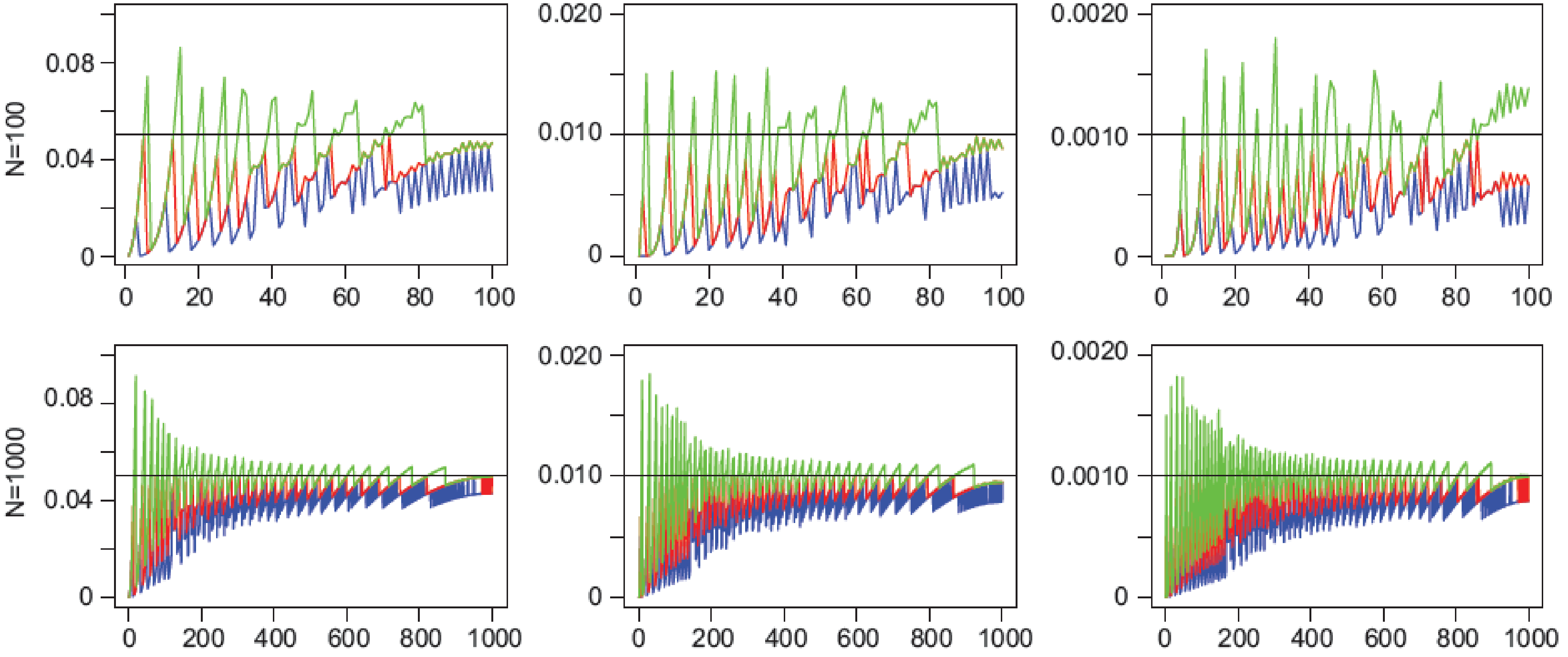


Figure 2 Type I error rate against minor allele count for different sample sizes (25, 50, 100 and 1000) and significance levels (0.05, 0.01, and 0.001) for exact tests with standard two-sided (red), doubled one-sided (blue) and mid *p*-values (green).

Power of Exact Test

If there is not HWE:

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

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 $C = 1/[1/(5!5!0!) + 1/(6!3!1!) + 1/(7!1!2!)]$.

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

The $\psi = 2$ column 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($n_{Aa} \| n_a = 16$)						
	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

		$\Pr(n_{Aa} \vert n_a = 14)$						
	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$(n_{Aa}\|n_a = 13)$						
	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($n_{Aa}\|n_a = 12$)						
	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

Graffelman and Moreno, 2013

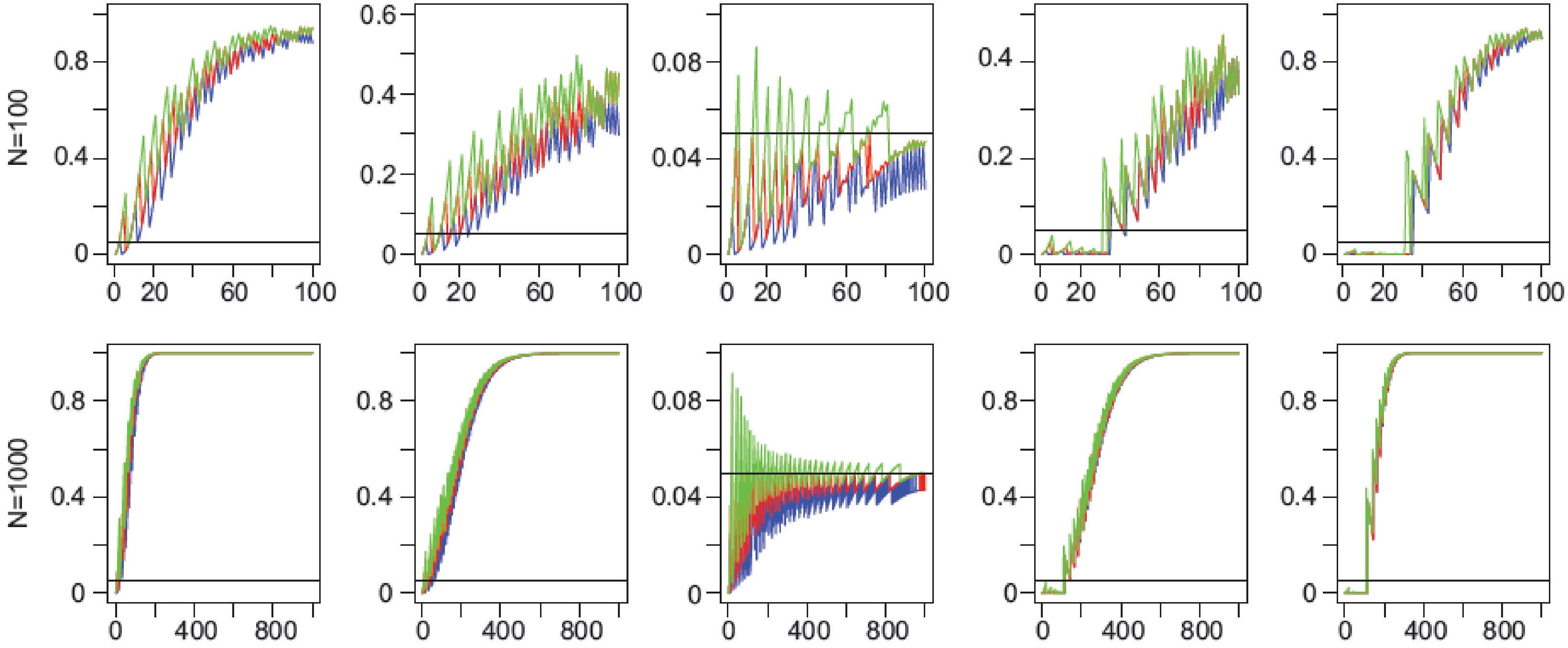


Figure 3 Power of HWP exact tests against minor allele count for different sample sizes (25, 50, 100 and 1000) and degree of disequilibrium (1, 2, 4, 8 and 16). Standard two-sided (red), doubled one-sided (blue) and mid *p*-values (green).

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.

This procedure is not needed for SNPs with only 2 alleles. The number of possible arrays is always less than bout half the sample size.

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.

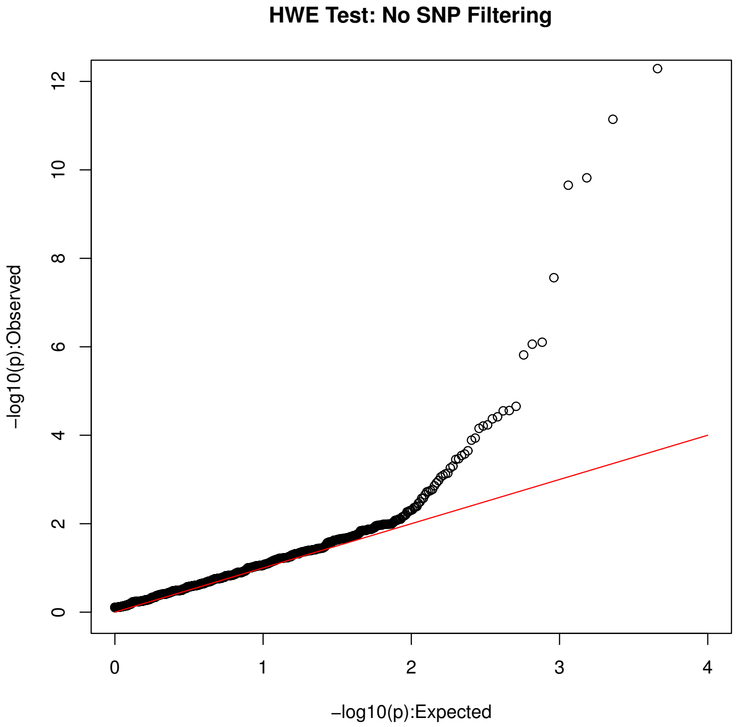
If $L = 15$, need $\alpha' = 0.0033$ in order for $\alpha = 0.05$.

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.

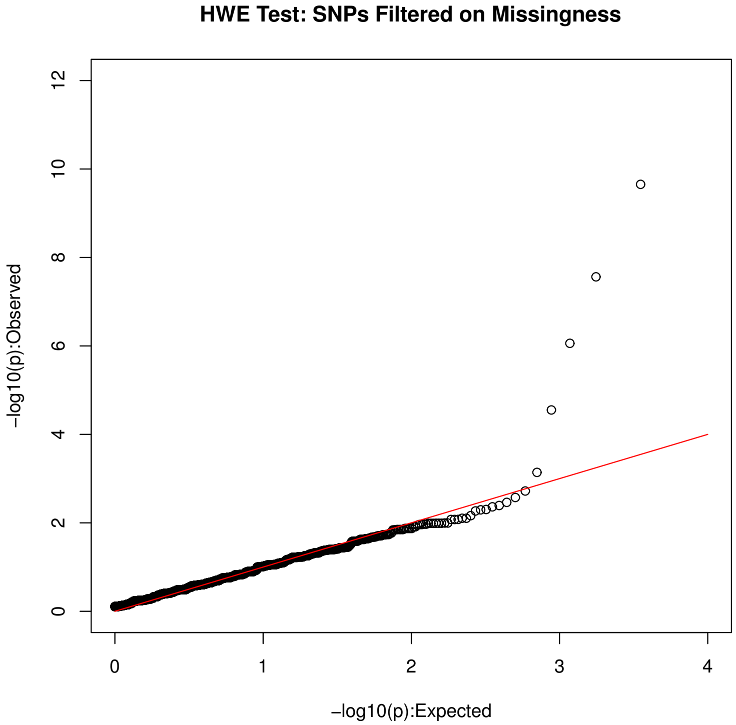
For a set of n tests, we would expect to see n evenly spread p values between 0 an 1 e.g. $1/n, 2/n, \ldots, n/n$. We plot the observed p-values against these expected values: the smallest against $1/n$ and the largest against 1. It is more convenient to transform to $-\log_{10}(p)$ to accentuate the 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.

Imputing Missing Data

Instead of discarding an individual for any SNP when there is no genotype call, it may be preferable to use neighboring SNPs to impute the missing values. This procedure has been applied to a study on pre-term birth (Graffelman et al., 2015, G3 (Genes, Genomes, Genetics) 5:2365-2373).

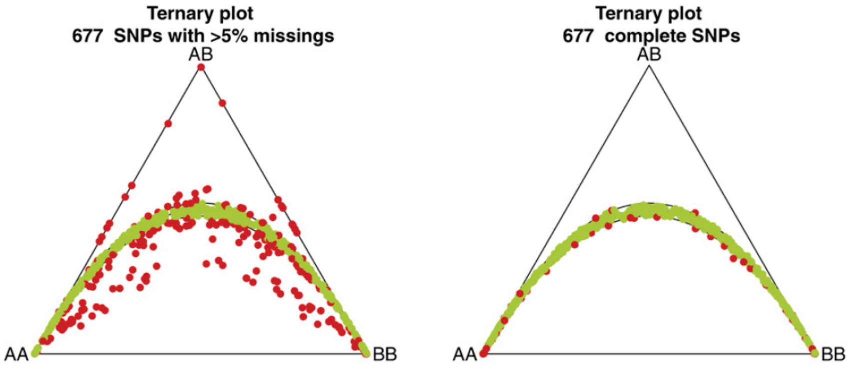
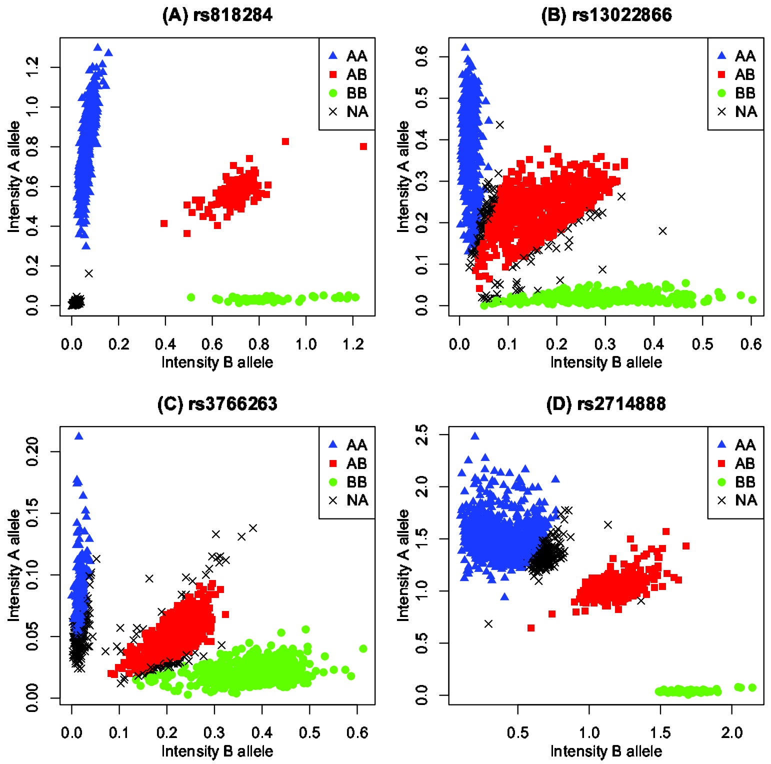


Figure 1 Left panel: ternary plot for 677 single-nucleotide polymorphisms (SNPs) with >5% missing. A total of 229 SNPs (34%) are significant in a χ^2 test. Right panel: 677 SNPs without missings taken at random. A total of 56 (8%) SNPs are significant in a χ^2 test. Significant markers are red and nonsignificant markers are green (α = 0.05).

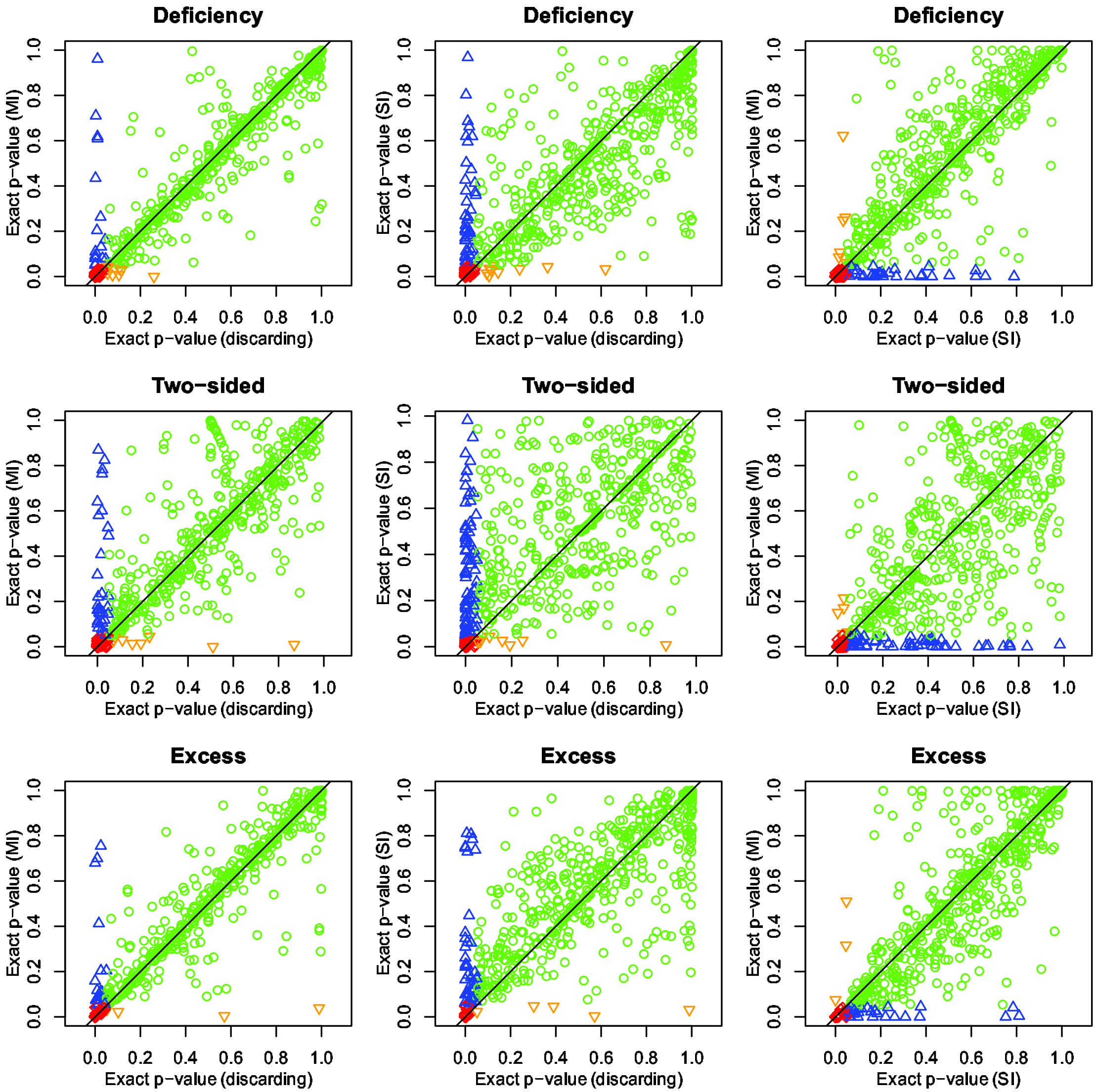
DeFinetti diagram: distance of point to side of triangle is frequency of genotype shown on opposite vertex.

Imputing Missing Data



SNP	Discard	Impute	Comment
rs818284	0.000	0.000	Null alleles
rs13022866	0.046	0.571	Het deficiency
rs3766263	0.020	0.539	Het excess
rs2714888	0.192	0.007	Hom deficiency

Graffelman et al., 2015



HWE Test for X-linked Markers

Under HWE, allele frequencies in males and females should be the same. Best to examine the difference when testing for HWE.

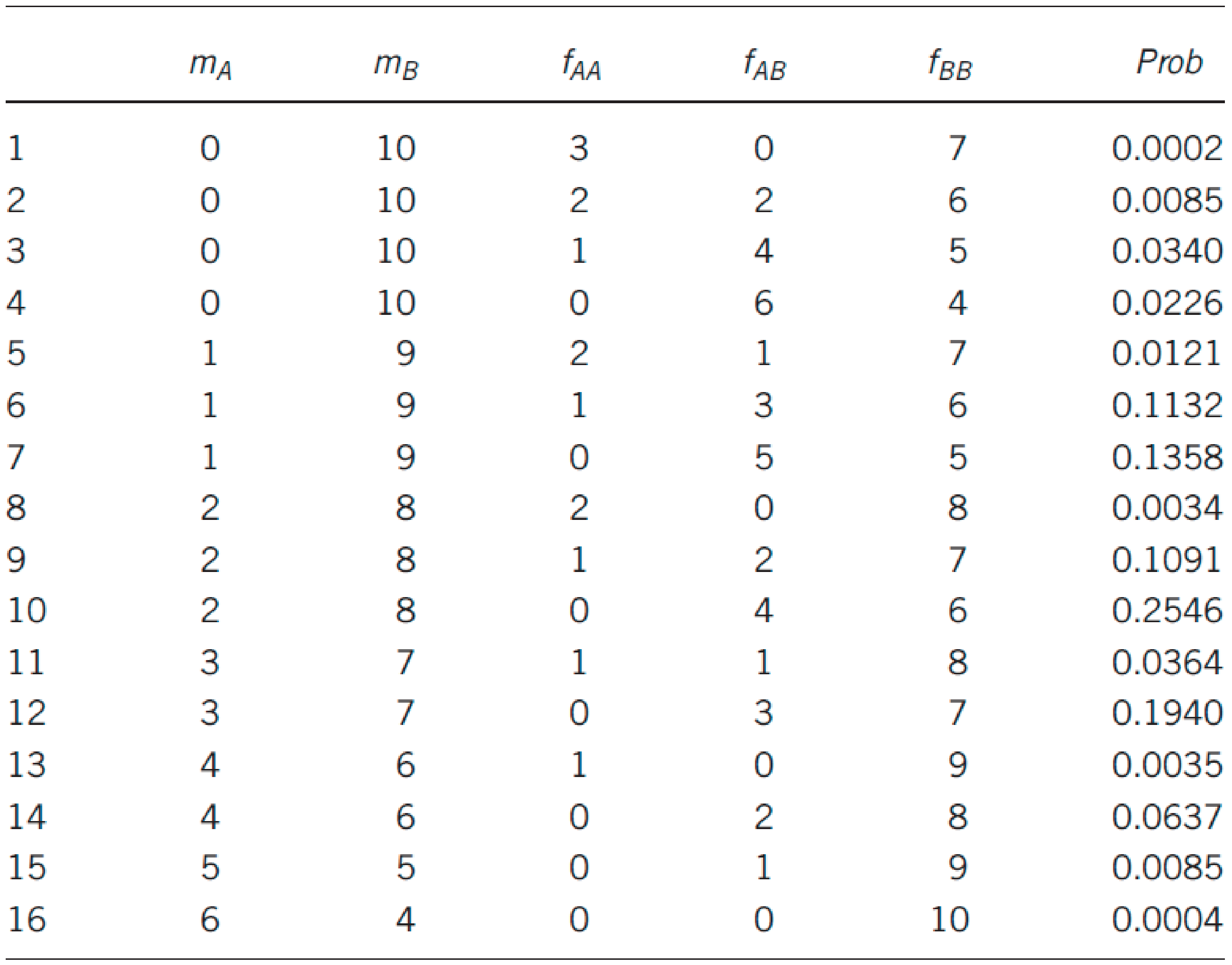
If a sample has n_m males and n_f females, and if the males have m_A, m_B alleles of types A, B, and if females have f_{AA}, f_{AB}, f_{BB} genotypes AA, AB, BB, then the probability of the data, under HWE, is

$$\frac{n_A!n_B!n_m!n_f!}{m_A!m_B!f_{AA}!f_{AB}!f_{BB}!n_t!}2^{f_{AB}}$$

where $n_t = n_m + 2n_f$.

(Graffelman and Weir, 2016, Heredity 116:558-568).

Example: 10 males, 10 females, 6 A alleles



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

Graffelman and Weir, 2016. Possible Scenarios

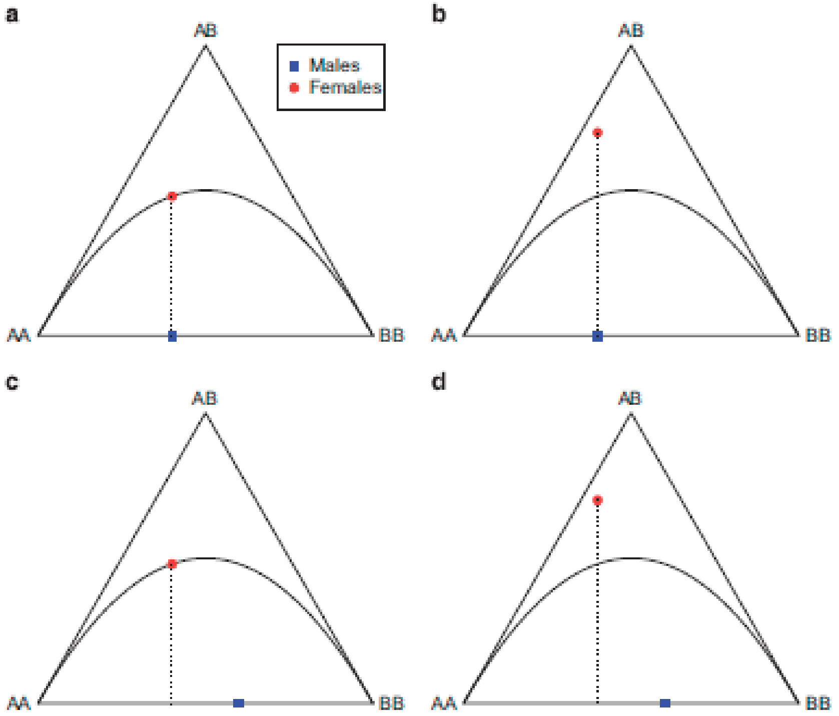


Figure 2 Hardy–Weinberg (dis)equilibrium for a biallelic marker on the X chromosome. **(a)** Equilibrium. **(b)** Disequilibrium due to deviating female genotype frequencies. **(c)** Disequilibrium due to nonhomogeneous allele frequencies. **(d)** Disequilibrium due to deviating female genotype frequencies and nonhomogeneous allele frequencies.

Graffelman and Weir, 2016. Simulated Data

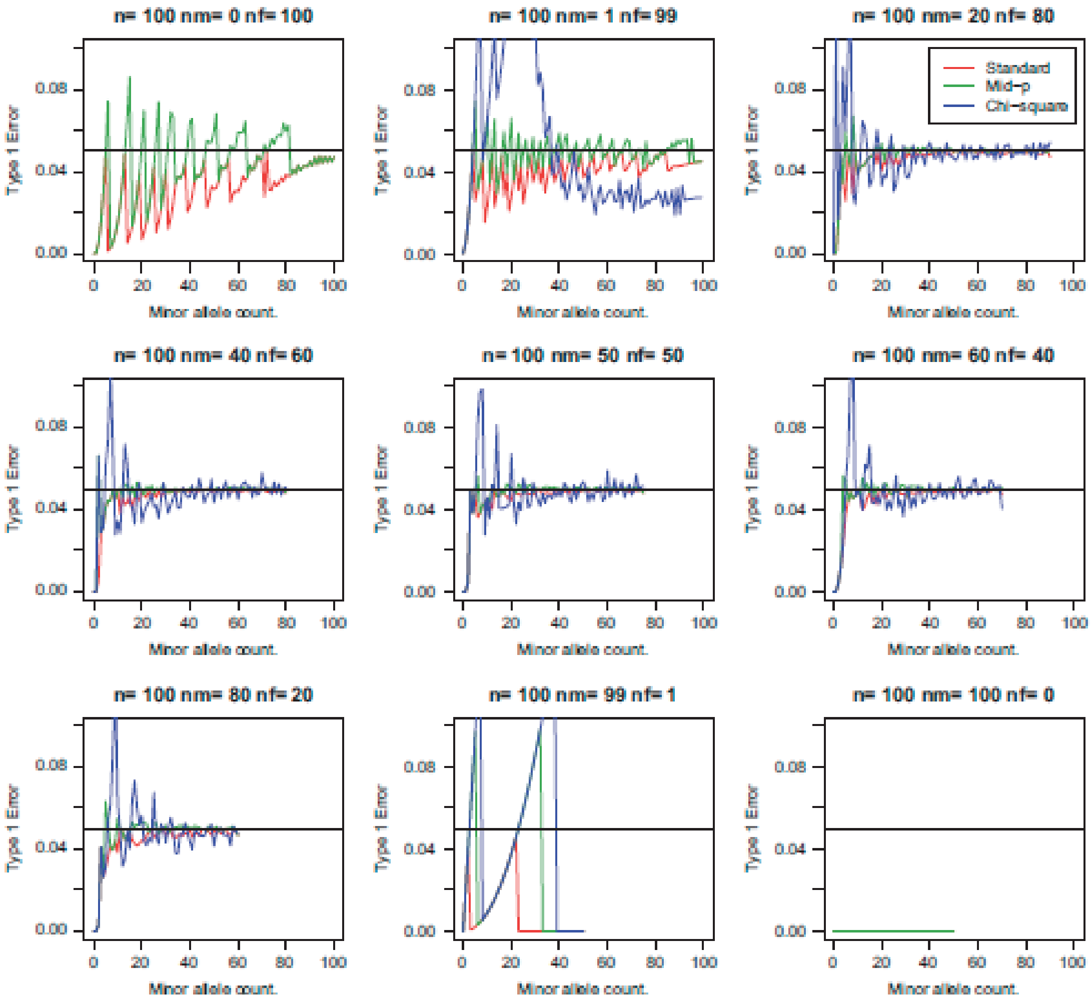


Figure 4 Type I error rate of X-chromosomal tests for HWP as a function of sex ratio and MAF. The Type I error rate of an all-individual test for HWP is plotted for the exact test with the standard *P*-value (red), the exact test with the mid *P*-value (green) and the χ^2 test without continuity correction (blue). *n*, sample size (100); nf, number of females; nm, number of males.

Graffelman and Weir, 2016. Real Data

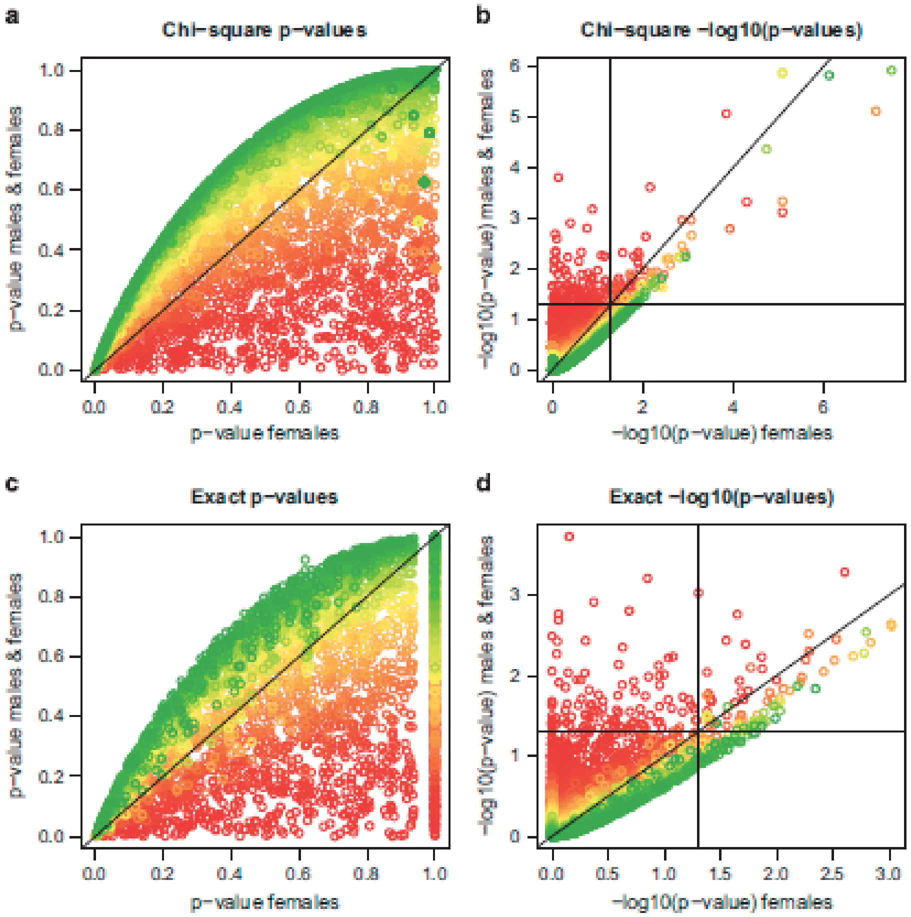


Figure 6 Scatter plots of *P*-values for χ^2 tests and exact tests for HWE using females only and using both males and females for 4158 SNPs at the X chromosome of the venous thrombosis database. The horizontal and vertical black lines in (**b**) and (**d**) corresponds to a significance level of 5%. Points colored according to their significance level in Fisher's test for equality of allele frequencies (range 0–1 from red to green).

Graffelman and Weir, 2016. Real Data

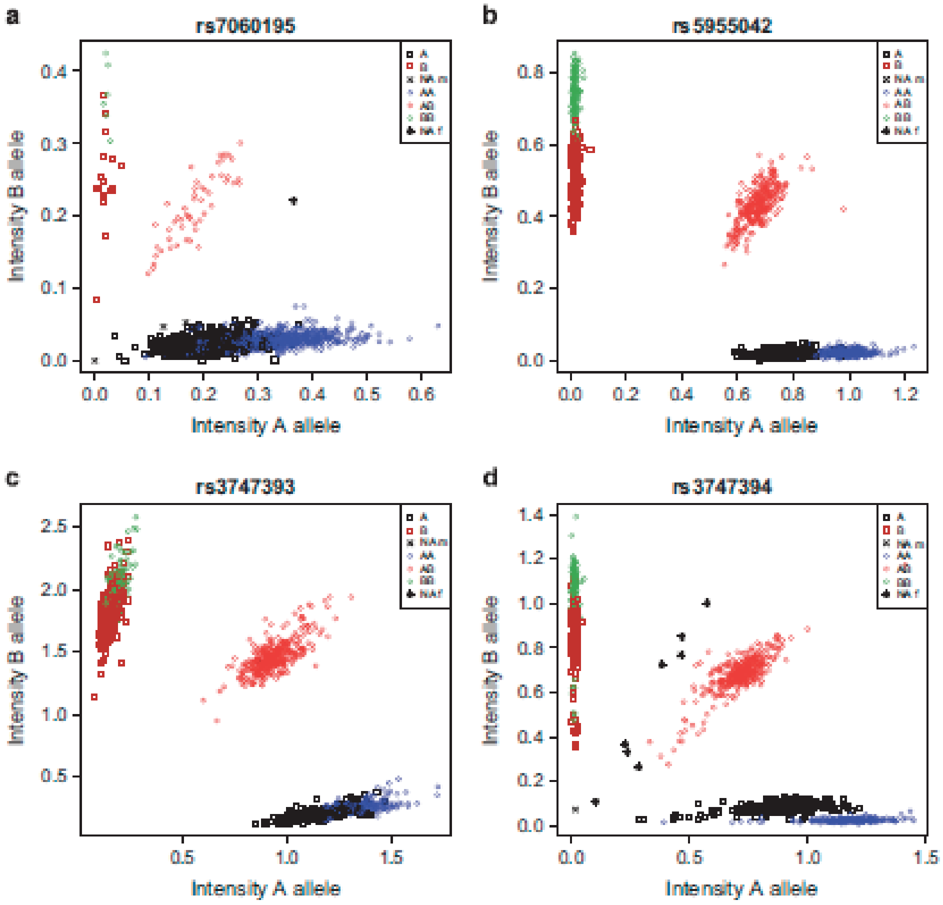


Figure 8 Cluster plots of allele intensities of four SNPs of the venous thrombosis database that are significant (**a**, **b**, **c**) in an all-individual exact test for HWE.

Separate Male and Female Counts

The X-linked test can be extended to autosomal markers when genotype counts are recorded separately for males and females.

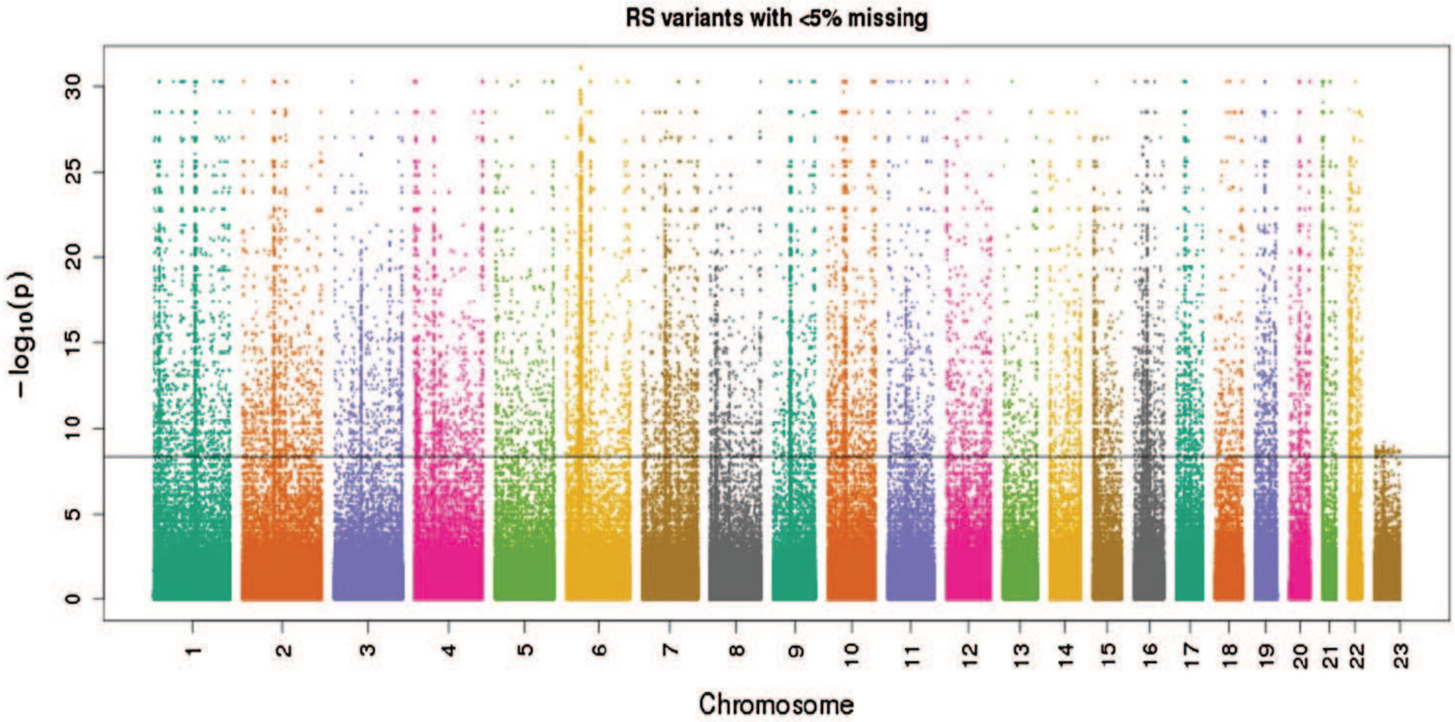
SNV Data

Sequence-based variants have many low-frequency alleles that are susceptible to effects of copy-number variation.

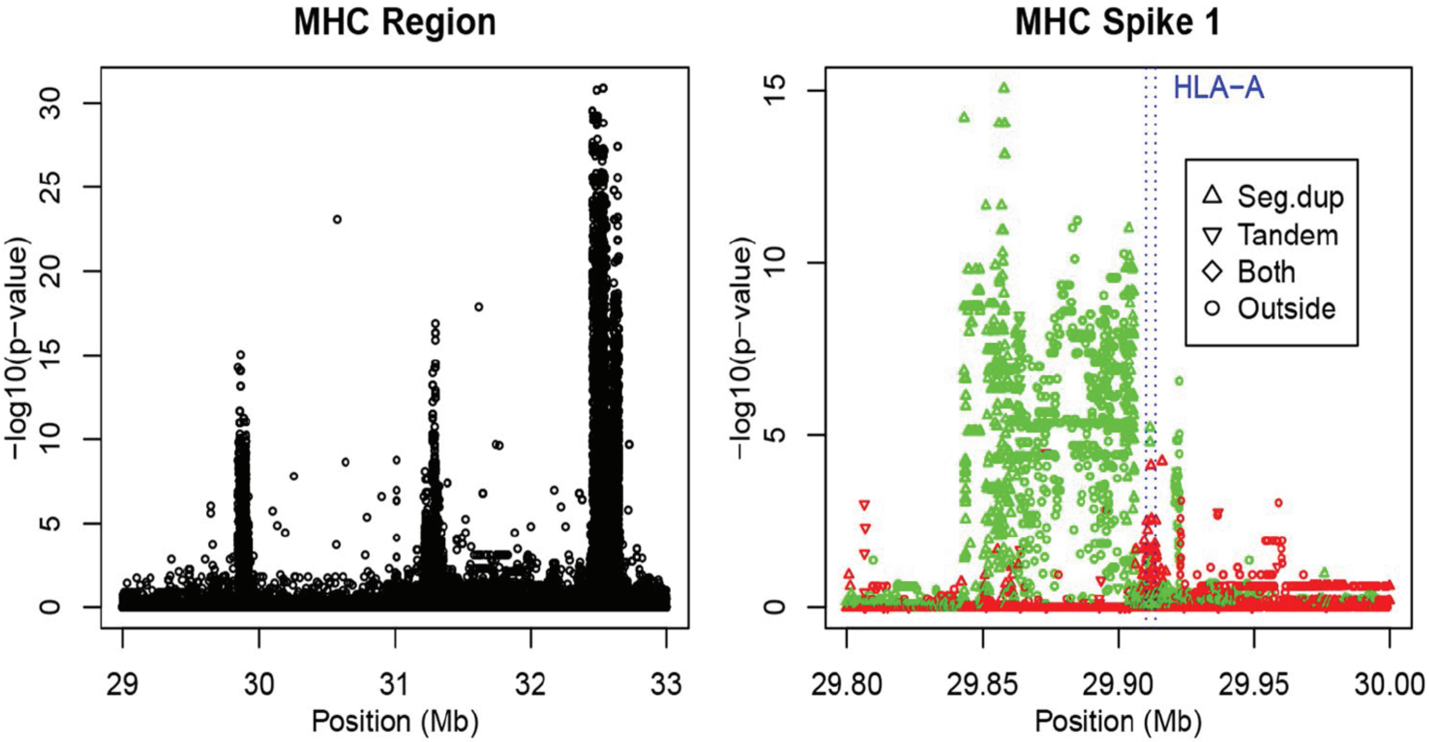
Recent survey of 1000Genomes data revealed more departures from HWE than expected by chance, and many of these reflect an apparent heterozygote excess. SNP-array data often shoe HWE departures from heterozygote deficiency.

(Graffelman et al., 2017. Human Genetics 136:77-741.)

Whole Genome HWE Tests



MHC Region HWE Tests



Green: heterozygote deficiency. Red: heterozygote excess.

Copy Number Variants

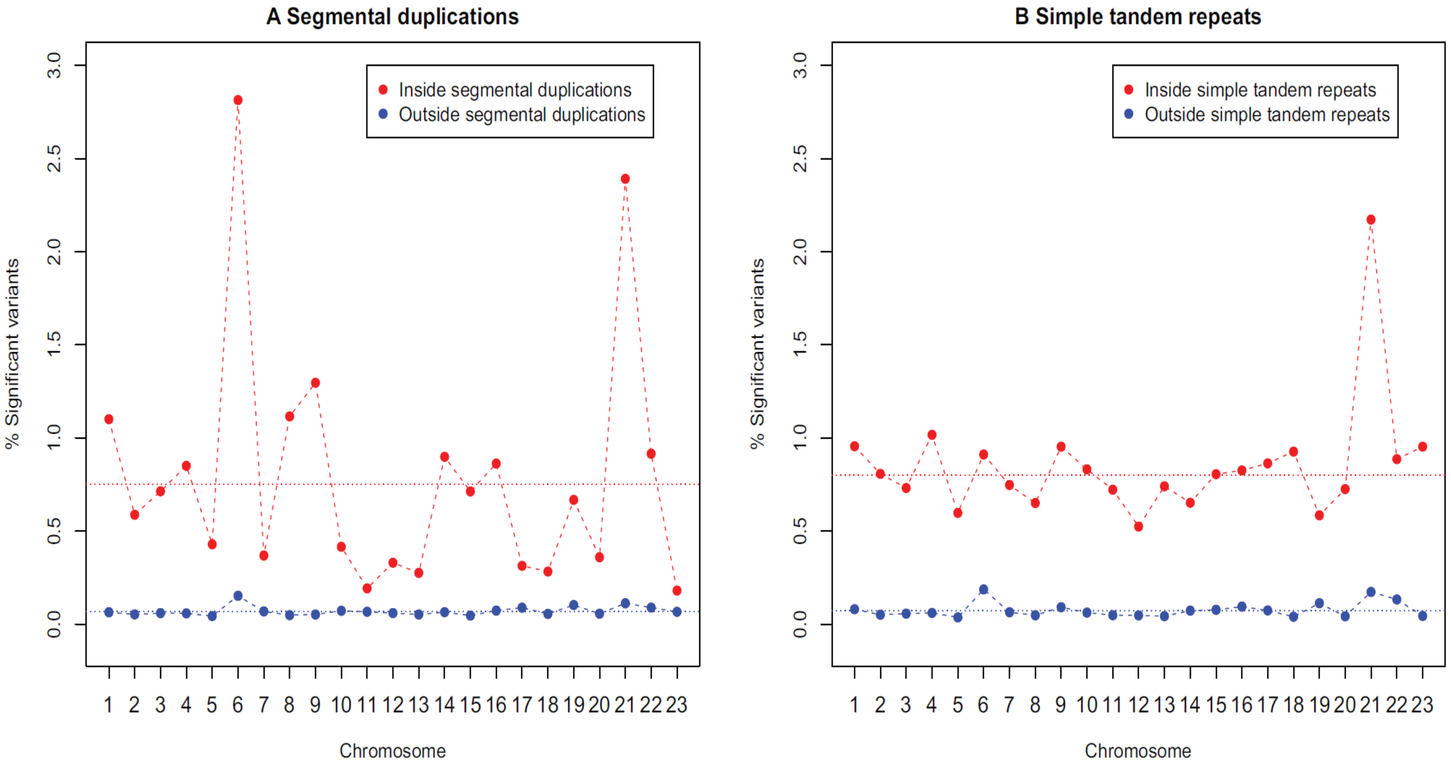
JPT sample has 3.8% of variants in segmental duplications and 3% in simple tandem repeats. HWD rates 11 times higher in these regions: reflecting sequencing problems due to multiple copies of a variant, leading to heterozygote excess.

"Segmental duplications (SDs) are segments of DNA with near-identical sequence.

A microsatellite is a tract of repetitive DNA in which certain DNA motifs (ranging in length from 25 base pairs) are repeated, typically 5-50 times."

[Wikipedia]

Copy Number Variants



SNV-HWE Conclusions

Significant HWE results may indicate copy number variation - excluding them may also exclude disease-associated variants.

NGS data have heterozygote excess, reflecting copy number variation.

SNP array data have heterozygote deficiency, reflecting null alleles.

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

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

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

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

Estimation of LD

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

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

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

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

Testing LD

Write MLE of D_{AB} as

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

and use "Delta method" to find

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

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

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

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

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

Goodness-of-fit Test

The test statistic for the 2×2 table

$$\begin{array}{cc|c} n_{AB} & n_{Ab} & n_A \\ n_{aB} & n_{ab} & n_a \\ \hline n_B & n_b & n \end{array}$$

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

$$\begin{aligned}
\mathcal{E}(x_j) = p_A,\ \ j = 1,2 \quad &, \quad \mathcal{E}(y_j) = p_B \ \ j = 1,2 \\
\mathcal{E}(x_j^2) = p_A,\ \ j = 1,2 \quad &, \quad \mathcal{E}(y_j) = p_B \ \ j = 1,2 \\
\mathcal{E}(x_1x_2) = P_{AA} \quad &, \quad \mathcal{E}(y_1y_2) = P_{BB} \\
\mathcal{E}(x_1y_1) = P_{AB} \quad &, \quad \mathcal{E}(x_2y_2) = P_{AB} \\
\mathcal{E}(x_1y_2) = P_{A,B} \quad &, \quad \mathcal{E}(x_2y_1) = P_{A,B}
\end{aligned}$$

Write

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

Composite Linkage Disequilibrium

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

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

and

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

Composite Linkage Disequilibrium

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

where

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

This does not require phased data.

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

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

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

Example

For the data

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

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

Example

The estimated composite disequilibrium coefficient is

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

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

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

Multi-locus Disequilibria: Entropy

It is difficult to describe associations among alleles at several loci. One approach is based on information theory.

For a locus with sample frequencies $\tilde{p}_u$ for alleles A_u the entropy is

$$H_A = -\sum_u \tilde{p}_u \ln(\tilde{p}_u)$$

For independent loci, entropies are additive: if haplotypes A_uB_v have sample frequencies $\tilde{P}_{uv}$ the two-locus entropy is

$$H_{AB} = -\sum_u \sum_v \tilde{P}_{uv} \ln(\tilde{P}_{uv}) = -\sum_u \sum_v \tilde{p}_u\tilde{p}_v[\ln(\tilde{p}_u) + \ln(\tilde{p}_v)] = H_A + H_B$$

so if $H_{AB} \neq H_A + H_B$ there is evidence of dependence. This extends to multiple loci.

Conditional Entropy

If the entropy for a multi-locus profile A is H_A then the conditional probability of another locus B, given A, is $H_{B|A} = H_{AB} - H_A$.

In performing meaningful calculations for Y-STR profiles, this suggests choosing a set of loci by an iterative procedure. First choose locus L_1 with the highest entropy. Then choose locus L_2 with the largest conditional entropy $H(L_2|L_1)$. Then choose L_3 with the highest conditional entropy with the haplotype L_1L_2, and so on.

Conditional Entropy: YHRD Data

Added	Entropy			Added	Entropy		
Marker	Single	Multi	Cond.	Marker	Single	Multi	Cond.
YS385ab	4.750	4.750	4.750	DYS481	2.962	6.972	2.222
DYS570	2.554	8.447	1.474	DYS576	2.493	9.318	0.871
DYS458	2.220	9.741	0.423	DYS389II	2.329	9.906	0.165
DYS549	1.719	9.999	0.093	DYS635	2.136	10.05	0.053
DYS19	2.112	10.08	0.028	DYS439	1.637	10.10	0.024
DYS533	1.433	10.11	0.010	DYS456	1.691	10.12	0.006
GATAH4	1.512	10.12	0.005	DYS393	1.654	10.13	0.003
DYS448	1.858	10.13	0.002	DYS643	2.456	10.13	0.002
DYS390	1.844	10.13	0.002	DYS391	1.058	10.13	0.002

This table shows that the most-discriminating loci may not contribute to the most-discriminating haplotypes. Furthermore, there is little additional discriminating power from Y-STR haplotypes beyond 10 loci.

Population Structure and Relatedness

Population STR 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

What questions can we answer with these data, and how?

HapMap III SNP Data

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

HapMap SNP Data

Some allele frequencies are:

SNP	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MXL	MKK	TSI	YRI
1	0.4789	0.8375	0.9000	0.9143	0.8133	0.8631	0.5060	0.8169	0.5263	0.8506	0.4049
2	0.0704	0.0932	0.4684	0.4357	0.2831	0.4085	0.1084	0.0423	0.1382	0.1104	0.0525
3	0.5563	0.8735	0.9000	0.9143	0.8373	0.8795	0.5663	0.8310	0.6355	0.9156	0.4907
4	0.3944	0.1512	0.1125	0.1214	0.2831	0.1548	0.4819	0.2817	0.2924	0.2338	0.3988
5	0.3732	0.5957	0.6076	0.6812	0.5602	0.4695	0.2530	0.4718	0.3676	0.5909	0.3405
6	0.6690	0.8272	0.9000	0.9071	0.6988	0.7976	0.7952	0.7143	0.8187	0.7597	0.7362
7	0.6197	0.0216	0.4375	0.4500	0.1084	0.4643	0.6024	0.1268	0.4532	0.0390	0.7270
8	0.3803	0.9784	0.5625	0.5357	0.8916	0.5357	0.3795	0.8732	0.5205	0.9610	0.2669
9	0.2183	0.7407	0.4750	0.5000	0.6566	0.4167	0.2439	0.5915	0.4006	0.6908	0.1265
10	0.0986	0.0031	0.0886	0.0286	0.0120	0.0952	0.3012	0.0286	0.3588	0.0519	0.1933

What questions can we answer with these data, and how?

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

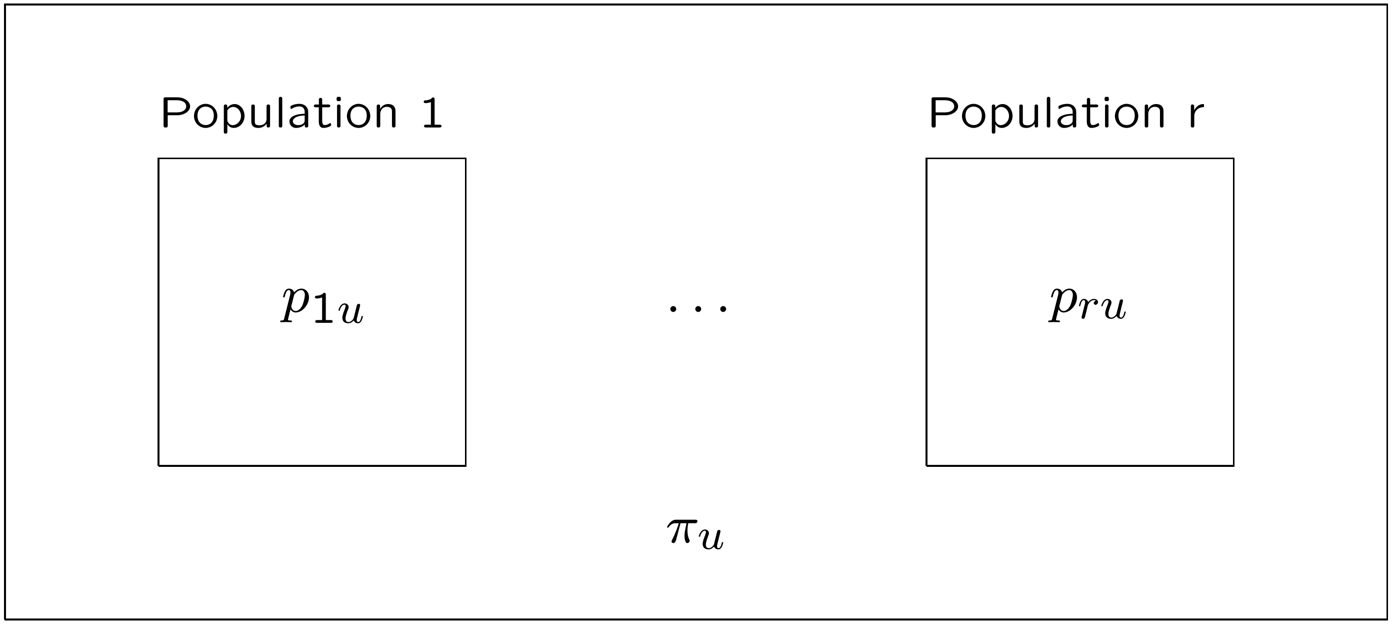
Possible to approach these data from purely statistical viewpoint.

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.

Genetic Analysis: Frequencies of Allele A_u

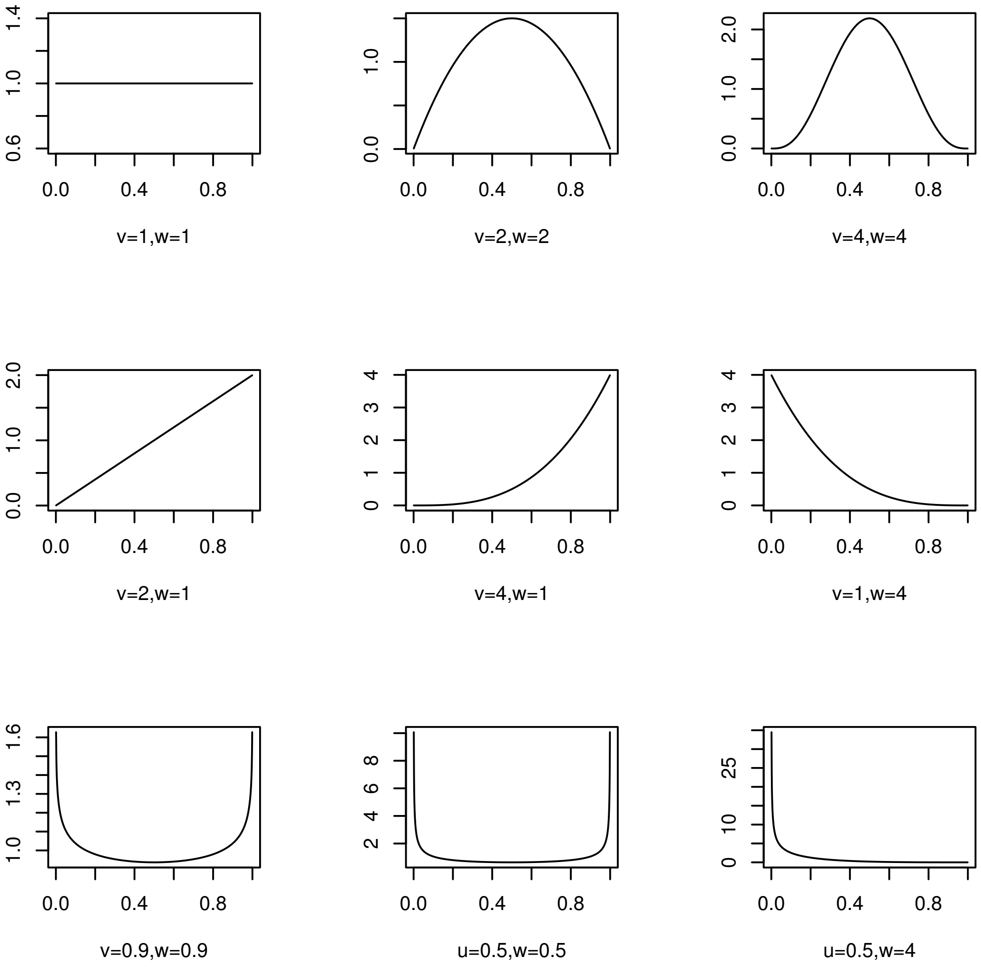


Among samples of n_i alleles from population i: counts for allele u follow a binomial distribution with mean p_{iu} and variance $n_i p_{iu}(1-p_{iu})$. Sample allele frequencies $\tilde{p}_{iu}$ have expected values p_{iu} and variances $p_{iu}(1-p_{iu})/n_i$.

Among replicates of population i: p_{iu} values follow a Beta distribution with mean π_u and variance $\pi_u(1-\pi_u)\theta^i$.

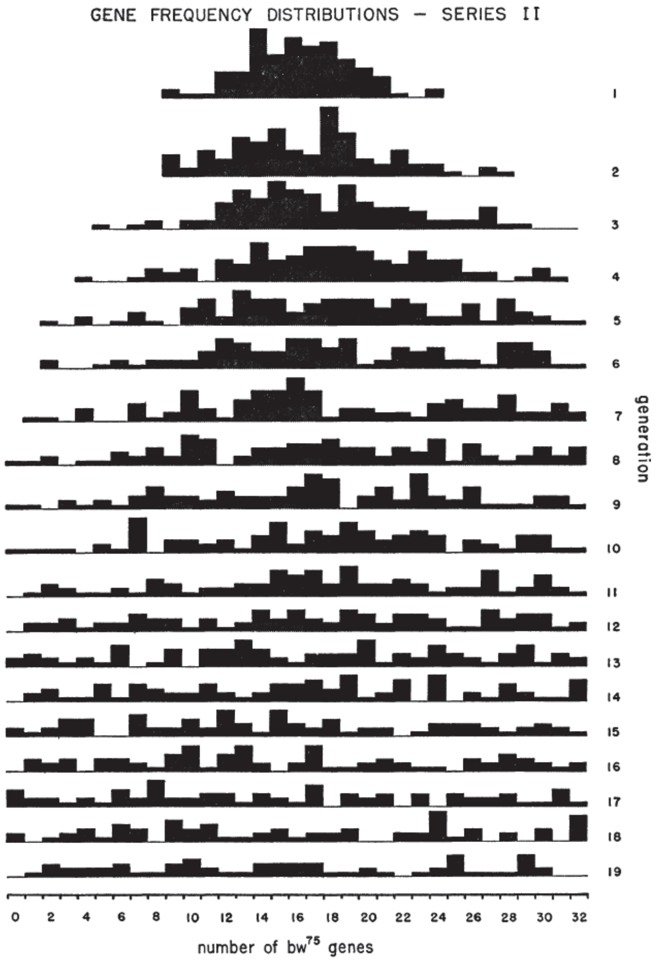
Beta distribution: Theoretical

The beta probability density is proportional to $p^{\mathsf{v}-1}(1-p)^{\mathsf{w}-1}$ and can take a variety of shapes.



Beta distribution: Experimental

The beta distribution is suggested by a *Drosophila* experiment with 107 replicate populations of size 16, starting with all heterozygotes, by P. Buri (Evolution 10:367, 1956).


GENE FREQUENCY DISTRIBUTIONS — SERIES II
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
generation
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
number of bw^{75} genes

What is θ?

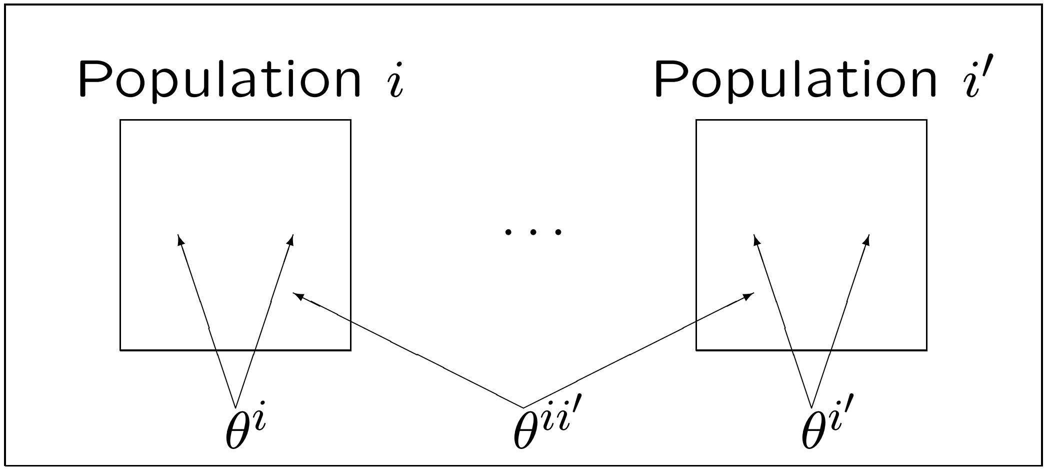
Two ways of thinking about θ.

It measures the extra degree of relatedness of individuals because they belong to the same population. We think of this reflecting long-term evolutionary history, but can use the same logic for people related because they are in the same family: first cousins have a θ value of 0.0625.

θ also helps to measure the variance of allele frequencies over populations.

These notes rest on “A unified characterization of population structure and relatedness” by Weir and Goudet, to be published in the August 2017 issue of Genetics. It was made available on May 26, 2017 in Genetics Early Online.
https://doi.org/10.1534/genetics.116.198424

What is θ?



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

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

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

Predicted Values of the θ's: Pure Drift

Our estimation procedure for the θ's holds for all evolutionary scenarios, but the theoretical values of the θ's do depend on the history of the sampled populations.

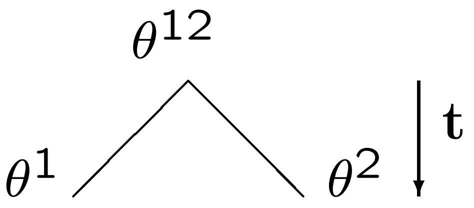
In the case of pure drift, where population i has constant size N_i and there is random mating, t generations after the population began drifting from an ancestral population in which $\theta^i = 0$

$$\theta^i(t) \;=\; 1 - \left(1 - \frac{1}{2N_i}\right)^t$$

If t is small relative to large N_i's, $\theta^i(t) \approx t/(2N_i)$, and $\theta^W(t) \approx t/(2N_h)$ where N_h is the harmonic mean of the N_i.

Drift Model: Two Populations

Now allow ancestral population itself to have ibd alleles with probability θ^{12} (the same value as for one allele from current populations 1 and 2):

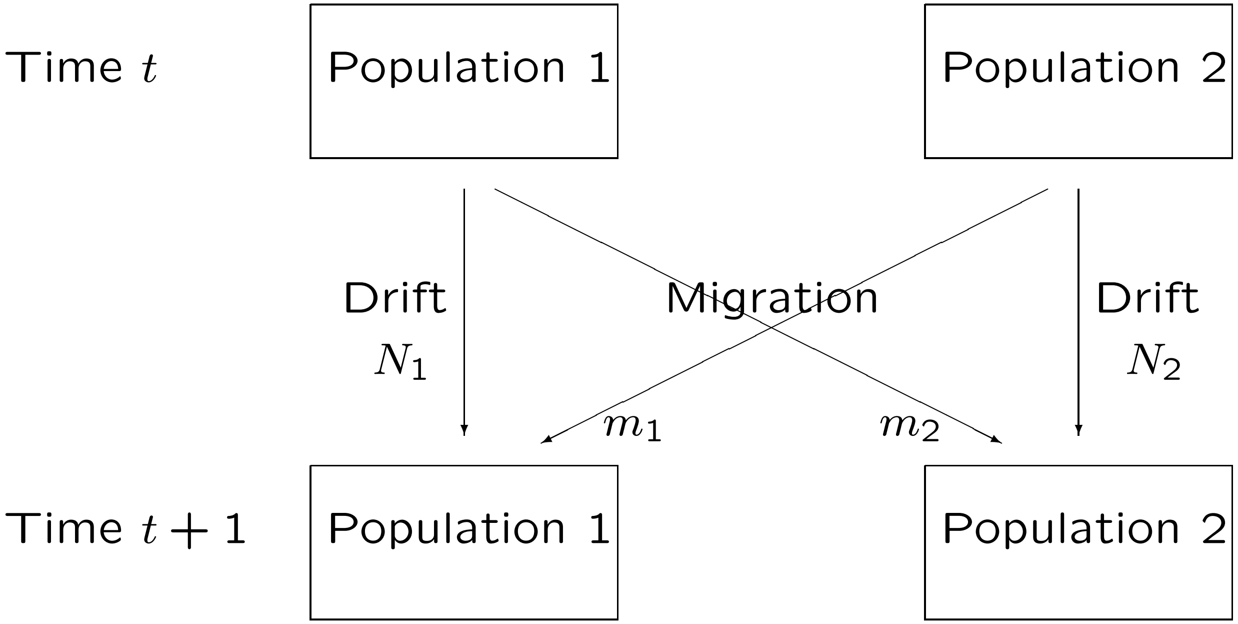


$$\theta^i \;=\; 1-(1-\theta^{12})\left(\frac{2N_i-1}{2N_i}\right)^t, \;\; i=1,2$$

We avoid needing to know the ancestral value θ^{12} by making θ^1, θ^2 *relative to* θ^{12}:

$$\beta^i = \frac{\theta^i-\theta^{12}}{1-\theta^{12}} \;=\; 1-\left(\frac{2N_i-1}{2N_i}\right)^t \approx \frac{t}{2N_i}, \;\; i=1,2$$

Two populations: drift, migration, mutation



There is also a probability μ that an allele mutates to a new type.

Drift, Mutation and Migration

For populations 1 or 2 with sizes N_1 or N_2, if m_1 or m_2 are the proportions of alleles from population 2 or 1, the changes in the θ's from generation t to $t+1$ are

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

where $\phi^i(t) = 1/(2N_i) + (2N_i-1)\theta^i(t)/(2N_i)$ and μ is the infinite-allele mutation rate.

It is possible that both of β^1, β^2 are positive, or that one of them is negative and the other one positive.

Drift and Mutation

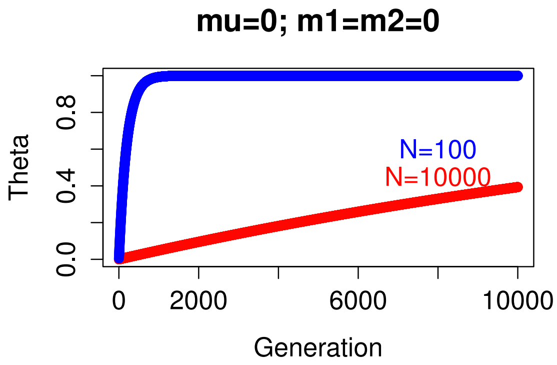
If there is no migration, the θ's tend to equilibrium values of

$$\begin{aligned} \widehat{\theta}^{1} &\approx \frac{1}{1+4N_1\mu} \\ \widehat{\theta}^{2} &\approx \frac{1}{1+4N_2\mu} \\ \widehat{\theta}^{12} &= 0 \end{aligned}$$

so $\beta^i = \theta^i, \ i = 1, 2.$

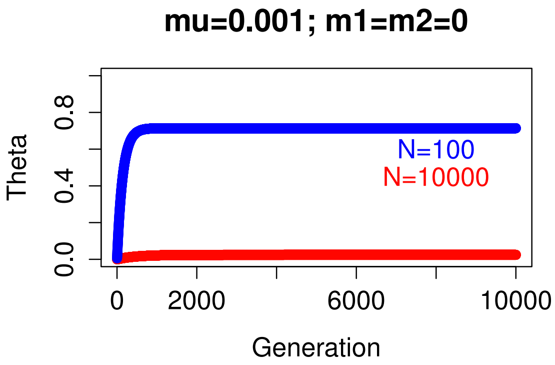
Drift, Mutation and Migration

The θ's are non-negative, but one of the β's may be negative.



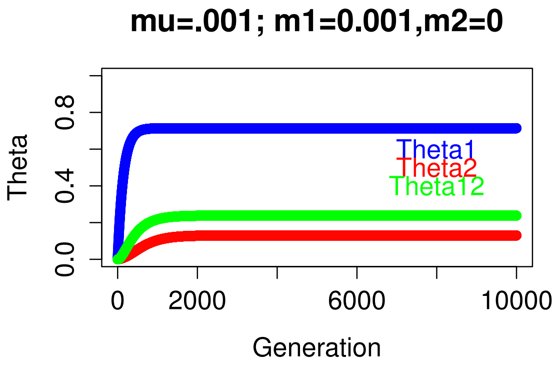
Drift Only

$\beta^1, \beta^2 > 0$



Drift and Mutation

$\beta^1, \beta^2 > 0$



Drift, Mutation and Migration
$\beta^1 > 0, \beta^2 < 0$

Genotypes vs Alleles

So far we have ignored individual genotypic structure, leading to an analysis of population allele frequencies as opposed to genotypic frequencies.

θ^i is the probability two alleles drawn randomly from population i are ibd, and $\theta^{ii'}$ is the probability an allele drawn randomly from population i is ibd to an allele drawn from population i'.

Within population i, we define θ^i_{jj} as the probability that two alleles drawn randomly from individual j are ibd, and $\theta^i_{jj'}$ as the probability that allele drawn randomly from individual j is ibd to an allele from individual j'.

Coancestry vs Inbreeding

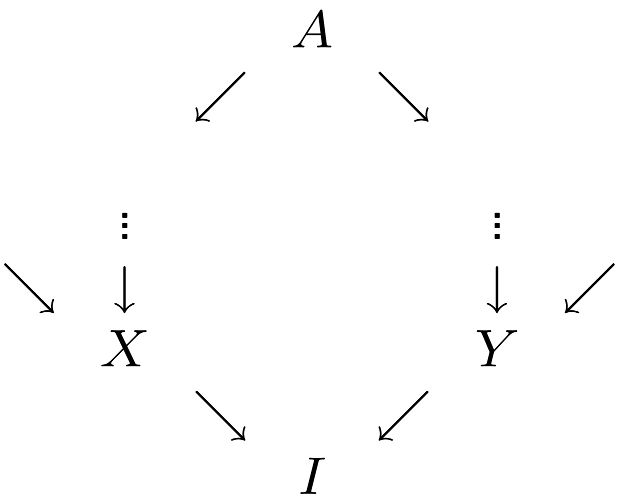
The coancestry of individuals j, j' in population i is the probability an allele from j is ibd to an allele from j'. This is $\theta^i_{jj'}$.

The inbreeding of individual j in population i is the probability the two alleles in that individual are ibd. Write this as F^i_j.

Two alleles drawn from individual j are equally likely to be the same allele or different alleles:

$$\theta^i_{jj} \ = \ \frac{1}{2}\left(1 + F^i_j\right)$$

Predicted Values: 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

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

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

Average Coancestries

The average over all pairs of distinct individuals, $j \neq j'$, of the coancestries $\theta^i_{jj'}$ is written as θ^i_S. The average of this over populations is θ^S. These are probabilities for individuals.

When there is random mating and Hardy-Weinberg equilibrium in a population, any pair of distinct alleles in a population (within or between individuals) is equivalent and then the average ibd probability for all these pairs is written as θ^i_W, where W means within populations. The average over populations is θ^W. These are probabilities for distinct alleles.

The ibd probability for any allele from population i and any allele from population i' is $\theta^{ii'}_B$, where B means between populations. Averaging over all pairs of distinct populations gives θ^B.

Relative Inbreeding: F_{IS}

For population i, the inbreeding coefficient for individual j, *relative to* the identity of pairs of alleles between individuals in that population, is

$$\beta_j^i = \frac{F_j^i - \theta_S^i}{1 - \theta_S^i}$$

The average over individuals within this population is the population-specific F_{IS}^i, and it compares within-indivdual ibd to between-individual ibd in the same population. It is the quantity being addressed by Hardy-Weinberg testing in population i.

If the reference set of alleles is for pairs of individuals within populations, averaged over populations, then the average relative inbreeding coefficient is $\beta_{IS} = (F^I - \theta^S)/(1 - \theta^S)$ where F^I is the average of F_j^i over individuals j and populations i. It is generally called F_{IS}.

Relative Inbreeding: F_{IT}

For population i, the inbreeding coefficient for individual j, *relative to* the identity of pairs of alleles from different populations averaged over all pairs of populations, is

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

The average over individuals within this population is the population-specific F^i_{IT}. The average of these over all populations is the total inbreeding coefficient $F_{IT} = (F^I - \theta^B)/(1 - \theta^B)$.

Relative Coancestry: F_{ST} for Individuals

For population i, the coancestry of individuals j, j' relative to the coancestry for all pairs of individuals in that population is

$$\beta^i_{jj'} = \frac{\theta^i_{jj'} - \theta^i_S}{1 - \theta^i_S}$$

and these average zero over all pairs of individuals in the population.

If the reference set is all pairs of alleles, one from each of two populations,

$$\beta^i_{jj'} = \frac{\theta^i_{jj'} - \theta^B}{1 - \theta^B}$$

The average β^i_{ST} over all pairs of individuals in population i is the population-specific F^i_{ST}, and averaging this over populations gives the global $F_{ST} = (\theta^S - \theta^B)/(1 - \theta^B)$. It is the ibd probability between individuals within populations relative to the ibd probability between populations.

F_{IS}, F_{IT}, F_{ST} for Individuals

When individuals are distinguished:

$$
\begin{aligned}
(1 - F_{IT}) &= (1 - F_{IS})(1 - F_{ST}) \\
F_{IS} &= \frac{F_{IT} - F_{ST}}{1 - F_{ST}}
\end{aligned}
$$

This classic result also holds for population-specific values

$$
\begin{aligned}
(1 - F^i_{IT}) &= (1 - F^i_{IS})(1 - F^i_{ST}) \\
F^i_{IS} &= \frac{F^i_{IT} - F^i_{ST}}{1 - F^i_{ST}}
\end{aligned}
$$

Relative Coancestry: F_{ST} for Alleles

For random union of gametes, when individual structure is not needed, the ibd probability θ^i_W for any distinct pair of alleles within population i relative to the ibd probability between populations is

$$\beta^i_{WT} \ = \ \frac{\theta^i_W - \theta^B}{1 - \theta^B}$$

This is the population-specific F^i_{ST} for alleles.

Averaging over populations:

$$\beta_{WT} \ = \ \frac{\theta^W - \theta^B}{1 - \theta^B}$$

and this is the global F_{ST} for alleles.

Estimation of β's

Now that the various measures for individuals and populations are defined, it is straightforward to estimate them from genotypic or allelic data.

For any allele we set its indicator variable x_u to be 1 if that allele is of type u and to be 0 if it is not of type u. Then the expected value of x_u or of x_u^2 is the allele frequency π_u. For any two distinct alleles, the expected value of the product of their x_u's is $\pi_u^2 + \pi_u(1 - \pi_u)\theta$ where θ is the ibd probability for those two alleles.

We examine pairs of alleles to see whether or not they are the same type: i.e. they match. When two alleles match, the product of their indicators is 1 and the expected value of the proportion of matching pairs depends on θ. We manipulate matching proportions to give estimates of the θ's.

Allelic Matching Proportions for Individuals

For individual j in population i the proportion of its distinct alleles that match is either 0 or 1. It is convenient to work with allele dosages X^i_{ju} for the number of its alleles that are of type u and write the matching proportion as

$$\tilde{M}^i_j = \frac{1}{2}\sum_u X^i_{ju}(X^i_{ju}-1)$$

For biallelic SNPs, the reference allele dosage is X^i_j (with values 0, 1 or 2) and the matching proportion is $(X^i_j-1)^2$ - this is still 0 or 1. The expected value over replicates of the population is

$$\mathcal{E}(\tilde{M}^i_j) = M_T + (1-M_T)F^i_j$$

where $M_T = \sum_u \pi^2_u$ is unknown.

We avoid having to know M_T by estimating F^i_j relative to either pairs of alleles from different individuals (i.e. coancestry) in the same population, or pairs of alleles from different populations.

Allelic Matching Proportions for Pairs of Individuals

The proportion of pairs of alleles, one from individual j and one from j' in population i that match is 0, 0.5 or 1. These proportions can be expressed in terms of allele dosages

$$\tilde{M}^i_{jj'} \;=\; \frac{1}{4}\sum_u X^i_{ju}X^i_{j'u}$$

For biallelic SNPs, with reference allele dosages X^i_j the matching proportion is $[1+(X^i_j-1)(X^i_{j'}-1)]/2$ - this is still 0, 0.5 or 1. The expected value over replicates of the population is

$$\mathcal{E}(\tilde{M}^i_{jj'}) \;=\; M_T+(1-M_T)\theta^i_{jj'}$$

Averaging over all pairs of individuals, the observed and expected matching proportion are

$$\tilde{M}^i_S=\frac{1}{n_i(n_i-1)}\sum_{j=1}^{n_i}\sum_{j'\neq j}^{n_i}\tilde{M}^i_{jj'} \quad , \quad \mathcal{E}(\tilde{M}^i_S)=M_T+(1-M_T)\theta^i_S$$

Within-population Inbreeding Estimates

The inbreeding coefficient of an individual can be estimated as its allelic matching proportion relative to the matching proportion for pairs of individuals in the same population:

$$\widehat{\beta}_j^i = \frac{\tilde{M}_j^i - \tilde{M}_S^i}{1 - \tilde{M}_S^i} \quad , \quad \mathcal{E}(\widehat{\beta}_j^i) = \frac{F_j^i - \theta_S^i}{1 - \theta_S^i}$$

These estimates may be negative.

Averaging over individuals, with $\tilde{M}_I^i = \sum_{j=1}^{n_i} \tilde{M}_j^i / n_i$,

$$\widehat{\beta}_{IS}^i = \widehat{F}_{IS}^i = \frac{\tilde{M}_I^i - \tilde{M}_S^i}{1 - \tilde{M}_S^i} \quad , \quad \mathcal{E}(\widehat{\beta}_I^i) = F_{IS}^i = \frac{F_I^i - \theta_S^i}{1 - \theta_S^i}$$

If there are data from several populations, we may use $\tilde{M}^S$ as the average within-population coancestry, averaged over populations, so that inbreeding is estimated relative to this average.

Within-population Coancestry Estimates

The coancestry coefficient of two individuals can be estimated as their allelic matching proportion relative to the matching proportion for pairs of individuals in the same population:

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

These estimates have been constructed to have an average value of zero over pairs of individuals in a population.

Comparison with Standard Inbreeding and Coancestry Estimates

For SNPs, omitting population superscripts, we estimate inbreeding and coancestry by

$$\hat{\beta}_j = \frac{(X_j - 1)^2 - \tilde{M}_S}{1 - \tilde{M}_S} \quad , \quad \mathcal{E}(\hat{\beta}_j) = \frac{F_j - \theta_S}{1 - \theta_S}$$

$$\hat{\beta}_{jj'} = \frac{\frac{1}{2}[1 + (X_j - 1)(X_{j'} - 1)] - \tilde{M}_S}{1 - \tilde{M}_S} \quad , \quad \mathcal{E}(\hat{\beta}_{jj'}) = \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}$$

where $\tilde{M}_S = \sum_{j=1}^{n} \sum_{j' \neq j}^{n} \tilde{M}_{jj'}/[n(n-1)]$.

Standard estimates use sample allele frequencies $\tilde{p} = \sum_{j=1}^{n} X_j^i/(2n)$:

$$\begin{aligned} \hat{\beta}_j &= \frac{(X_j - 2\tilde{p})^2}{2\tilde{p}(1-\tilde{p})} - 1 \\ \hat{\beta}_{jj'} &= \frac{(X_j - 2\tilde{p})(X_{j'} - 2\tilde{p})}{4\tilde{p}(1-\tilde{p})} \end{aligned}$$

Standard estimate expectations

The standard estimates have expectations that depend on the inbreeding and coancestry coefficients of all members of the sample. This can result in pairs of individuals having β rankings that differ from their θ rankings.

$$\mathcal{E}(\hat{\beta}_{jj'}) = \frac{(\theta_{jj'} - \psi_j - \psi_{j'} + \theta_S) - \frac{1}{n}(\theta_{jj} + \theta_{j'j'} - \psi_j - \psi_{j'} - F_S + \theta_S)}{(1 - \theta_S) - \frac{1}{n}(F_S - \theta_S)}$$

where $F_S = \sum_{j=1}^{n} \theta_{jj}/n$ is the average of all within-individual coancestries $(1 + F_j)/2$, $\psi_j = \sum_{j'=1, j'\neq j}^{n} \theta_{jj'}/(n-1)$ is the average coancestry of individual j to all other individuals, and $\theta_S = \sum_{j=1}^{n} \psi_j/n$.

When $j = j'$, we have $\theta_{jj} = (1 + F_j)/2$.

Matching Proportions for Populations

Population-structure measures are estimated by comparing within-population allelic matching proportions to those between populations.

When genotypic data are available, the between-individual within-population matching proportions use the averages over all pairs of individuals within the population, from above, is

$$\tilde{M}_S^i \;=\; \frac{1}{4n_i(n_i-1)}\sum_u \sum_{j=1}^{n_i} \sum_{j'\neq j}^{n_i} X_{ju}^i X_{j'u}^i$$

and the average over pairs of populations of between-population matching proportions is

$$\tilde{M}^B \;=\; \frac{1}{r(r-1)}\sum_u \sum_{i=1}^{r} \sum_{i'\neq i}^{r} \tilde{p}_{iu}\tilde{p}_{i'u}$$

Genotype-based F_{ST} Estimates

The within- and between-population matching proportions for genotypic data set have expectations

$$\begin{aligned} \mathcal{E}(\tilde{M}_S^i) &= M_T + (1 - M_T)\theta_S^i \\ \mathcal{E}(\tilde{M}^B) &= M_T + (1 - M_T)\theta^B \end{aligned}$$

so the population-specific, genotype-based, F_{ST} estimate is

$$\hat{\beta}_{ST}^i = \hat{F}_{ST}^i \quad = \quad \frac{\tilde{M}_S^i - \tilde{M}^B}{1 - \tilde{M}^B}$$

Averaging over populations, the genotype-based global F_{ST} estimate is

$$\hat{\beta}_{ST} = \hat{F}_{ST} \quad = \quad \frac{\tilde{M}^S - \tilde{M}^B}{1 - \tilde{M}^B}$$

SNP-genotype-based F_{ST} Estimates

For SNPs, with allele dosages X:

$$\begin{aligned}
\tilde{M}_S^i &= \frac{1}{2n_i(n_i-1)} \sum_{j=1}^{n_i} \sum_{j' \neq j}^{n_i} [1 + (X_j^i - 1)(X_{j'}^i - 1)] \\
\tilde{M}^B &= \frac{1}{r(r-1)} \sum_{i=1}^{r} \sum_{i' \neq i}^{r} [\tilde{p}_i \tilde{p}_{i'} + (1 - \tilde{p}_i)(1 - \tilde{p}_{i'})]
\end{aligned}$$

Allele-based Matching Proportions within Populations

When the genotypic structure of data is ignored, or not known, allelic data can be used to estimate F_{ST}.

If alleles from n_i individuals are observed for population i, the observed matching proportion of allele pairs within this population is

$$\begin{aligned} \tilde{M}_W^i &= \frac{1}{2n_i(2n_i-1)}\sum_u \sum_{j=1}^{2n_i} \sum_{j'\neq j}^{2n_i} x_{ju}x_{j'u} \\ &= \frac{2n_i}{2n_i-1}\sum_u \tilde{p}_{iu}^2 - \frac{1}{2n_i-1} \end{aligned}$$

where $\tilde{p}_{iu}$ is the sample frequency for allele u for this population.

The expected value of this over replicates of the population is

$$\mathcal{E}(\tilde{M}_W^i) = M_T + (1-M_T)\theta_W^i$$

where $M_T = \sum_u \pi_u^2$.

Allele-based Matching Proportions between Populations

The observed proportion of matching allele pairs between populations i and i' is

$$\begin{aligned}\tilde{M}_B^{ii'} &= \frac{1}{4n_in_{i'}}\sum_u \sum_{j=1}^{2n_i}\sum_{j'=1}^{2n_{i'}} x_{ju}x_{j'u} \\ &= \sum_u \tilde{p}_{iu}\tilde{p}_{i'u}\end{aligned}$$

The expected value of this over replicates of the population is

$$\mathcal{E}(\tilde{M}_B^{ii'}) = M_T + (1-M_T)\theta_B^{ii'}$$

and, averaging over all pairs of populations

$$\mathcal{E}(\tilde{M}^B) = M_T + (1-M_T)\theta^B$$

Allele-based Estimate of F_{ST}

We avoid having to know M_T by considering allele-pair matching within a population relative to the allele-pair matching between pairs of populations:

$$\widehat{\beta}^i_{WT} = \widehat{F}^i_{ST} \;=\; \frac{\tilde{M}^i_W - \tilde{M}^B}{1 - \tilde{M}^B}$$

and this has expected value $F^i_{WT} = (\theta^i_W - \theta^B)/(1 - \theta^B)$ which is the population-specific value.

Average over populations:

$$\widehat{F}_{WT} = \widehat{\beta}_{WT} \;=\; \frac{\tilde{M}^W - \tilde{M}^B}{1 - \tilde{M}^B}$$

and this has expected value $F_{WT} = (\theta^W - \theta^B)/(1 - \theta^B)$ which is the global value.

Simple Computing Equations for F_{ST}

For large sample sizes and r populations:

$$\begin{aligned}
\tilde{M}_W^i &\approx \sum_u \tilde{p}_{iu}^2 \\
\tilde{M}^W &= \frac{1}{r}\sum_{i=1}^{r} \tilde{M}_W^i = \sum_u \bar{p}_u^2 + \frac{r-1}{r}\sum_u s_u^2
\end{aligned}$$

where $\bar{p}_u = \sum_{i=1}^r \tilde{p}_{iu}/r$ is the mean allele frequency over populations, and $s_u^2 = \sum_{i=1}^r (\tilde{p}_{iu} - \bar{p}_u)^2/(r-1)$ is the variance of allele frequencies over populations.

For all sample sizes:

$$\begin{aligned}
\tilde{M}_B^{ii'} &= \sum_u \tilde{p}_{iu}\tilde{p}_{i'u} \\
\tilde{M}^B &= \frac{1}{r(r-1)}\sum_{i=1}^{r}\sum_{i'\neq i}^{r} \tilde{M}_B^{ii'} = \sum_u \bar{p}_u^2 - \frac{1}{r}\sum_u s_u^2
\end{aligned}$$

Simple Allele-based Estimates for F_{ST}

The allele-based population-specific estimates are

$$\widehat{F}^i_{WT} \quad = \quad \frac{\sum_u [(\tilde{p}^2_{iu} - \bar{p}^2_u) + s^2_u]}{\sum_u [\bar{p}_u(1 - \bar{p}_u) + \frac{1}{r} s^2_u}$$

The corresponding global estimates are

$$\widehat{F}_{WT} \quad = \quad \frac{\sum_u s^2_u}{\sum_u [\bar{p}_u(1 - \bar{p}_u) + \frac{1}{r} s^2_u]}$$

SNP-allele-based Estimates for F_{ST}

For SNPs with two alleles, write $\tilde{p}_i$ for the reference allele sample frequency for population i. In this case, $\sum_u \bar{p}_u(1-\bar{p}_u) = 2\bar{p}(1-\bar{p})$ where $\bar{p} = \sum_{i=1}^r \tilde{p}_i/r$, and $\sum_u s_u^2 = 2s^2$ where $s^2 = \sum_{i=1}^r(\tilde{p} - \bar{p})^2/(r-1)$.

The population-specific estimates are

$$\hat{F}_{WT}^i \;=\; \frac{\tilde{p}_i^2 - \bar{p}^2 + s^2}{\bar{p}(1-\bar{p}) + \frac{1}{r}s^2}$$

The global estimates are

$$\hat{F}_{WT} \;=\; \frac{s^2}{\bar{p}(1-\bar{p}) + \frac{1}{r}s^2}$$

SNP-allele-based Estimates of F_{ST} for Two Populations

For SNP studies with two populations, with reference allele sample frequencies $\tilde{p}_1, \tilde{p}_2$, $\bar{p} = (\tilde{p}_1 + \tilde{p}_2)/2$ and $s^2 = (\tilde{p}_1 - \tilde{p}_2)^2/2$.

The population-specific estimates are

$$\hat{\beta}^1_{WT} = \hat{F}^1_{WT} = \frac{(\tilde{p}_1 - \tilde{p}_2)(2\tilde{p}_1 - 1)}{\tilde{p}_1(1 - \tilde{p}_2) + \tilde{p}_2(1 - \tilde{p}_1)}$$

$$\hat{\beta}^2_{WT} = \hat{F}^2_{WT} = \frac{(\tilde{p}_2 - \tilde{p}_1)(2\tilde{p}_2 - 1)}{\tilde{p}_1(1 - \tilde{p}_2) + \tilde{p}_2(1 - \tilde{p}_1)}$$

One of these two may be negative.

The global estimate is

$$\hat{\beta}_{WT} = \hat{F}_{WT} = \frac{(\tilde{p}_1 - \tilde{p}_2)^2}{\tilde{p}_1(1 - \tilde{p}_2) + \tilde{p}_2(1 - \tilde{p}_1)}$$

This must be positive.

Multiple Loci

The unweighted estimators for locus l are of the form

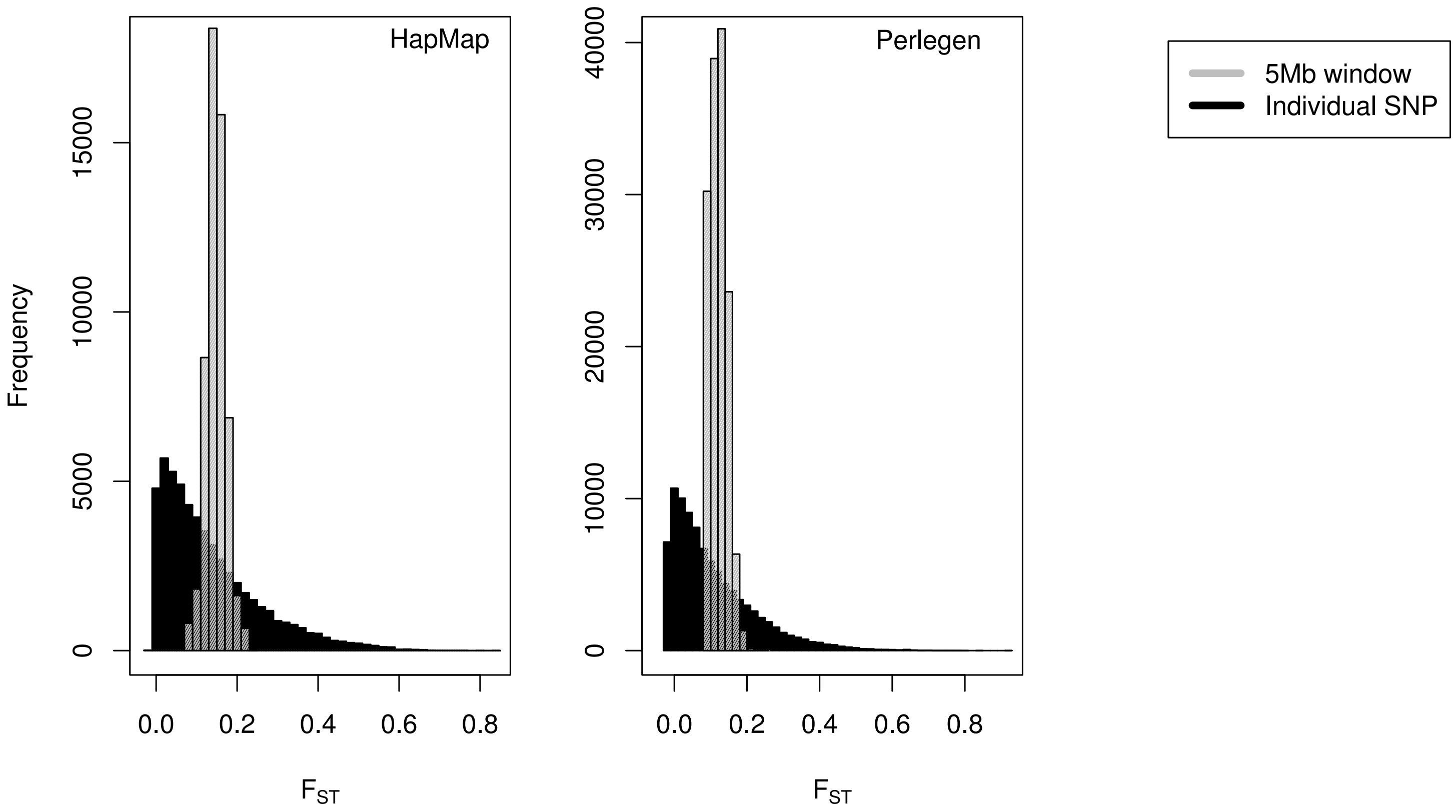
$$\hat{\beta}_l = \frac{\text{Numerator}_l}{\text{Denominator}_l}$$

There are several ways to combine estimates over loci. Here, we weight by the denominators: With several loci, these can be extended to

$$\hat{\beta} = \frac{\sum_l \text{Numerator}_l}{\sum_l \text{Denominator}_l}$$

and these estimate β if each locus has the same value of the θ's. Otherwise they estimate a weighted average of the different θ values, where the weights are functions of the allele frequencies at the loci in the sum.

Effect of Number of Loci



Worldwide Autosomal-STR Survey

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

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

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

$$\widehat{\beta}^{i}_{WT_l} = \frac{\tilde{M}^{i}_{WT_l} - \tilde{M}^{B}_{l}}{1 - \tilde{M}^{B}{}_{l}}$$

The "T" may refer to the group of populations with the same continental ancestry, or it may refer to the entire set of populations.

STR Survey: $\hat{\beta}$ Values for Groups and Loci

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

Coancestry is relative, not absolute

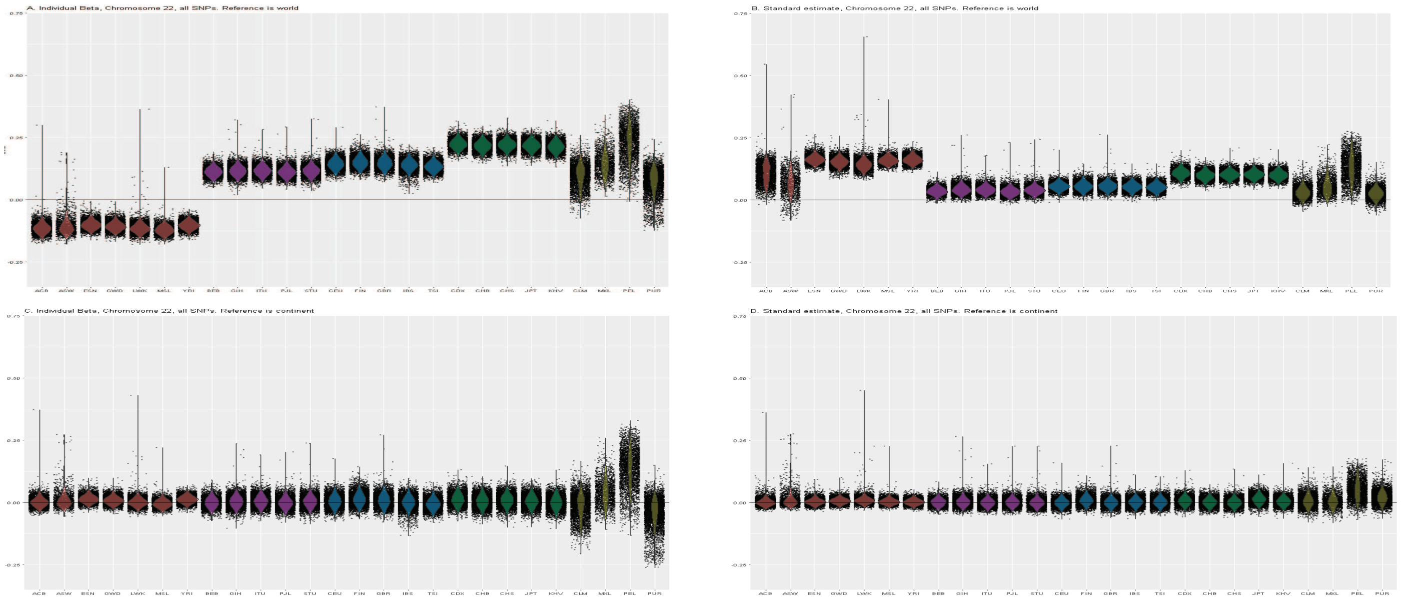
The new relative coancestry estimates have been applied to 1000 Genomes data, and compared to standard estimates, shown on next slide.

For the whole world, all 26 populations, as reference the beta estimates show a relatively narrow range of values within each African population (AFR) and lower African values than in the rest of the world, as expected from our understanding of higher genetic diversity within African than non-African populations from the migration history of modern humans. This pattern was not shown by the standard estimates - those estimates showed higher coancestry among African individuals than among non-Africans.

The wide plots for the Admixed American populations (AMR) reflect the admixture within those populations, with greater relatedness reflecting more ancestral commonality. When each continental group is used as a reference, all populations show low coancestry, except for the admixed AMR.

Coancestry is relative, not absolute

Top row: Whole world reference. Bottom row: Continental group reference.



Beta estimates

Standard estimates

Chromosome 22 data from 1000 Genomes.

Continents (left to right): AFR, SAS, EUR, EAS, AMR

Populations (l to r):**AFR**: ACB, ASW, ESN, GWD, LWK, MSL, YRI;
SAS: BEB, GIH, ITU, PJL, STU; **EUR**: CEU, FIN, GBR, IBS, TSI;
EAS: CDX, CHB, CHS, JPT; **AMR**: KHV, CLM, MXL, PEL, PUR

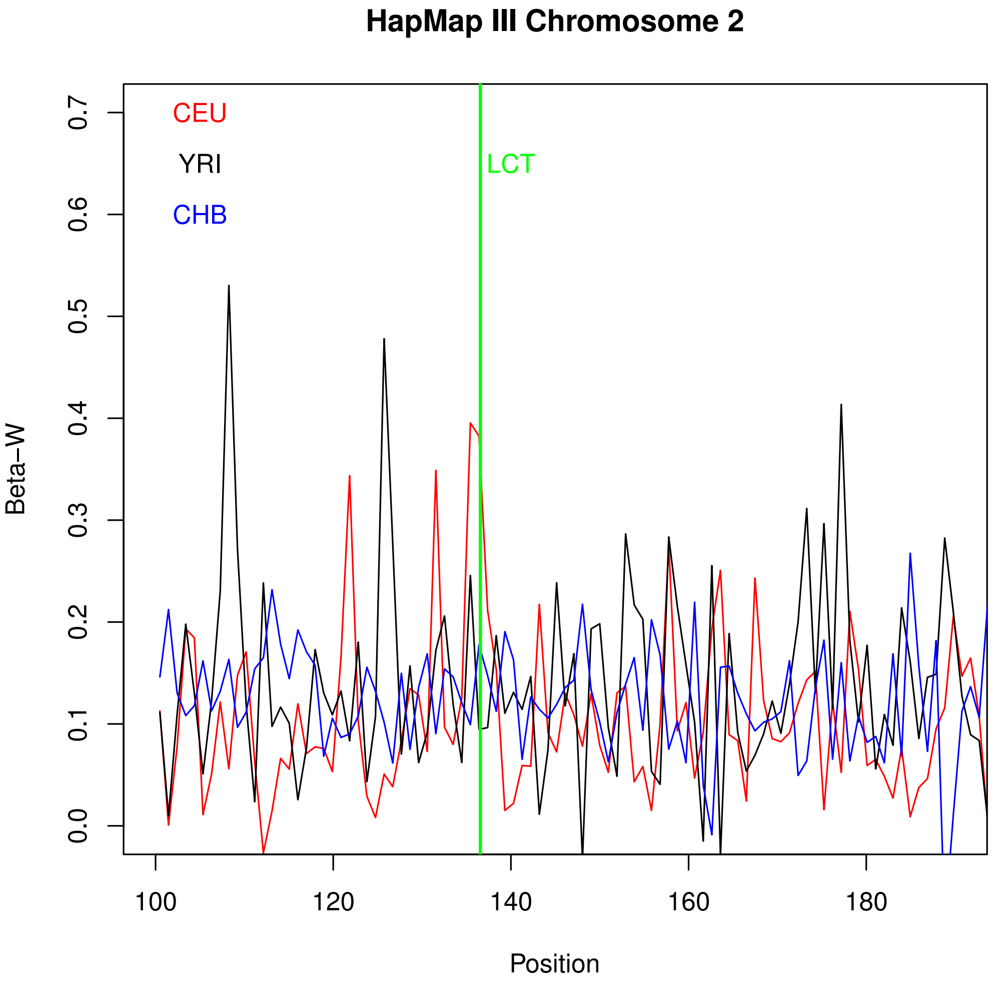
F_{ST} is relative, not absolute

Using data from the 1000 genomes, using 1,097,199 SNPs on chromosome 22.

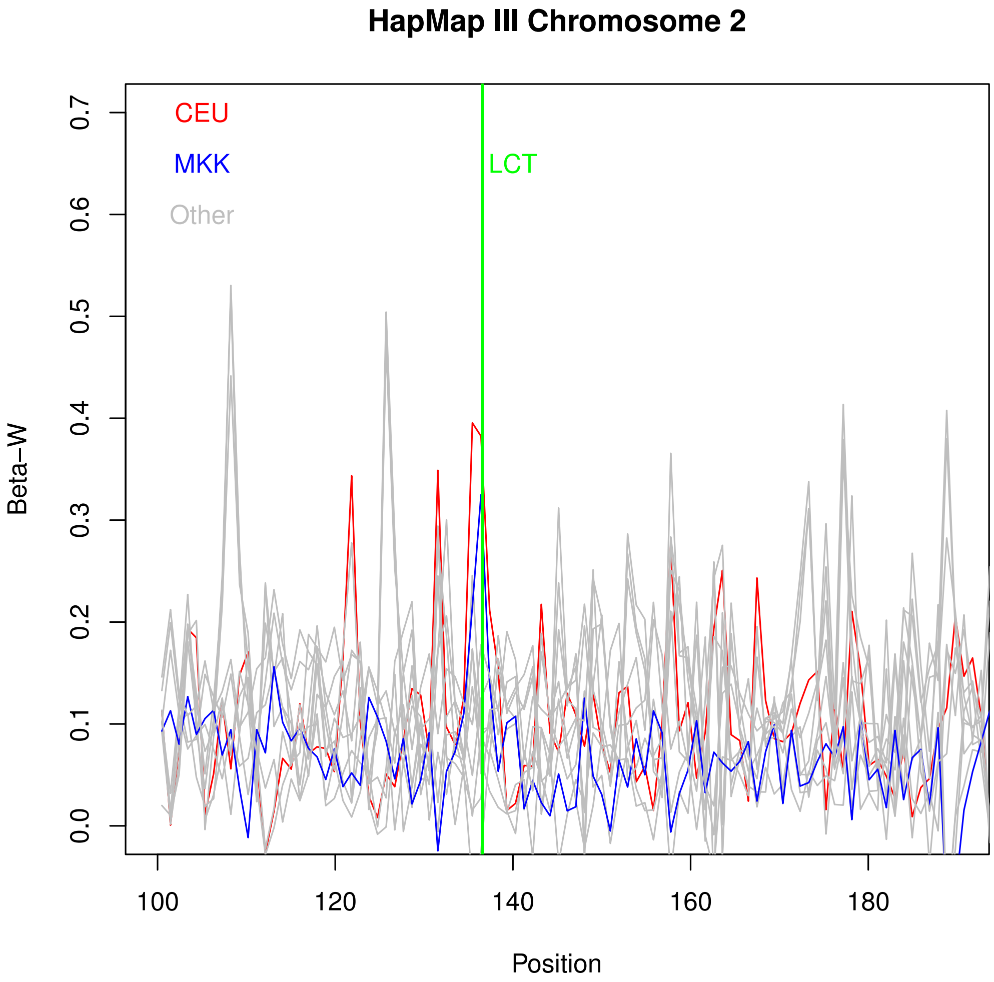
For the samples originating from Africa, there is a larger F_{WT}, $\widehat{\beta}_{WT} = 0.013$, with Africa as a reference set than there is, $\widehat{\beta}_{WT} = -0.099$, with the world as a reference set. African populations tend to be more different from each other on average than do any two populations in the world on average.

The opposite was found for East Asian populations: there is a smaller F_{WT}, $\widehat{\beta}_{WT} = 0.013$ with East Asia as a reference set than there is, $\widehat{\beta}_{WT} = 0.225$ with the world as a reference set. East Asian populations are more similar to each other than are any pair of populations in the world.

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



$\widehat{\beta}_{WT}$ 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

Weir & Cockerham 1984 Model

W&C assumed all populations have equal evolutionary histories ($\theta^i = \theta$, all i) and are independent ($\theta^{ii'} = 0$, all $i' \neq i$), and they worked with overall allele frequencies that were weighted by sample sizes

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

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

Weir & Cockerham 1984 Model

Two mean squares were constructed for each allele:

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

These have expected values

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

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

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

Weir & Cockerham 1984 Estimator

Under the β approach described here, the Weir and Cockerham estimator has expectation

$$\mathcal{E}(\widehat{\theta}_{\text{WC}}) = \frac{\theta_c^W - \theta_c^B + Q}{1 - \theta_c^B + Q} \quad \text{instead of} \quad \frac{\theta^W - \theta^B}{1 - \theta^B}$$

where

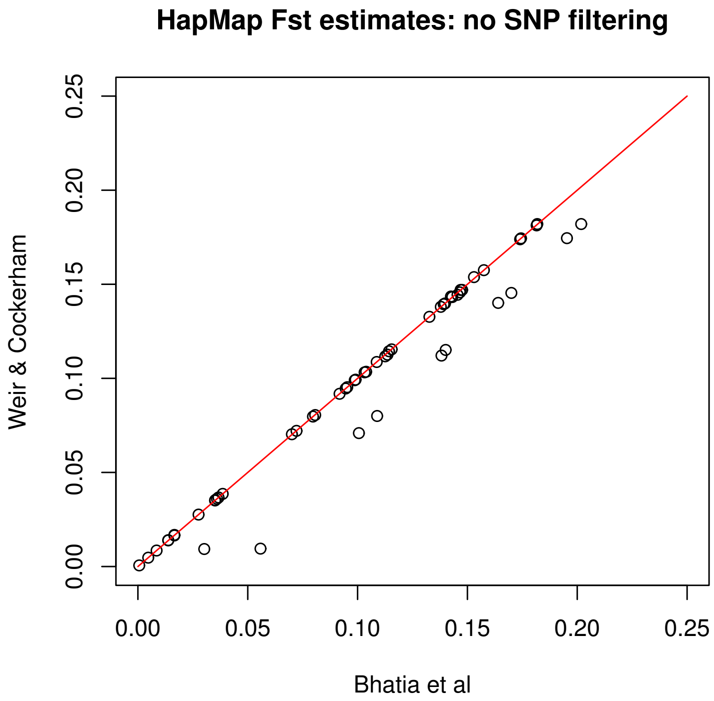
$$\theta_c^W = \frac{\sum_i n_i^c \theta^i}{\sum_i n_i^c} \quad , \quad \theta_c^B = \frac{\sum_{i \neq i'} n_i n_{i'} \theta^{ii'}}{\sum_{i \neq i'} n_i n_{i'}}$$

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

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

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

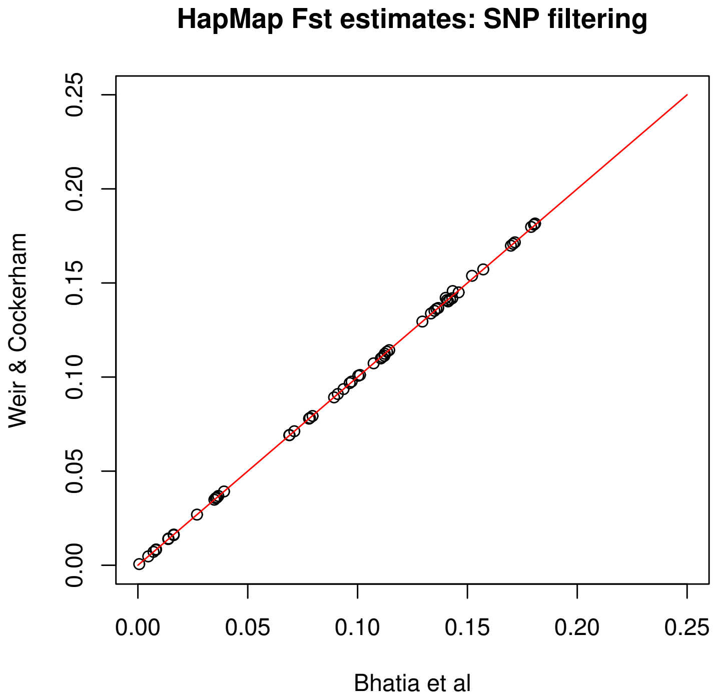
WC84 vs Beta Allele-based Estimators



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

(Bhatia et al, 2013, Genome Research 23:1514-1521.)

WC vs Unweighted Estimator



F_{WT} 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.

(Bhatia et al, 2013, Genome Research 23:1514-1521.)

MLE for Individual Inbreeding Coefficients

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

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

From Bayes' theorem then

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

MLE for Individual Inbreeding Coefficients

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

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

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

Toy Example

Suppose 5 loci have genotypes

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

.

Then the updated estimate is

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

If all the $p_l = 0.5$,

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

and this converges to $F = 0.2$.

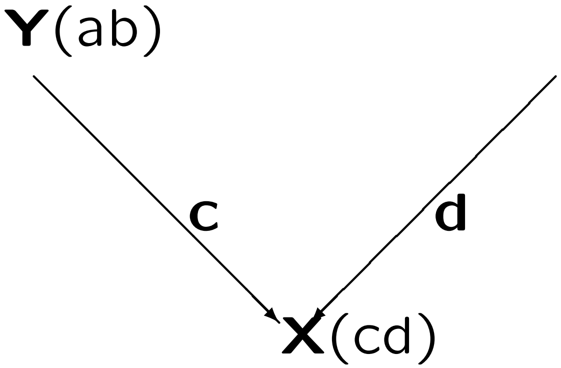
k-coefficients

The coancestry coefficient is the probability of a pair of alleles being ibd.

For joint genotypic frequencies, and for a more detailed characterization of relatedness of two non-inbred individuals, we need the probabilities that they carry 0, 1, or 2 pairs of ibd alleles. For example: their two maternal alleles may be ibd or not ibd, and their two paternal alleles may be ibd or not.

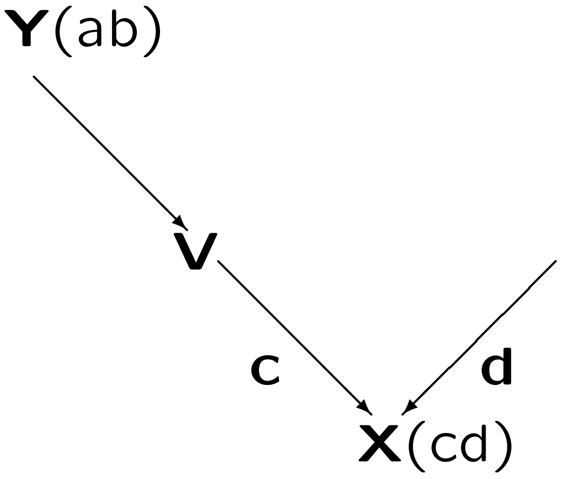
The probabilities of two individuals having 0, 1 or 2 pairs of ibd alleles are written as k_0, k_1, k_2 and $\theta = \frac{1}{2}k_2 + \frac{1}{4}k_1$.

Parent-Child



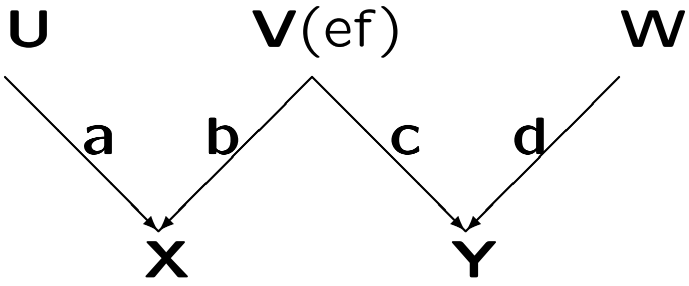
$$\Pr(c \equiv a) = 0.5, \quad \Pr(c \equiv b) = 0.5, \quad k_1 = 1$$

Grandparent-grandchild



$$\text{Pr}(c \equiv a) = 0.25, \quad \text{Pr}(c \equiv b) = 0.25, \quad k_1 = 0.5 \& k_0 = 0.5$$

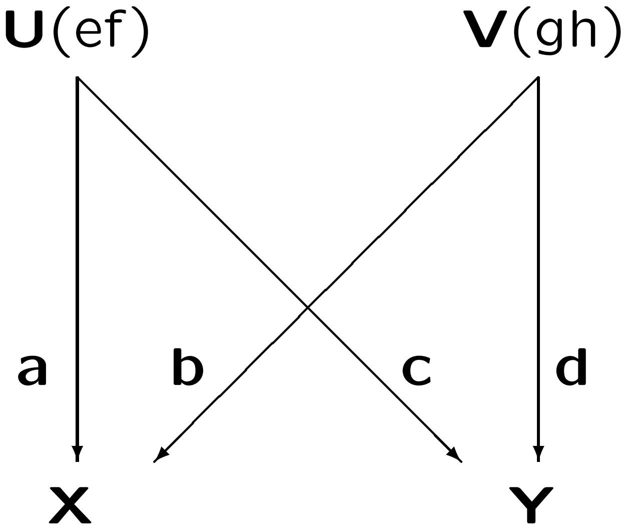
Half sibs



		0.5	0.5
		$c \equiv e$	$c \equiv f$
0.5	$b \equiv e$	0.25	0.25
0.5	$b \equiv f$	0.25	0.25

Therefore $k_1 = 0.5$ so $k_0 = 0.5$.

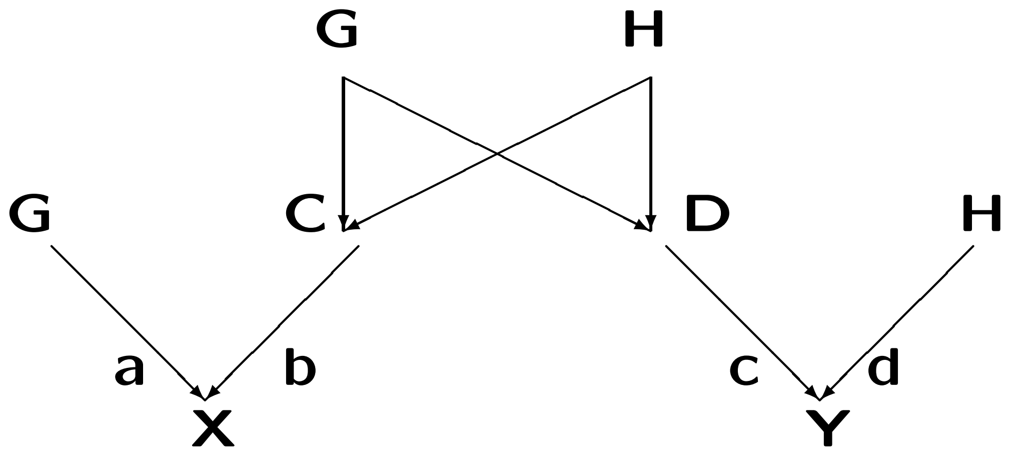
Full sibs



		0.5	0.5
		$b \equiv d$	$b \not\equiv d$
0.5	$a \equiv c$	0.25	0.25
0.5	$a \not\equiv c$	0.25	0.25

$$k_0 = 0.25, k_1 = 0.50, k_2 = 0.25$$

First cousins



Non-inbred Relatives

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

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

Method of Moments for Relatedness Coefficients

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

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

MOM Approach: k_0

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

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

sum over loci l to get

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

This gives a moment estimate

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

MOM Approach: k_1

From

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

sum over loci to get

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

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

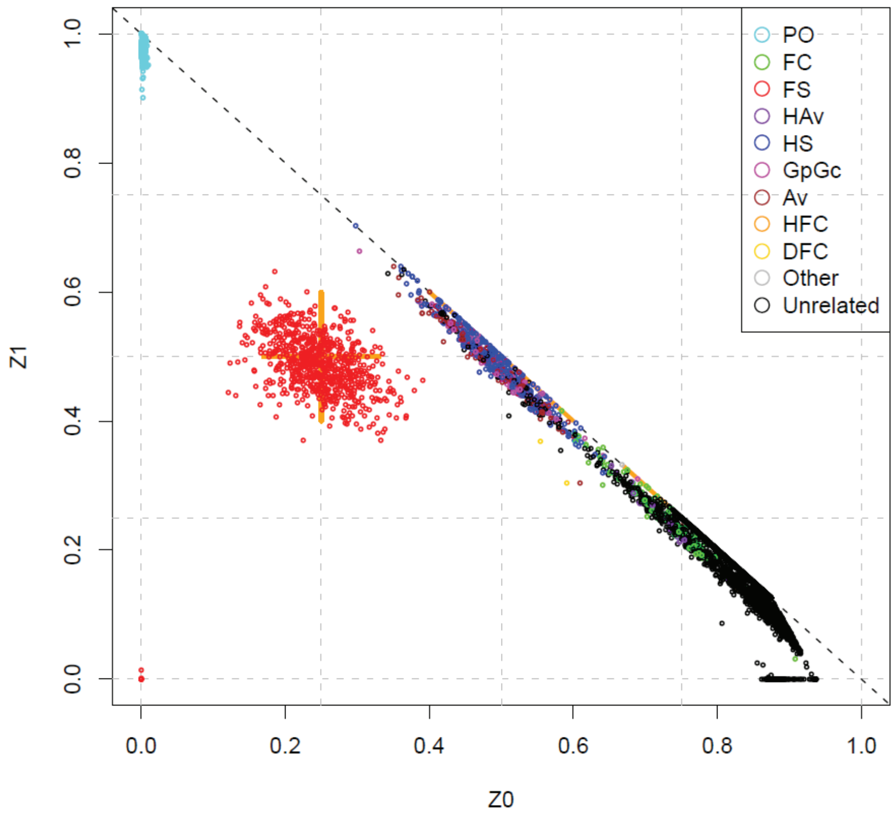
$$\text{Pr}(\text{IBD}=1) = \frac{N_1-\sum_l 4p_l(1-p_l)[p_l^2+(1-p_l)^2]\,\text{Pr}(\text{IBD}=0)}{\sum_l 2p_l(1-p_l)}$$

MOM Approach: k_2

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

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

PLINK Example



MLE for Relatedness Coefficients

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

G	$\Pr(G) = \sum_i \Pr(G\|D_i)k_i$
MM, MM	$k_2p_M^2 + k_1p_M^3 + k_0p_M^4$
mm, mm	$k_2p_m^2 + k_1p_m^3 + k_0p_m^4$
MM, mm	$2k_0p_M^2p_m^2$
MM, Mm	$2k_1p_M^2p_m + 4k_0p_M^3p_m$
mm, Mm	$2k_1p_Mp_m^2 + 4k_0p_Mp_m^3$
Mm, Mm	$2k_2p_Mp_m + k_1p_Mp_m + 4k_0p_M^2p_m^2$

MLE for Relatedness Coefficients

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

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

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

The updated estimates are obtained by averaging over L loci:

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

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

Toy Example

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

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

and

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

The updating equations are:

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

Toy Example

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

Toy Example

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

“RELPAIR” calculations

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

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

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

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

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.